Explainable Emotion Recognition Using a Valence Arousal Dominance Calibrated IndoBERT for Indonesian Counseling Chatbots

Jimmy Agustian Loekito
Department of Information System, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia | Department of Computer Engineering, Universitas Kristen Maranatha, Bandung, Indonesia
7026231002@student.its.ac.id

Aris Tjahyanto
Department of Information System, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
aristj@its.ac.id (corresponding author)

Rarasmaya Indraswari
Department of Information System, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
raras@its.ac.id

Received: 15 January 2026 | Revised: 6 March 2026 and 30 March 2026 | Accepted: 10 April 2026

 DOI: https://doi.org/10.48084/etasr.17581

ABSTRACT

Emotion recognition in text-based counseling is challenging as emotional cues are often subtle, implicit, and context-dependent. In contrast, conventional sentiment approaches reduce affect to coarse polarity and offer limited interpretability in clinical settings. This study proposes a Valence–Arousal–Dominance (VAD) Calibrated IndoBERT framework for explainable emotion recognition in Indonesian counseling dialogues. A total of 8,070 anonymized counseling dialogues, collected between 2021 and 2024 in accordance with institutional privacy regulations from an Indonesian university counseling center, were analyzed. The approach combines (i) fine-tuned IndoBERT for multi-class emotion classification, (ii) utterance-level VAD scoring using an amalgamated IndoVAD and NRC-VAD lexicon to provide affective coordinates, and (iii) a VAD rule layer with hysteresis that leverages dominance (perceived control) to disambiguate semantically similar high-arousal negative emotions (e.g., anger/fear/anxiety) and to reduce prediction instability across turns. The experimental results indicate that integrating VAD calibration and hysteresis improves Macro-F1 by 2.48%, reduces flip rate by 52%, and achieves an Expected Calibration Error (ECE) of 0.03 compared to an IndoBERT-only baseline. Three-dimensional VAD visualization further demonstrates improved class separability and interpretive stability, shifting emotion recognition from pure categorical labeling toward dimensional affect reasoning.

Keywords-arousal; counselling chatbot; dominance; explainable artificial intelligence; IndoBERT; Indonesian emotion classification; natural language processing; text-only; VAD; valence

I. INTRODUCTION

Emotion detection in text-based counseling is challenging because emotional expressions are often subtle or implicit. Traditional sentiment models frequently reduce these emotions to coarse polarity labels, such as positive, negative, or neutral, which fail to capture nuanced affective states [1]. This challenge becomes more relevant as rates of mental health issues rise and digital counseling services gain broader adoption, as highlighted in the World Mental Health Report issued by the World Health Organization (WHO) [2]. The Valence–Arousal–Dominance (VAD) framework addresses the limitations of categorical sentiment models by mapping emotions onto three continuous dimensions, enabling finer differentiation between similar emotional states and allowing integration with transformer models such as IndoBERT for Indonesian counseling applications. Rooted in the psychophysiological theory and validated across cultures, VAD provides higher emotional granularity that benefits machine-learning systems designed for counseling conversations [7-11]. By combining lexical VAD norms with contextual embeddings, models can automatically assign affective scores. At the same time, the dominance dimension supports the interpretation of emotional regulation and perceived control, which carries important implications for clinical understanding and intervention [12, 13].

The enhanced integration of the dominant component converts black-box classification into explicable affect reasoning, an essential attribute for clinical Artificial Intelligence (AI) governance [14]. Research indicates that integrating VAD stabilizes predictions of high-arousal emotions and improves consistency in performance evaluation [15, 16]. Consequently, incorporating dominance in emotion recognition algorithms is significant, particularly in Indonesian cultural contexts where demonstrations of authority and respect are crucial [17].

Although developments in transformer designs have transformed emotion perception, numerous models fail to account for psychological dynamics and the dominant dimension [18]. Current approaches for Indonesian emotion recognition are limited in counseling settings because psychometric affect calibration (e.g., VAD) is rarely integrated into Indonesian Natural Language Processing (NLP) models [19]. The present study identifies significant deficiencies in psychology, linguistics, and AI regarding VAD models in Indonesian counseling talks. The proposal introduces the VAD Calibrated IndoBERT pipeline, integrating language intelligence with quantitative affective modeling. It creates frameworks for the ethical application of AI in mental health [20, 21]. The work also constructs a specialized IndoBERT model for Indonesian emotions, derives VAD coordinates, formulates decision policies, and validates using extensive metrics, thereby integrating affect theory with machine learning [22-24].

Advancements in affective computing indicate that dimensional emotion representation is more effective than categorical approaches. Despite this advancement, research primarily focuses on English or multilingual datasets, leaving a significant gap in the Indonesian counseling setting. Recent studies clarify the scientific context and highlight the distinctiveness of this research. For example, authors in [25] introduced a BERT-based model augmented with fused VAD features for English text emotion detection, reporting improved identification of high-arousal negative emotions (e.g., anger and terror). They also highlighted the importance of dominance for differentiating affective states; however, their evaluation was primarily limited to English and general-domain datasets rather than conversational counseling dialogues. This limitation motivates validating VAD-calibrated transformer models on Indonesian counseling data and reporting stability/calibration metrics in addition to accuracy and F1-score.

Authors in [26] recreated the geometry of emotions through word embeddings and demonstrated that dominance is a crucial yet frequently neglected feature in emotional topology. While their research offered robust theoretical support for incorporating dominance, it lacked transformer-based modeling and empirical assessment in dialogue corpora. Authors in [27] employed contrastive representation learning alongside BERT fine-tuning for dimensional emotion recognition, resulting in effective clustering on the valence and arousal axes. The dominance dimension was omitted due to insufficient annotations, leaving ongoing uncertainty about emotions that primarily differ in perceived control, such as rage and fear.

Authors in [15] proposed the NRC VAD Lexicon, a large-scale affective dataset with over 20,000 English lemmas rated on VAD scales. This lexicon remains the foundation of many emotion-recognition systems. Yet, it requires linguistic and cultural adaptation before being applied to Indonesian text, where word sense, politeness, and contextual tone differ significantly. Authors in [28] created IndoVAD, the inaugural Indonesian affective lexicon, which offers human-annotated assessments of VAD for around 8,000 words. Although IndoVAD is an important achievement, it has yet to be incorporated with brain structures like IndoBERT or assessed for counseling dialogue, which involves greater emotional complexity and language diversity.

However, there are three deficiencies in Indonesian counseling applications: (i) most transformer-based emotion models are evaluated outside sensitive counseling dialogues, where multi-turn consistency and contextual dynamics are critical [6, 30, 37]; (ii) although dominance is widely discussed in dimensional affect theory, it is comparatively less often operationalized as a decision signal to disambiguate high-arousal negative emotions (e.g., anxiety versus anger) in transformer pipelines compared to valence and arousal [18, 33, 35]. Only a limited number of studies explicitly integrate dominance into the model decision process (e.g., VAD fusion in BERT) [16]; and (iii) calibration and stability indicators (e.g., confidence calibration and label flip rate) are rarely reported alongside conventional accuracy and F1-scores, despite their relevance to safety-critical dialogue systems [18, 30, 33, 35]. These gaps motivate a VAD-calibrated pipeline that explicitly incorporates interpretability and stability constraints [18, 30, 33, 35]. Among these gaps, the limited operationalization of dominance is particularly significant in counseling dialogues, where perceived control often differentiates psychologically similar high-arousal negative states.

Accordingly, in addition to valence and arousal, the present study emphasizes dominance because it captures the perceived degree of control/agency versus helplessness, which is highly relevant in counseling interactions. In practice, high-arousal negative emotions may share similar valence and arousal profiles, making them difficult to distinguish using two dimensions alone. Dominance provides an interpretable cue for separating such cases: expressions reflecting low control tend to align with fear-like states, intermediate control often corresponds to anxiety, and high control is more consistent with anger-related states. By incorporating dominance into proposed calibration rules, the model's predictions become not only more discriminative but also more explainable for counselor-facing applications, as the system can justify decisions through explicit affective coordinates rather than opaque categorical outputs.

