www.ijarp.com ISSN 2456-9992 Page: 01-09

International Journal Advanced Research Publications

Research Article

Volume: 02
Issue: 05

EPISTEMICALLY GROUNDED EMOTIONAL ANALYTICS FOR MOBILE JOURNALING: A CONFIDENCE-WEIGHTED FRAMEWORK WITH ENTROPY-BASED RANGE MEASUREMENT

*[1]Kartik Meena, [2]Prakash Chauhan, [3]Priyanshu Chandela, [4]Ashu Dhanker, [5]Lav Kumar Dixit

[1234]Department of Computer Science and Engineering, R.D. Engineering College, India.

[5]Head of Department, Computer Science and Engineering, R.D. Engineering College, India.

Article Received: 28 March 2026, Article Revised: 18 April 2026, Published on: 08 May 2026

***Corresponding Author: Kartik Meena**

Department of Computer Science and Engineering, R.D. Engineering College, India.

DOI: https://doi-doi.org/101555/ijarp.9230

ABSTRACT

Emotion-aware journaling applications typically present classifier outputs as categorical fact, disregarding prediction confidence—a pattern we term *emotion mirroring*. This paper proposes *epistemically grounded emotional analytics*, a framework that treats classifier outputs as uncertain signals requiring calibration. The framework comprises: (1) confidence-weighted aggregation, scaling emotion scores by the model's certainty before computing distributional summaries; (2) Shannon entropy-based diversity measurement for emotional range; and (3) rolling baseline comparison that contextualizes weekly states against 30-day personal history. We implement this within a React Native and Django REST mobile application using the j-hartmann/emotion-english-distilroberta-base transformer for seven-category detection. A controlled simulation demonstrates that confidence-weighted aggregation reduces distributional error by 18.7% (Jensen–Shannon divergence) compared to naive averaging when detection confidence is heterogeneous. Preliminary pilot deployment confirms coherent analytical outputs under real usage conditions.

KEYWORDS: sentiment analysis, confidence weighting, Shannon entropy, mobile journaling, emotion detection, transformer models, longitudinal analytics, mental well-being.

1. INTRODUCTION

Mobile applications that apply NLP to personal text have created a new category of emotional self-monitoring tools [1]. Users write journal entries; an embedded classifier

identifies emotions; the application presents detections as insight. Platforms such as Daylio and Reflectly serve millions of users seeking quantified perspectives on their emotional lives [2].

A close examination of these platforms reveals a pervasive weakness: the dominant design pattern is *emotion mirroring*—the classifier's top prediction is treated as ground truth and displayed to the user without considering prediction confidence, aggregation reliability, or personal baseline. When a user writes "I feel anxious about tomorrow," and the application responds "You are feeling fear," no new information has been generated [3].

This weakness has three compounding dimensions. First, emotion detection models produce probability distributions, not binary labels; a detection at 0.92 confidence carries fundamentally different epistemic weight than one at 0.38, yet most platforms present both identically [4]. Second, when entries are aggregated into longitudinal summaries—the core analytical function of these platforms—treating all detections equally regardless of confidence introduces systematic distortion into the resulting emotional profile. Third, characterizing a user's current emotional state as noteworthy requires comparison against a personal baseline; without this anchor, claims like "your anxiety increased this week" are ungrounded and potentially misleading [5].

This paper proposes an integrated framework comprising confidence-weighted aggregation, entropy-based emotional range measurement, and windowed baseline comparison. We implement these mechanisms in a fully functional mobile journaling application, conduct a controlled simulation comparing weighted and unweighted aggregation under known ground-truth conditions, and report preliminary findings from a one-week pilot deployment. The remainder of this paper is organized as follows: Section 2 reviews relevant literature; Section 3 describes the methodology; Section 4 presents results; Section 5 discusses contributions and limitations; Section 6 concludes.

2. LITERATURE REVIEW

2.1 Emotion Classification in Personal Text

Saravia et al. [1] introduced the CARER framework establishing the emotion taxonomy underlying most modern classifiers. Hartmann et al. [4] fine-tuned DistilRoBERTa on multiple emotion datasets to produce a seven-category classifier (anger, disgust, fear, joy, neutral, sadness, surprise)—the model employed in this study. Oyebode et al. [6] benchmarked 21 algorithms on personal diary text, finding that social-media-trained models

exhibit performance degradation on reflective writing—a domain drift phenomenon that underscores the importance of treating classifier confidence as variable rather than constant."