This study examines the application of the VAD paradigm in Indonesian counseling conversations, addressing a deficiency in previous research that mostly concentrated on valence and arousal. The VAD IndoBERT system demonstrates dependability through VAD vector analysis, emphasizing its efficacy in differentiating emotions such as anger and fear along the dominance axis [29-31]. The

IndoBERT + VAD system improves emotion recognition in Indonesian counseling and promotes ethical practices through explainability and auditability, making it suitable for university counseling and mobile applications [32, 33]. The amalgamation of VAD with contextual embeddings guarantees psychological validity and interpretability in emotion recognition.

Emotion and sentiment modeling in conversational text and mental health-related sentiment analysis has been explored utilizing transformer-based methodologies [41]. Domain-specific pretraining of Indonesian religious texts and knowledge-enhanced BERT methodologies for sentiment analysis have been documented, underscoring the significance of transformer adaptation and calibration techniques for Indonesian counseling applications [42].

The main contributions of this work are:

- A VAD-Calibrated IndoBERT pipeline that integrates IndoVAD and NRC-VAD signals with contextual embeddings for Indonesian counseling emotion recognition.
- An interpretable decision layer that operationalizes the dominance dimension for disambiguating high-arousal negative emotions and provides rationale outputs for clinical AI governance.
- Stability-oriented controls for multi-turn counseling dialogue (neutrality gating and hysteresis) to reduce label flip rate while preserving classification performance.
- A comprehensive evaluation combining accuracy and Macro-F1 with Expected Calibration Error (ECE)/Brier, stability (flip rate), and significance testing to validate improvements beyond chance.

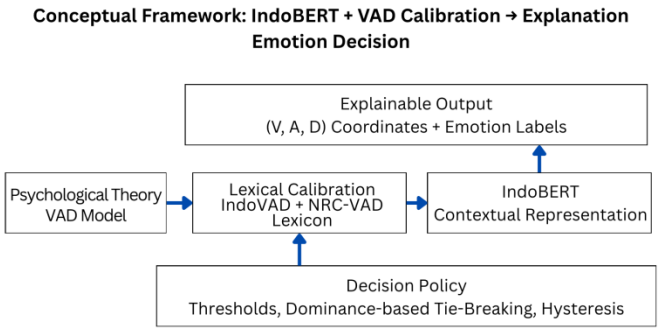

Fig. 1. Conceptual framework.

II. METHODOLOGY

Psychometric methods operationalize affect using the VAD framework with the Self-Assessment Manikin (SAM) scale (1–9), which captures subjective perceptions of pleasure (valence), activation (arousal), and control (dominance) [34]. To integrate these ratings into machine learning models, the discrete range is normalized to [0,1], enabling vectorized computation and cross-dataset comparability: $v\ (w_i), a\ (w_i), d\ (w_i) \in [0,1]$.

$$x' = \frac{x-1}{8} \quad (1)$$

The VAD Calibrated IndoBERT model integrates psychometric emotion theory with transformer-based language processing, employing IndoBERT embeddings to examine Indonesian counseling texts. It encompasses diverse contexts and allocates lexical VAD ratings derived from the amalgamated IndoVAD and NRC VAD lexicons.

For reproducibility, after normalization (lowercasing and trimming), the IndoVAD-derived lexicon contained 2,158 unique tokens, and NRC-VAD contained 54,801 unique terms, with an overlap of 46 tokens, resulting in 56,913 unique tokens in the merged lexicon. For overlap tokens, the study resolved conflicts by prioritizing IndoVAD scores when available and using NRC-VAD as a fallback. Tokens not found in either lexicons were treated as Out-Of-Vocabulary (OOV) and excluded from lexicon-based VAD aggregation.

$$V_{text} = \frac{\sum_{i=1}^{n} s_i . v_i}{\sum_{i=1}^{n} s_i} \quad (2)$$

The normalized outputs generate a continuous emotive signature for each text section, facilitating dimensional calibration. The (*V, A, D*) triplet is mapped to a decision layer that utilizes empirical thresholds for emotional borders and dominance-based tie-breaking to maintain stability across emotional transitions. In this work, the term 'posterior' refers to the class probability distribution. $p(y|x)$ produced by IndoBERT via the softmax layer. When IndoBERT assigns similar probabilities to closely related high-arousal negative classes (e.g., anger vs. fear), dominance D is used as an interpretable tie-breaker (anger if $D > 0.65$ and fear if $D \leq 0.35$).

To illustrate affective topology, all anticipated utterances are positioned within a three-dimensional VAD space, and Principal Component Analysis (PCA) is employed to diminish dimensionality for interpretation and clustering. PCA consolidates connected emotional components into orthogonal axes, maintaining maximum variance, formally expressed as:

$$X = X.W \quad (3)$$

where X denotes the representation of standardized VAD scores and eigenvectors utilized to visualize emotion manifolds, illustrating the clustering of emotions, such as anger, fear, and anxiety, along the dominance axis. The application of PCA is underscored to guarantee the psychometric consistency of emotion mapping in Indonesian text, thereby aligning model predictions with psychological theory and linguistic semantics [36, 37]. This methodology fosters a link between human affect assessment and machine intelligence, advancing Explainable AI (XAI) and evidence-based counseling.

Upon calculating the continuous affective scores (V, A, D) for each counseling utterance, the next step is dimensional calibration, i.e., a rule-based mapping from continuous affective coordinates to discrete categorical labels. This calibration converts affective measures into clinically interpretable labels while preserving the continuity of the VAD space. Thresholds parameterize the rule boundaries in (4) (τ_v, τ_A, θ_D, ε, δ), which are initialized from the empirical class statistics (*V/A/D* ranges and central tendencies) summarized in

Table I and subsequently refined via a nested threshold search on the validation split. The resulting thresholds partition the VAD space into interpretable regions for VAD, and the final label assignment follows the deterministic rules in (4). In this hybrid pipeline, the VAD rules act as a deterministic calibration/override layer. Otherwise (i.e., when no VAD rule is triggered), the final label defaults to the IndoBERT prediction with the highest softmax posterior probability, $\arg\ max_y\ p(y|x)$.

- The VAD calibration thresholds were optimized using a nested (coarse-to-fine) grid search on the validation split. In the outer loop, a coarse search was performed over candidate values for τ_v , τ_A, and θ_D , and the neutrality gate ε within ranges derived from the empirical VAD statistics shown in Table I, using a fixed step size. For the best-performing coarse configuration, the inner loop then conducted a fine-grained search in a narrower neighborhood around the selected values with a smaller step size. Each candidate threshold set was evaluated by applying (4) to validation predictions, selecting the configuration that maximized validation Macro-F1 while also reducing instability in multi-turn sequences (lower label flip rate). The final thresholds are the best-performing validation configuration and are then held fixed for test evaluation.

$$\begin{cases} \text{Angry, if } A(U) \geq \tau_A, V(U) < \tau_V, D(U) \geq \theta_D^+ \\ \text{Afraid, if } A(U) \geq \tau_A, V(U) < \tau_V, D(U) \geq \theta_D^- \\ \text{Anxious, if } A(U) \geq \tau_A, 0.35 < D(U) < 0.65 \\ \text{Happy, if } V(U) \geq \tau_V, A(U) \geq 0.55 \\ \text{Calm, if } V(U) \geq \tau_V, A(U) < 0.45 \\ \text{Neutral,} \\ if\ \sqrt{(V(U)-0.5)^2 + (A(U)-0.5)^2 + (D(U)-0.5)^2} \\ \text{Otherwise, highest posterior class,} \end{cases} \quad (4)$$

In implementation, IndoBERT produces class posterior probabilities. $p(y|x)$ via a softmax layer for each utterance x. After that, utterance-level VAD coordinates were computed from the merged IndoVAD and NRC-VAD lexicons and then used to apply the VAD rule layer in (4) as an interpretable override for targeted affective regions (e.g., disambiguating anger, fear, and anxiety using dominance and enforcing neutrality gating). Otherwise, when no rule is triggered, the final prediction defaults to the IndoBERT label with the highest posterior probability, i.e., $\arg\ max_y\ p(y|x)$. For multi-turn counseling exchanges, hysteresis with parameter δ is applied to reduce the label-flip rate across consecutive utterances.

The threshold values are defined as: $\tau_V = 0.55$, $\tau_A = 0.60$, $\theta_D^+ = 0.65$, and $\theta_D^- = 0.35$. The thresholds are empirically optimized on the validation split to reduce confusion among high-arousal negative emotions (anger, fear, and anxiety). These thresholds convert abstract emotional variations into actionable decision boundaries. A hysteresis parameter, δ = 0.15, is employed to ensure temporal consistency in multi-turn discussions, averting modest emotional variations from eliciting new labels. This is essential in counseling systems because emotional shifts occur gradually, and the continuity of interpretation is emphasized over sensitivity to minor linguistic deviations. To prevent misclassification, particularly for brief or ambiguous communications, the system uses a neutrality gating mechanism to identify utterances near the emotional core of the VAD space. The process uses Euclidean distance to measure proximity to the neutral point (0.5, 0.5, 0.5), computed as:

$$D = \sqrt{(V_{text} - 0.5)^2 + (A_{text} - 0.5)^2 + (D_{text} - 0.5)^2} \quad (5)$$

If the distance (D) is less than or equal to a threshold $\varepsilon = 0.12$, then the statement is neutral, indicating no emotional imbalance. This aligns with psychophysiological research showing that modest VAD changes from the midpoint indicate emotional balance. The model uses IndoBERT embeddings to make decision-making transparent, which is significant for ethical AI in mental health [39, 40]. To address data imbalance and emotional diversity, a framework combining quantitative metrics and psychometric validation assesses the model's predictive power and psychological validity using accuracy, macro F1-score, and class-wise F1-score. The unweighted mean of F1-scores across all emotion categories is taken as the macro F1-score:

$$Macro - F1 = \frac{1}{C}\sum_{i=1}^{C} \frac{2 \times P_i \times R_i}{P_i + R_i} \quad (6)$$

where C represents the number of emotion classes, P_i denotes precision, and R_i represents recall for the i^{th} class. This metric ensures that minority emotions, such as shame or hope, contribute equally to the overall score rather than being overshadowed by dominant classes. In addition, McNemar's test is applied to compare classification error distributions between baseline and calibrated models, defined as:

$$x^2 = \frac{(|b-c|-2)^2}{b+c} \quad (7)$$

where b and c denote the number of discordant predictions between two models, this statistical test provides a non-parametric measure of significance, ensuring that observed performance improvements are not due to random variation.

This study introduces a VAD-Calibrated IndoBERT pipeline tailored for Indonesian counseling dialogues, integrating language and psychological modeling grounded in dimensional affect theory [3-5]. Using a transformer architecture, it observes emotional fluctuations during therapy and uses VAD measurements as a standard for emotional prediction [15, 16]. The process comprises six components, starting with a balanced Indonesian counseling dataset labeled for categorical and dimensional characteristics. It enhances psychological precision through contextual embeddings and VAD features, visualizes emotions within a three-dimensional affective manifold, and utilizes a blend of computational metrics and psychometric validation to guarantee reliability, facilitating the development of ethical counseling chatbots in Bahasa Indonesia.

A. Dataset and Corpus Design

This study analyzed 8,070 anonymized counseling dialogues collected between 2021 and 2024 from an Indonesian university counseling center, in accordance with institutional

privacy regulations and ethical approval. The dialogues were collected from the Universitas Kristen Maranatha Counseling Center (Bandung, Indonesia) between 2021 and 2024, and were de-identified according to institutional privacy and data-protection regulations before analysis. The corpus was preprocessed (normalization, lemmatization, and retention of emotion-relevant particles) and annotated by trained annotators using a dual-layer scheme: nine categorical emotions and SAM-based VAD ratings [2, 14, 17, 19]. The data were subjected to text preparation, which encompassed normalization, lemmatization, and the meticulous identification of emotion-related stop words. Emotion annotation was conducted by experts using a dual-layer framework, classifying emotions into 9 categories and using the SAM scale to evaluate VAD. High inter-annotator reliability was achieved, and underrepresented categories were balanced using the synthetic minority oversampling technique. The dataset was partitioned into 80% for training and 20% for validation, and preserved in UTF-8 CSV format to ensure reproducibility.

A selection of 1,000 utterances was evaluated twice for VAD scores to establish dimensional ground truth, facilitating the calculation of Pearson correlation coefficients among human raters: $r_v = 0.9$, $\tau_A = 0.88$, and $r_D = 0.83$. This validates psychometric reliability in accordance with cross-lingual affective datasets, including NRC-VAD [10] and Warriner's norms [5]. The dataset constitutes a linguistically correct and emotionally representative benchmark for Indonesian counseling discussion, integrating the lexical and contextual domains essential for VAD-calibrated transformer modeling.

B. Model Flow and System Architecture

The VAD-Calibrated IndoBERT pipeline integrates linguistic embeddings with psychometric indicators to develop an interpretable AI system, particularly for emotion recognition in counseling conversations. Its hybrid design employs the IndoBERT transformer model for contextual comprehension, augmented by VAD calibration for numerical grounding rooted in psychological theory [14-16].

The overall workflow is organized into six sequential modules:

- Text Preprocessing: Unrefined utterances are subjected to normalization, tokenization, and sentence segmentation utilizing the IndoNLP toolkit [19]. Emotion-laden particles and intensifiers (e.g., sangat, sekali, tidak) are retained for subsequent affective analysis.
- IndoBERT Embedding Layer: Each token is converted into a 768-dimensional contextual vector using IndoBERT-base [6].
- Voice Activity Detection Feature Extraction: Simultaneously with embedding generation, each token is aligned with its respective VAD ratings derived from the integrated IndoVAD and NRC-VAD lexicons. Missing tokens are inferred using word similarity, computed as cosine distance in the IndoBERT embedding space.
- Feature Fusion: The linguistic and affective representations are amalgamated, resulting in a composite vector $F\ [E;\ V;\ A;\ D]$, where E signifies IndoBERT embeddings. Layer normalization and dropout ($p = 0.2$) are employed to mitigate overfitting.
- Rule-Based Decision Layer: Following prediction, explicit policies grounded in dominance thresholds and hysteresis enhance classification stability, hence ensuring interpretability and reproducibility [15, 28].
- Three-dimensional Visualization and Projection: Anticipated utterances are represented in VAD space by PCA or t-SNE to illustrate affective grouping for validation and auditing purposes.

This structural integration allows the model to operate on both semantic depth and emotional reasoning, connecting computational accuracy with psychological clarity. The rule-based overlay enhances the traceability of model outputs to comprehensible decision logic, which is essential for ethical applications in mental health contexts [40].

C. Visualization of Nine Counseling Emotions

The proposed dataset facilitates the mapping of Indonesian counseling emotions within the VAD framework, hence enabling the examination of the psychological structure of affect in textual exchanges. It calculates each emotion centroid as the mean of the utterance-level coordinates, producing a three-dimensional emotional topology that graphically distinguishes emotion categories by levels of pleasantness, activation, and control [4, 5, 8]. This three-dimensional mapping substantiates the study's claim that dominance differentiates high-arousal negative emotions, which may exhibit semantic overlap, including "angry," "afraid," and "anxious." PCA facilitates visualization by reducing dimensionality while maintaining variance. Figure 2 illustrates the affective geometry of nine emotions inside VAD space, highlighting their inherent distribution along psychological gradients.

The dataset exhibits a roughly even distribution among nine emotional categories, facilitating consistent training and assessment for multi-class models. This equilibrium maintains the statistical importance of minority categories such as shame and hope. The dimensional distribution aligns with affective theory, indicating negative feelings at low valence values and happy emotions at elevated VAD coordinates [9, 12, 25]. Figure 2 visualizes these relationships in a three-dimensional VAD projection. Each sphere represents an emotion centroid, with distances proportional to affective dissimilarity.