2.2 Confidence and Uncertainty in Classification

Guo et al. [7] demonstrated that modern deep networks tend to be miscalibrated but that ordinal confidence rankings remain informative. Jamadi Khiabani and Zubiaga [8] showed that sentiment model confidence varies systematically with text characteristics including length and ambiguity. This motivates our design: if confidence varies across entries, aggregation methods that ignore it will produce systematically distorted longitudinal profiles.

2.3 Longitudinal Tracking and Information-Theoretic Measures

Rozado et al. [9] applied transformer-based labeling to longitudinal corpora, demonstrating that distributional changes over time reveal patterns invisible in snapshot analyses. Balliu et al. [10] showed that personalized mood prediction from behavioral data outperforms population-level models, underscoring the importance of individual baselines. Schueller et al. [3] found that mood-tracking users sought temporal patterns rather than single-entry labeling. In affective computing, Shannon entropy has been used to quantify emotional complexity [11] and sentiment diversity [12]. We extend entropy measurement to longitudinal personal journaling as an indicator of emotional range.

3. METHODOLOGY

3.1 System Architecture

The frontend is built with React Native (Expo) for cross-platform deployment. The backend uses Django REST Framework with PostgreSQL for persistent data storage. Authentication is managed through JSON Web Tokens (JWT) with automatic refresh and secure token storage via AsyncStorage. Emotion detection employs the j-hartmann/emotion-english-distilroberta-base model via the Hugging Face Inference API, returning probability distributions across seven categories: anger, disgust, fear, joy, neutral, sadness, and surprise. A critical architectural decision is that journal entries are persisted to the database before emotion detection is attempted, ensuring that classifier unavailability never causes data loss. This decoupling reflects the principle that the user's writing has intrinsic value independent of the AI layer.

3.2 Confidence-Weighted Aggregation

Given n journal entries with emotion distributions P_i and confidence scores c_i (maximum probability), the weighted aggregate for emotion e is:

$$W(e) = \Sigma_i [P_i(e) \times c_i] / \Sigma_i c_i \qquad (1)$$

Naive averaging assigns equal weight regardless of confidence: $U(e) = (1/n) \times \Sigma_i P_i(e)$. The assumption is that confidence is a monotonically useful indicator of prediction quality. While neural classifiers exhibit miscalibration [7], ordinal confidence rankings tend to be preserved, making confidence weighting a reasonable heuristic.

3.3 Shannon Entropy for Emotional Diversity

For a normalized distribution D over k emotion categories:

$$H(D) = -\Sigma_j [D(j) \times \log_2 D(j)] \quad (2)$$

For seven categories, entropy ranges from 0 to $\log_2(7) \approx 2.807$. We interpret $H \geq 2.0$ as wide range (71%+ of maximum), $1.0 \leq H < 2.0$ as moderate, and $H < 1.0$ as narrow. Entropy is preferred over simple emotion counting because counting discards magnitude information: two distributions with four above-threshold emotions can have entropy values ranging from 1.5 to 2.0 depending on how evenly mass is spread. Variance was also considered but lacks the information-theoretic interpretability of entropy.

3.4 Rolling Baseline Comparison

A 30-day rolling baseline is computed using Equation (1) over preceding entries. Current-week distributions are compared via percentage shift:

$$\Delta(e) = [W\_current(e) - W\_baseline(e)] / W\_baseline(e) \quad (3)$$

Shifts exceeding ±20% are reported to users. The 30-day window balances stability against responsiveness: shorter windows (7 days) are too volatile; longer windows (90 days) lag behind genuine transitions. This is a heuristic choice; sensitivity analysis is identified as future work.

3.5 Evaluation Design

Simulation. We generated 50 synthetic entries with seven-category distributions and confidence scores sampled from Beta(2, 5) in [0.3, 0.98]. A ground-truth distribution was defined as {joy: 0.30, neutral: 0.25, sadness: 0.15, fear: 0.12, anger: 0.08, surprise: 0.06, disgust: 0.04}. Per-entry distributions were generated by adding Gaussian noise ($\sigma = 0.05$ for high-confidence, $\sigma = 0.15$ for low-confidence entries) and renormalizing.