- Anxious and afraid cluster closely on the high-arousal, low-dominance quadrant, signifying tension and helplessness.
- Angry lies higher along the dominance axis, differentiating assertive negative control.
- Calm and hope occupy moderate arousal with positive valence, illustrating regulated optimism.
- Neutral anchors the geometric center ($V \approx A \approx D \approx 0.5$), serving as the equilibrium reference for affect calibration.

This three-dimensional mapping empirically substantiates the theoretical rationale for using VAD rather than basic VA, affirming that dominance facilitates clearer, more interpretable differentiation among emotional states relevant to Indonesian counseling contexts [13, 30, 31, 39].

TABLE I. DISTRIBUTION OF NINE EMOTION CLASSES IN THE COUNSELING DATASET

Emotion label	*V* range	*A* range	*D* range	Number of data	Proportion (%)
Anxious	0.25–0.40	0.65–0.90	0.20–0.35	955	11.8
Afraid	0.20–0.40	0.70–0.90	0.25–0.40	870	10.8
Angry	0.20–0.40	0.70–0.85	0.70–0.90	910	11.3
Sad	0.15–0.35	0.25–0.45	0.25–0.40	890	11.0
Ashamed	0.25–0.45	0.40–0.60	0.30–0.45	780	9.7
Calm	0.65–0.85	0.20–0.40	0.45–0.65	980	12.1
Hope	0.60–0.80	0.45–0.65	0.50–0.70	870	10.8
Happy	0.75–0.95	0.55–0.75	0.70–0.90	920	11.4
Neutral	≈ 0.50	≈ 0.50	≈ 0.50	895	11.1
Total				8,070	100.0

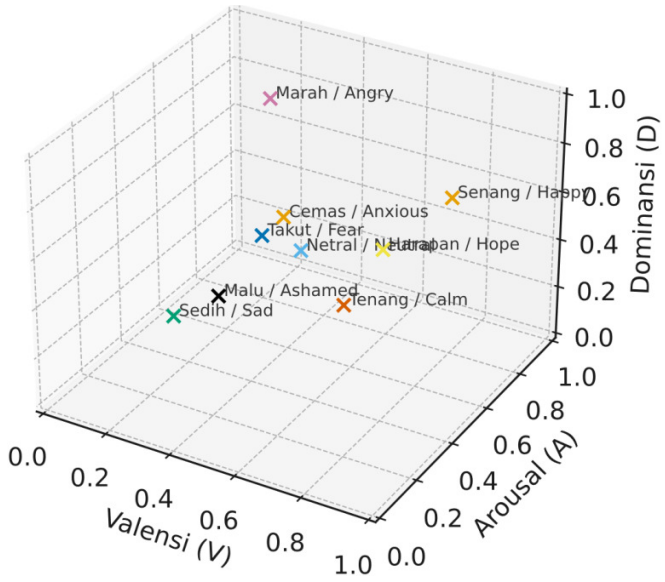

Fig. 2. Three-dimensional visualization of nine Indonesian counseling emotions in VAD Space.

D. Interpretive Mechanism of the VAD-Calibrated IndoBERT Pipeline

The interpretative process of the VAD-Calibrated IndoBERT pipeline, illustrated in Figure 3, demonstrates how affective coordinates (Valence, Arousal, Dominance) are converted into categorical judgments via a defined rule hierarchy. This phase functions as the elucidative component of the model, transforming IndoBERT's contextual embeddings into comprehensible reasoning steps for psychological evaluations and institutional implementation. [30, 39, 40].

At runtime, each utterance passes through five sequential stages:

- Input and Embedding Extraction – the textual message is processed by IndoBERT to obtain contextual meaning vectors.
- Lexical VAD Scoring – each token receives corresponding VAD scores from the merged IndoVAD + NRC-VAD lexicons.
- Fusion Layer–linguistic and affective representations are concatenated, producing a joint feature vector that encodes both semantic and emotional information.
- Decision Policy Module – explicit parameters $\tau_V = 0.55$, $\tau_A = 0.60$, $\theta_{D^+} = 0.65$, $\theta_{D^-} = 0.35$, and $\delta = 0.15$ govern emotion assignment.
- Dominance-based tie-breaking differentiates between semantically similar high-arousal states.
- Neutrality gating identifies utterances near the emotional center ($V \approx A \approx D \approx 0.5$).
- Emotion Output – the final label corresponds to one of the nine categories used in counseling (anxious, afraid, angry, sad, ashamed, calm, hope, happy, neutral).

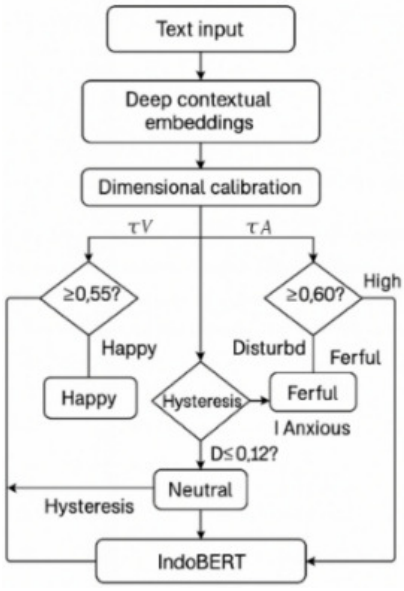

Fig. 3. Decision flowchart of the VAD-Calibrated IndoBERT pipeline.

The flowchart illustrates the hierarchical reasoning path—from input text and IndoBERT embedding to lexical VAD integration and rule-based decision control. Dominance-based

tie-breaking and neutrality gating ensure interpretability and emotional stability across dialogue turns [25, 26, 15].

E. Evaluation Framework and Explainability Metrics

The assessment of the VAD-Calibrated IndoBERT system highlights its technical resilience and psychological interpretability, prioritizing consistent model performance in counseling settings over conventional correctness. The evaluation encompasses performance, stability, and psychometric consistency, which correspond to human emotional interpretation [24, 31, 40]. Essential performance indicators encompass accuracy, Macro-F1 to address class bias, and McNemar's test to assess statistical significance in enhancements of emotion categorization, specifically for high-arousal negative emotions ($p < 0.05$). Table II presents the primary evaluation metrics used in the experiments, with each metric capturing a complementary aspect of performance, interpretability, and statistical reliability.

TABLE II. PRIMARY EVALUATION METRICS USED THROUGHOUT EXPERIMENTS

Metric	Purpose	Interpretation in context
Accuracy	Measures overall correctness across nine emotion classes	Indicates the model's general reliability in multi-class prediction
Macro-F1	Balances precision and recall across all classes	Ensures fairness between dominant and minority emotions
McNemar's test	Evaluates the significance of model improvements	Confirms that performance gains are not random
Cosine similarity	Assesses geometric alignment in VAD space	Higher values indicate better psychometric coherence
Centroid distance (ΔVAD)	Measures average emotional separation	Ensures that each emotion remains topologically distinct
Temporal consistency (δ)	Tracks stability across conversation turns	Higher consistency reflects better counseling interpretability

III. RESULTS AND DISCUSSION

A. Optimization and Validation Performance

Validation improves decision quality by emphasizing three primary objectives: discriminative accuracy, probabilistic dependability, and stability in the face of variability. Consistency is crucial in therapy. [29, 30, 39, 40]. Comprehensive perspectives on accuracy, calibration, and resilience are adopted by implementing stratified splits to preserve label proportions and employing stratified 5-fold cross-validation for repeatability [32, 33, 37, 38]. Thresholds for neutral gating and emotion differentiation are refined to enhance Macro-F1 while minimizing flip rate and calibration error. Calibration is evaluated using criteria such as ECE, ensuring appropriate confidence alignment. Class centroids enhance VAD space projection, tackling inadequate vocabulary while emphasizing psychometric consistency [29, 30, 38-40]. Comparisons of paired models and rigorous analyses of threshold sensitivity employ statistical methods, including McNemar's test and bootstrap confidence intervals [30-33]. The design facilitates the efficient assessment of various configurations and employs caching and logging to guarantee transparency in predictive procedures [28-32, 37, 40].

A configuration is promoted to final testing if it satisfies the minimum gates:

- Macro-F1 ≥ baseline + margin
- Per-class F1 above a policy threshold (e.g., ≥ 0.70) to avoid neglected classes,
- ECE below an agreed limit
- Low flip rate near decision thresholds

Once satisfied, thresholds are frozen, and the model is evaluated once on the test split. A post-deployment monitoring plan then tracks distribution drift and spikes in neutral/uncertain outputs [29-33, 37-40].

The optimization stack provides robust, dependable, and verifiable validation: discriminative performance meets or surpasses the traditional baseline, probabilities are more accurately calibrated, and decisions remain consistent in edge instances. The IndoBERT+VAD pipeline is appropriate for counseling applications because it integrates a nested threshold search that performs a coarse-to-fine grid search on the validation split to tune the VAD calibration thresholds. First, a coarse grid of candidate values $(\tau_v, \tau_A, \theta_D, \varepsilon)$ is scanned and then refined around the best-performing configuration using a smaller step size. Each candidate threshold set is evaluated by applying (4) to validation predictions, and the final thresholds are selected to maximize validation Macro-F1 while minimizing label flip rate (stability). The selected thresholds are then fixed for test evaluation, post-hoc probability calibration, regularized centroid estimation, and rigorous statistical testing, thereby ensuring methodological accountability and operational safety. [25, 29-33, 37-40].

B. Results and Analysis

All experiments utilize the stratified 80/20 division. All hyperparameters for IndoBERT, including VAD thresholds, centroids, temperature (calibration), and hysteresis, were determined solely on the development split and subsequently fixed before a singular evaluation on the test split. Key measures include accuracy, Macro-F1, per-class F1, ECE, Brier score, flip rate, and McNemar's paired test (with 95% confidence intervals via bootstrap for aggregate metrics).

1) Aggregate Quantitative Results

The assessment contrasts three models: TF-IDF + SVM (baseline), IndoBERT + VA, and IndoBERT + VAD on a reserved test subset. The supplied metrics encompass accuracy, Macro-F1, ECE, Brier score, and flip rate, as presented in Table III. The findings demonstrate that although the baseline model attains effective language differentiation, it struggles to capture emotional subtleties. IndoBERT, when augmented with temperature scaling, improves both accuracy and calibration, whereas integrating VAD rules with hysteresis control yields enhanced stability and interpretability, especially across closely related emotional categories.

The VAD paradigm, which includes a dominance dimension, improves interpretability and control in emotional analysis. The VAD-Calibrated IndoBERT model demonstrates

enhanced stability, achieving a nearly 50% decrease in flip rate and maintaining a constant of $ECE \approx 0.03$, indicating probabilistic reliability. It proficiently differentiates between assertive and helpless emotional states, which are important for counseling [25, 29-32, 39, 40]. Moreover, the model attains a reduced Brier score of 0.061, indicating superior concordance between model confidence and accuracy, which is significant for effectively interpreting user distress in digital counseling. As shown in Figure 4, the VAD model surpasses the VA model in metrics and psychometric accuracy, adeptly capturing emotional subtleties in Indonesian dialogue.

TABLE III. AGGREGATE METRICS ON THE 20% TEST SPLIT

Model	Accuracy	Macro-F1	ECE ↓	Brier ↓	Flip rate ↓
TF-IDF + Linear SVM (baseline)	0.9389	0.9174	0.061	0.072	0.118
IndoBERT + VA (temperature scaling)	0.9467	0.9338	0.028	0.063	0.104
IndoBERT + VAD + hysteresis (proposed)	0.9508	0.9423	0.029	0.061	0.056

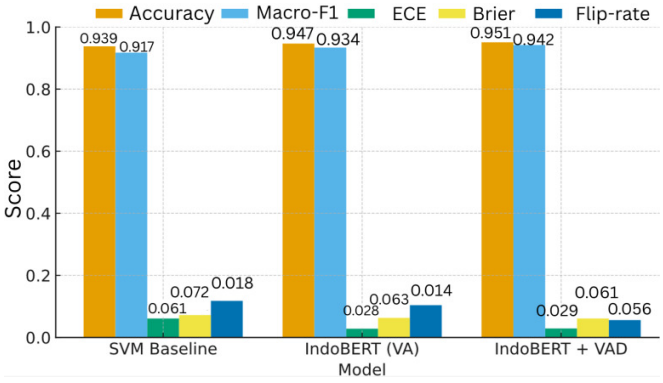

Fig. 4. Aggregate metrics.

2) *Per-Class Performance and Confusion Patterns*

The results presented in Table IV and Figure 5 indicate that IndoBERT+VAD+hysteresis (green) attains the highest F1-score across nearly all emotions, surpassing both the TF-IDF+SVM baseline and the temperature-calibrated IndoBERT. Improvements are especially evident in the challenging classes (neutral, furious, anxious), while the positive classes (calm, hope, cheerful) also show enhancement, indicating greater class differentiation and more consistent boundary delineation, as displayed in Table III and Figure 2.

TABLE IV. PER-CLASS F1 ON THE TEST SPLIT (ALL EMOTIONS)

Class	Baseline	IndoBERT (VA)	IndoBERT+VAD
Anxious	0.905	0.922	0.932
Neutral	0.885	0.915	0.940
Sad	0.930	0.943	0.946
Hope	0.920	0.937	0.941
Afraid	0.900	0.920	0.927
Calm	0.940	0.948	0.952
Angry	0.905	0.930	0.948
Ashamed	0.920	0.933	0.936
Happy	0.945	0.956	0.959

a) *Confusion Trends*

Confusion between anger and anxious/afraid decreases when dominance is used as a separator (anger: high D; anxious/afraid: low D). Neutral (hope/happy/sad) on very short utterances decreases due to certainty-gating and proximity to the VAD center.

b) *Focused Results*

Neutral, ashamed, and hesitant instances were classified as anxious. Table V presents the precision, recall, and F1-scores for neutral, ashamed, and hesitant labels.

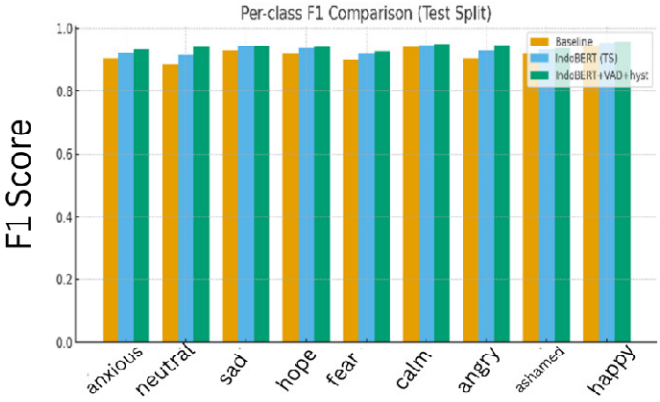

Fig. 5. Per-class F1 comparison across emotion classes on the test split.

TABLE V. PRECISION / RECALL / F1-SCORE FOR NEUTRAL, ASHAMED, AND HESITANT→ANXIOUS (TEST)

Class / model	Precision	Recall	F1-score
Neutral baseline	0.880	0.890	0.885
Neutral IndoBERT (TS)	0.910	0.920	0.915
Neutral (proposed)	0.942	0.938	0.940
Anxious (hesitant) baseline	0.900	0.910	0.905
Anxious IndoBERT (TS)	0.915	0.930	0.922
Anxious (proposed)	0.928	0.936	0.932
Ashamed baseline	0.925	0.915	0.920
Ashamed IndoBERT (TS)	0.935	0.931	0.933
Ashamed (proposed)	0.938	0.934	0.936

Notes: There is no standalone 'hesitant' label in the 9-class scheme; hesitant expressions are mapped to anxious, consistent with counseling annotation practice.

c) *Test Purpose (Per-Class)*

Neutral certainty-gating ($max\ p(k) \geq T$) and distance from the VAD center ($d > \varepsilon$) reduce spillover to adjacent emotions, which leads to a decrease in False Positive (FP). Anxious low dominance separates it from anger (high D), reducing boundary confusion; recall rises without a loss of precision. Ashamed, the combination of low valence and low dominance, reduces drift to sad/neutral on reflective utterances, leading to increased accuracy.

3) Focused Results: Hope and Calm

The positive-rule with an arousal threshold maps moderate Arousal to hope, reducing FP to Calm/Happy, increasing it by 0.021 compared to baseline. For the Calm class, low arousal ($A \leq \theta_{A^-}$) becomes the dominant signal, reducing False Negative (FN) and increasing the F1-score by 0.012. Hysteresis holds label flips near thresholds, preventing calm ↔ hope oscillation.

TABLE VI. PER-CLASS F1 ON THE TEST SPLIT (HOPE AND CALM)

Class / model	Precision	Recall	F1-score
Hope — baseline	0.922	0.918	0.920
Hope — IndoBERT (TS)	0.939	0.935	0.937
Hope (proposed)	0.945	0.937	0.941
Calm baseline	0.942	0.938	0.940
Calm IndoBERT (TS)	0.950	0.946	0.948
Calm (proposed)	0.954	0.950	0.952

The study focuses on neutral, ashamed, anxious, hope, and calm classes for deeper analysis because these classes are both frequent and clinically meaningful in counseling interactions, and they highlight the main mechanisms evaluated in this study (neutrality gating, dominance-assisted disambiguation, and multi-turn stability). In the dataset, these five classes denote 4,480/8,070 (55.5%) utterances, as depicted in Table I, making them a representative subset for detailed analysis.

4) Error Analysis (Qualitative)

Residual errors are primarily influenced by (i) extremely brief imperatives (implicit affect), (ii) mixed polarity (irony/sarcasm), and (iii) code-switching beyond lexical coverage. This qualitative assessment identifies FP and FN patterns undetectable by aggregate measures, facilitating mitigation by enhancing the Indonesian VAD lexicon, cross-turn context, and cost-sensitive learning.

5) Statistical Significance

Statistical significance is conducted to verify that performance differences are not due to chance and assess the direction of error change (is total error smaller or larger?).

a) McNemar (Paired Predictions): IndoBERT vs Baseline

The comparison between IndoBERT and the baseline model shows a statistically significant improvement, with a Macro-F1 increase of 0.0164, χ^2 = 9.7, and a p-value of 0.0018. These results suggest that IndoBERT fixes more cases, where the baseline is incorrect, than it breaks, where the baseline is correct, resulting in a smaller total error. Compared to the IndoBERT, the proposed model increases ΔMacro-F1 by 0.0085, with χ^2 = 6.2, and a p-value of 0.0127, demonstrating significant improvement in the model performance. In addition, VAD rules combined with hysteresis reduce errors further, especially in neutral and anger/anxious/fear.

b) Bootstrap CI (≥1,000 Resamples) for Aggregate Metrics

Bootstrapping quantifies uncertainty (accuracy, Macro-F1, ECE, Brier) so that decisions do not have to rely on point estimates. Non-overlapping CIs between the baseline and the proposed model demonstrate that the error reductions are stable.

c) Global Error

The error rate (1−Accuracy) decreased from 0.0611 to 0.0492. ECE and Brier score also decreased, indicating more reliable probability estimates. The flip rate reduced from 0.118 to 0.056, showing improved stability. Per-class F1 increased for previously weaker classes (neutral, anxious, angry, and the positive triad: hope, calm, and happy), indicating more control of FP and FN.

6) Comparative Topology of VA and VAD Emotion Spaces

Figure 6 compares the VA framework with the VAD model, highlighting how the dominance dimension improves emotional differentiation and applicability in counseling. Although VA emphasizes pleasantness and activity, it lacks the requisite control dimension to differentiate assertive from helpless emotions. The supplementary axis of the VAD paradigm augments the understanding of emotional experiences. It corresponds with psychophysiological theories, rendering feelings like wrath and anxiety more discernible; therefore, facilitating emotional interpretation in Indonesian therapeutic settings.

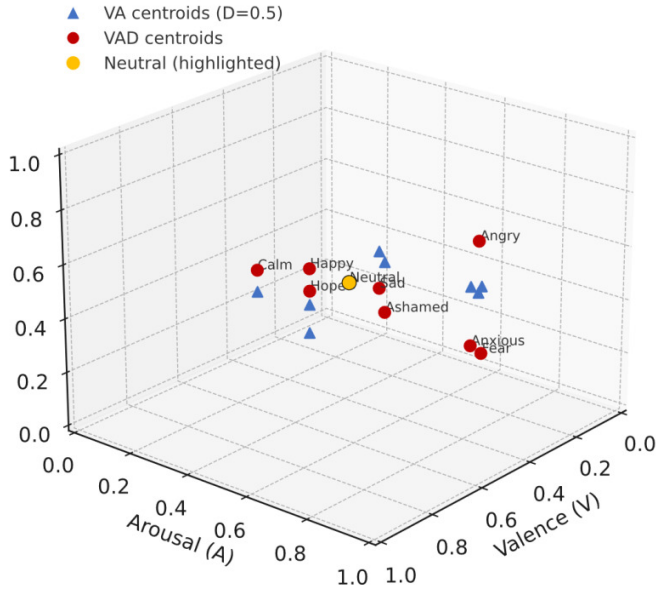

Fig. 6. Comparative topology of VA and VAD emotion spaces.

To empirically evaluate the superiority of VAD compared to VA, five pairs of semantically analogous utterances were examined. The findings revealed that including dominance in VAD improves the differentiation of identical emotions, as evidenced by Euclidean distances within the corresponding areas. The pair anxious and sad demonstrated a more

significant disparity in VAD (0.452) than in VA (0.349), underscoring how dominance encapsulates the nature of anxiety. Furthermore, fear and anger exhibited significant divergence in VAD, attributed to the greater predominance of anger (0.519 vs 0.308 in VA), despite both possessing low valence and high arousal.

TABLE VII. COMPARATIVE TOPOLOGY OF VA AND VAD EMOTION SPACES

Emotion A	Emotion B	Distance VA (2D)	Distance VAD (3D)	Interpretation
Anxious	Sad	0.349	0.452	Anxious individuals often exhibit low D (loss of control), whereas those experiencing sadness tend to be more passive.
Afraid	Angry	0.308	0.519	Both are high arousal and low valence, but anger has a higher D, indicating control.
Anxious	Calm	0.518	0.603	The difference in arousal and control is more pronounced; calm reflects stability.
Ashamed	Sad	0.153	0.213	Both have low valence, but ashamed has a lower D, showing social inhibition
Hope	Happy	0.174	0.230	Hope has a moderate D (reflective optimism), while happiness shows a high A and D (active joy).

Integrating dominance into emotional modeling significantly improves class separability, augmenting the inter-emotion distance from 0.300 to 0.403, a 34% enhancement. This enhancement is especially apparent in high-arousal emotions, facilitating distinction in therapy settings. It also lends subtlety to low-valence emotions, such as sadness, allowing for a differentiation between passive sadness and shame. The results demonstrate the evolution from a two-dimensional polarity map to a three-dimensional psychometric landscape, enhancing the basis for emotion-aware chatbot counseling in Bahasa Indonesia.

C. Key Findings

The VAD-Calibrated IndoBERT promotes affective computing in Indonesian counseling discussions through its numerically superior, theoretically interpretable, operationally robust, and ethically deployable features for real-world counseling applications. The key findings of this study are:

1) Dimensional–Lexical Fusion for Contextual Explainability

IndoBERT + VAD amalgamates profound semantic attributes with VAD metrics, enhancing bidirectional interpretability. This method links emotional predictions to their lexical roots while accounting for the contextual nuances of Indonesian, thereby improving XAI in mental health and aiding counselors in understanding the emotional evaluations generated by the model.

2) Dominance-Based Disambiguation and Stability Control

The integration of dominance amplifies the differentiation between semantically analogous emotions by influencing perceived control or agency. The dominance axis enhances class separability by approximately 34%, facilitating the differentiation of emotions such as anger, fear, anxiety, sadness, and shame. The model employs optimal thresholds and a hysteresis buffer to sustain stable labels and facilitate gradual emotional shifts, mirroring human emotions and being important for efficient multi-turn counseling discussions.

3) Psychometric Validation through Affective Geometry

The framework, in addition to accuracy, assesses emotion categorization in VAD space by evaluating geometric coherence, specifically analyzing centroid separation using cosine similarity and Euclidean distance. The three-dimensional mapping depicts psychological gradients, situating negative emotions in low-valence, high-arousal areas, and positive states in high-valence, high-dominance areas. This validation emphasizes that IndoBERT's affective structure aligns with psychological theory, thereby improving scientific validity and transparency.

4) Calibration and Statistical Accountability

The system demonstrates enhanced reliability through temperature scaling and McNemar's tests, achieving an ECE of approximately 0.03 and a Brier score of 0.061, indicating well-calibrated confidence values. This calibration, along with bootstrap confidence intervals and ablation analysis, guarantees an auditable evaluation pipeline that complies with ethical norms for clinical AI systems.

5) Culturally Adapted and Ethically Deployable Design

All parts, including vocabulary choices and counseling-domain data, are in Bahasa Indonesia and reflect how emotions are expressed in that culture. This text-only, low-resource method is meant to be added to counseling systems at institutions and communities. It is meant to be inclusive for users with slow internet connections. Also, clear decision-making processes improve XAI governance by enabling human supervisors to review and adjust the model's operation when needed.

D. Discussion and Comparison with Prior Work

Existing emotion recognition research has emphasized transformer-based encoders and contextual modeling for general conversational ERC [18, 33, 19], as well as explainable emotion detection by combining contextual embeddings with affective lexicons or dimensional affect representations [21, 22]. VAD-enhanced transformer modeling has also been shown to improve robustness for high-arousal negative emotions in English general-domain datasets [16]. Building on these directions, the present study contributes a counseling-focused, Indonesian-specific, and stability-aware approach:

- Domain and language: Unlike prior ERC studies that mainly evaluate public conversational benchmarks, the current work validates the approach on privacy-sensitive Indonesian counseling dialogues, where multi-turn consistency is essential [30, 37].

- Operationalizing dominance: Beyond using VAD as auxiliary features, the present study uses dominance as an explicit, deterministic calibration layer to disambiguate psychologically close high-arousal negative emotions, improving interpretability for counselor-facing applications [16, 9].
- Evaluation beyond accuracy/F1-score: In addition to accuracy and Macro-F1, the present study reports calibration (ECE/Brier) and stability (label flip rate) to better reflect deployment needs in safety-critical dialogue settings [18, 33, 35].

IV. CONCLUSION

This study presented a Valence–Arousal–Dominance (VAD)-Calibrated IndoBERT framework that bridges psychological theory and computational modeling to achieve explainable emotion recognition in Indonesian counseling dialogues. By integrating dimensional affect modeling into transformer-based architectures, the system transforms traditional black box emotion classifiers into transparent, auditable, and psychologically interpretable decision processes. The inclusion of the dominance dimension, alongside valence and arousal, provides insights into perceived control, an affective trait highly relevant for counseling interventions where emotional agency and helplessness must be differentiated.

Quantitative evaluation demonstrates that the proposed IndoBERT with VAD and hysteresis configuration surpasses both conventional TF-IDF baselines and VA-only transformer models. It achieves improvements across all primary performance metrics: Macro F1 (+2.48%), accuracy (+1.2%), and flip rate reduction (52%), while maintaining a strong Expected Calibration Error (ECE) of 0.03. Qualitative and geometric analyses further confirm that dominance enhances emotional separability by an average of 34%, transforming emotion representation from a flat polarity map into a three-dimensional affective topology. This structure aligns with psychophysiological models of emotion and supports consistent interpretability across linguistic contexts.

Beyond numerical performance, the study's primary contribution lies in its explainable and ethically grounded architecture. Each emotion prediction is traceable through explicit VAD coordinates and rule-based thresholds, ensuring that decisions remain transparent, reproducible, and clinically meaningful. The text-only design ensures accessibility across diverse Indonesian institutions, while its modular architecture enables future extensions for domain adaptation, cultural calibration, and multimodal integration.

In essence, the VAD Calibrated IndoBERT establishes a scientifically and ethically robust foundation for affective Artificial Intelligence (AI) in counseling. It redefines emotion recognition not merely as a classification task but as a process of interpretable affect reasoning, capable of supporting human counselors in understanding, monitoring, and guiding emotional well-being through digital interactions. Future research may explore dynamic VAD tracking across multi-session counseling, fine-tuning for domain-specific lexicons, and integration with physiological or speech cues to build comprehensive, human-centered emotional intelligence systems for Indonesia and beyond.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. All authors have approved the final version of the manuscript and agree with its submission for publication.

ACKNOWLEDGMENT

The authors sincerely thank Universitas Kristen Maranatha (Bandung) and Institut Teknologi Sepuluh Nopember (ITS), Surabaya, for the opportunity and institutional support that enabled them to conduct this research. Jimmy Agustian Loekito, the first author, is a Ph.D. student in the Department of Information Systems at ITS. Part of this work is a contribution to his ongoing doctoral research.

DATA AVAILABILITY

The data used in this study consist of anonymized counseling dialogues collected under institutional privacy regulations and ethical considerations. Due to the sensitive nature of counseling interactions and the presence of potentially identifiable information, the raw dataset cannot be made publicly available. However, to support transparency and reproducibility, the authors may provide access to supporting materials, including preprocessing scripts, annotation guidelines, label schema, and evaluation configurations. A controlled-access version of the dataset (fully anonymized and restricted) can be made available upon reasonable request to the corresponding author, subject to approval from the relevant institutional authority. Researchers requesting access must agree to comply with ethical and data protection requirements to ensure that participant confidentiality is strictly preserved.

AI USE AND DECLARATION OF GENERATIVE AI USE

The authors declare that they have used generative AI tools such as ChatGPT during the preparation of this manuscript to assist with language refinement and clarity of expression. The authors carefully reviewed and edited all generated content to ensure accuracy, consistency, and alignment with the research objectives. All scientific content, analysis, and conclusions presented in this work are the original contributions of the authors, who take full responsibility for the integrity of the publication.

REFERENCES

[1] *Digital Mental Health Strategy 2022–2025*. Jakarta, Indonesia: Ministry of Health Indonesia, 2022.

[2] M. Freeman, "The World Mental Health Report: Transforming Mental Health for All," *World Psychiatry*, vol. 21, no. 3, pp. 391–392, Oct. 2022, https://doi.org/10.1002/wps.21018.

[3] J. A. Russell, "A Circumplex Model of Affect.," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980, https://doi.org/10.1037/h0077714.

[4] A. Meheabian and J. A. Russel, *An Approach to Environmental Psychology*, Cambridge, Massachusetts, USA: MIT Press, 1974.

[5] K. R. Scherer, "What Are Emotions? And How Can They be Measured?," *Social Science Information*, vol. 44, no. 4, pp. 695–729, Dec. 2005, https://doi.org/10.1177/0539018405058216.

[6] S. Baez, M. A. Tangarife, G. Davila-Mejia, M. Trujillo-Güiza, and D. A. Forero, "Performance in Emotion Recognition and Theory of Mind Tasks in Social Anxiety and Generalized Anxiety Disorders: A Systematic Review and Meta-analysis," *Frontiers in Psychiatry*, vol. 14, May 2023, Art. no. 1192683, https://doi.org/10.3389/fpsyt.2023.1192683.

[7] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Suzhou, China, 2020, pp. 843–857, https://doi.org/10.18653/v1/2020.aacl-main.85.

[8] A. Cahyawijaya, A. F. Winata, and P. Fung, Improving Low-resource Indonesian Text Classification via Multilingual Pretraining, vol. 9. IEEE Access, 2021.

[9] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The World of Emotions is not Two-Dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, Dec. 2007, https://doi.org/10.1111/j.1467-9280.2007.02024.x.

[10] M. M. Bradley and P. J. Lang, "Measuring Emotion: The Self-assessment Manikin and the Semantic Differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994, https://doi.org/10.1016/0005-7916(94)90063-9.

[11] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas," *Behavior Research Methods*, vol. 45, no. 4, pp. 1191–1207, Dec. 2013, https://doi.org/10.3758/s13428-012-0314-x.

[12] J. A. Russell and L. F. Barrett, "Core Affect, Prototypical Emotional Episodes, and Other Things Called Emotion: Dissecting the Elephant.," *Journal of Personality and Social Psychology*, vol. 76, no. 5, pp. 805–819, 1999, https://doi.org/10.1037/0022-3514.76.5.805.

[13] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971, https://doi.org/10.1037/h0030377.

[14] K. Zhou, Y. Zhang, and T. Chen, "Emotion Recognition with Contextualized Attention and Transformer-based Encoders," vol. 1, Art. no. 102111, 2023.

[15] S. Mohammad, "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)*, Melbourne, Australia, 2018, pp. 174–184, https://doi.org/10.18653/v1/P18-1017.

[16] S. Chen, C. Li, and X. Zeng, "VAD Emotion Distribution Augmented BERT for Fine-Grained Emotion Recognition," *International Journal of Computer Science*, vol. 52, no. 9, pp. 3448–3458.

[17] T. Bänziger, D. Grandjean, and K. R. Scherer, "Emotion Recognition from Expressions in Face, Voice, and Body: The Multimodal Emotion Recognition Test (MERT).," *Emotion*, vol. 9, no. 5, pp. 691–704, 2009, https://doi.org/10.1037/a0017088.

[18] S. Gayathri and G. Maragatham, "A Survey on Text Based Emotion Detection," *International Journal of Creative Research Thoughts*, vol. 9, no. 3, pp. 2320–2882, Mar. 2021.

[19] W. Shen, J. Chen, X. Quan, and Z. Xie, "DialogXL: All-in-One XLNet for Multi-Party Conversation Emotion Recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, pp. 13789–13797, May 2021, https://doi.org/10.1609/aaai.v35i15.17625.

[20] X. Xu, X. Cheng, C. Chen, H. Fan, and M. Wang, "Emotion Recognition from Multi-Channel EEG via an Attention-Based CNN Model," in *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery*, vol. 153, N. Xiong, M. Li, K. Li, Z. Xiao, L. Liao, and L. Wang, Eds. Cham: Springer International Publishing, 2023, pp. 285–292.

[21] H. Pajupuu, K. Kerge, and R. Altrov, "Lexicon-Based Detection of Emotion in Different Types of Texts: Preliminary Remarks," *Eesti Rakenduslingvistika Ühingu aastaraamat Estonian Papers in Applied Linguistics*, vol. 0, no. 8, pp. 171–184, 2012, https://doi.org/10.5128/ERYa8.11.

[22] D. Lim, K. Lee, J. Jo, H. Lim, H. Bae, and C. Kang, "Web-Based Platform for Quantitative Depression Risk Prediction via VAD Regression on Korean Text and Multi-Anchor Distance Scoring," *Applied Sciences*, vol. 15, no. 18, Sep. 2025, Art. no. 10170, https://doi.org/10.3390/app151810170.

[23] Y. Sūn, "BERT-Based Knowledge Distillation for Sentiment Analysis Model," *Computer Science and Application*, vol. 13, no. 10, pp. 1938–1947, 2023, https://doi.org/10.12677/CSA.2023.1310192.

[24] A. Devgan, "Contextual Emotion Recognition Using Transformer-Based Models." Aug. 01, 2023, https://doi.org/10.36227/techrxiv.23804025.v1.

[25] X. Shen, X. Liu, X. Hu, D. Zhang, and S. Song, "Contrastive Learning of Subject-Invariant EEG Representations for Cross-Subject Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2496–2511, Jul. 2023, https://doi.org/10.1109/TAFFC.2022.3164516.

[26] H. Plisiecki and A. Sobieszek, "Emotion Topology: Extracting Fundamental Components of Emotions from Text Using Word Embeddings," *Frontiers in Psychology*, vol. 15, Oct. 2024, Art. no. 1401084, https://doi.org/10.3389/fpsyg.2024.1401084.

[27] Y. Huo and Y. Ge, "Multi-Task Emotion Recognition Based on Dimensional Model and Category Label," *IEEE Access*, vol. 12, pp. 75169–75179, 2024, https://doi.org/10.1109/ACCESS.2024.3404990.

[28] A. Sianipar, P. Van Groenestijn, and T. Dijkstra, "Affective Meaning, Concreteness, and Subjective Frequency Norms for Indonesian Words," *Frontiers in Psychology*, vol. 7, Dec. 2016, https://doi.org/10.3389/fpsyg.2016.01907.

[29] R. Prabowo and M. Thelwall, "Sentiment Analysis: A Combined Approach," *Journal of Informetrics*, vol. 3, no. 2, pp. 143–157, Apr. 2009, https://doi.org/10.1016/j.joi.2009.01.003.

[30] C. Wan, M. Labeau, and C. Clavel, "EmoDynamiX: Emotional Support Dialogue Strategy Prediction by Modelling MiXed Emotions and Discourse Dynamics," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Albuquerque, NM, USA, 2025, pp. 1678–1695, https://doi.org/10.18653/v1/2025.naacl-long.81.

[31] K. R. Scherer, "The Dynamic Architecture of Emotion: Evidence for the Component Process Model," *Cognition and Emotion*, vol. 23, no. 7, pp. 1307–1351, Nov. 2009, https://doi.org/10.1080/02699930902928969.

[32] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 5370–5381, https://doi.org/10.18653/v1/P19-1534.

[33] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019, https://doi.org/10.1109/ACCESS.2019.2929050.

[34] K. R. Scherer and H. G. Wallbott, "Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning," *Journal of Personality and Social Psychology*, vol. 66, no. 2, pp. 310–328, 1994, https://doi.org/10.1037/0022-3514.66.2.310.

[35] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A Transformer-Based joint-encoding for Emotion Recognition and Sentiment Analysis," in *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, Seattle, USA, 2020, pp. 1–7, https://doi.org/10.18653/v1/2020.challengehml-1.1.

[36] M. Khan, P.-N. Tran, N. T. Pham, A. El Saddik, and A. Othmani, "MemoCMT: Multimodal Emotion Recognition Using Cross-Modal Transformer-Based Feature Fusion," *Scientific Reports*, vol. 15, no. 1, Feb. 2025, Art. no. 5473, https://doi.org/10.1038/s41598-025-89202-x.

[37] H. Zhang and D. Niu, "Path-Aware Knowledge Injection for Fine-Grained Emotion Recognition in Mental Health Counseling," *IEEE Access*, vol. 13, pp. 144798–144813, 2025, https://doi.org/10.1109/ACCESS.2025.3599547.

[38] A. S. Cowen and D. Keltner, "Self-Report Captures 27 Distinct Categories of Emotion Bridged by Continuous Gradients," *Proceedings of the National Academy of Sciences*, vol. 114, no. 38, Sep. 2017, https://doi.org/10.1073/pnas.1702247114.

[39] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International Affective Picture System (IAPS): Technical Manual and Affective Ratings," *International Affective Picture System*, 1997.

[40] P. Ekman, "Facial Expression and Emotion," *American Psychologist*, vol. 48, no. 4, pp. 384–392, 1993, https://doi.org/10.1037/0003-066X.48.4.384.

[41] P. Mullangi *et al.*, "Sentiment and Emotion Modeling in Text-based Conversations utilizing ChatGPT," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20042–20048, Nov. 2024, https://doi.org/10.48084/etasr.9508.

[42] I. Darmawan, H. Elmunsyah, and D. D. Prasetya, "ALBERTIR: A BERT-Based Pretraining for Indonesian Religious Texts Using Qur'an and Hadith Translations," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 28307–28312, Oct. 2025, https://doi.org/10.48084/etasr.12977.