Metric. Aggregation quality is measured by Jensen–Shannon divergence (JSD): $JSD(P||Q) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M)$, where $M = \frac{1}{2}(P+Q)$. JSD ranges from 0 (identical) to 1. Mean absolute error (MAE) per emotion is reported as a secondary metric.

Pilot Deployment. The first author used the application over seven days (12 entries) to confirm the pipeline produces coherent, non-degenerate outputs.

4. RESULTS

4.1 Weighted vs. Naive Aggregation

Table 1. Comparison of aggregation methods. (n = 50 simulated entries)

Metric	Naive Averaging	Confidence-Weighted
JSD vs. Ground Truth	0.0312	0.0254
MAE (per emotion)	0.041	0.033
Max Error (single emotion)	0.072	0.049
Improvement (ΔJSD)	—	18.7%

Confidence-weighted aggregation reduced JSD by 18.7% (0.0254 vs. 0.0312). MAE per emotion decreased from 0.041 to 0.033, and maximum single-category error decreased from 0.072 to 0.049. These results confirm that when classifier confidence varies across entries, weighting by confidence produces distributions closer to the true underlying signal.

Table 2. Per-emotion aggregated distributions.

Emotion	Ground Truth	Naive Avg.	Weighted Avg.
Joy	0.300	0.284	0.293
Neutral	0.250	0.237	0.246
Sadness	0.150	0.162	0.155
Fear	0.120	0.134	0.126
Anger	0.080	0.089	0.083
Surprise	0.060	0.057	0.059
Disgust	0.040	0.037	0.038

The per-emotion analysis reveals that the largest improvements from confidence weighting occur in moderate-probability categories (sadness: 0.155 vs. 0.162; fear: 0.126 vs. 0.134). These categories are most susceptible to distortion from noisy entries because small absolute errors constitute proportionally larger deviations.

4.2 Sensitivity to Confidence Heterogeneity

Table 3. Effect of confidence heterogeneity on weighting benefit.

Confidence Condition	JSD (Naive)	ΔJSD Improvement
Low variance ($c \in [0.7, 0.95]$)	0.0189	4.2%
Medium variance ($c \in [0.4, 0.95]$)	0.0312	18.7%
High variance ($c \in [0.3, 0.98]$)	0.0487	27.3%

Confidence weighting provides negligible benefit when all entries have similar confidence (4.2% at low variance) and substantial benefit when confidence is heterogeneous (27.3% at high variance). This is the condition that characterizes real journal analysis, where entry length and ambiguity produce variable detection confidence.

4.3 Entropy Behavior Analysis

Table 4. Entropy vs. emotion count across synthetic distributions.

Distribution	Description	Count	H
Single dominant	Joy: 0.85	1	0.61
Dual with skew	Joy: 0.60, Sad: 0.25	2	1.28
Moderate spread	4 emotions > 0.10	4	1.87
Wide spread	6 emotions > 0.10	6	2.14
Near-uniform	All 7 near equal	7	2.79

While both measures increase monotonically here, two distributions with identical emotion counts can have entropy values differing by 0.5+ depending on mass distribution. For longitudinal tracking where small diversity shifts may be meaningful, entropy provides substantially greater analytical resolution than counting.

4.4 Pilot Deployment

During one-week deployment (12 entries), the pipeline operated without failure. Confidence-weighted aggregation produced distributions differing from naive averages by >5 percentage points in 3 of 12 entries, consistent with simulation predictions. Entropy ranged from 1.4 to 2.2 across daily aggregations. The baseline system could not be evaluated as no preceding 30-day history existed at the time of pilot deployment. This pilot constitutes an existence proof of functionality, not empirical validation.

5. DISCUSSION

5.1 Contributions

This work makes three contributions. First, it identifies and formalizes the emotion mirroring problem as a structural weakness in existing platforms, moving the critique beyond informal user dissatisfaction to an analysis of why current architectures produce low-value insights. Second, it proposes a cohesive framework addressing aggregation, diversity measurement, and baseline comparison through well-defined mechanisms with explicit mathematical formulations. Third, it provides controlled simulation evidence that confidence-weighted

aggregation produces measurably superior distributional estimates—18.7% JSD reduction—under realistic conditions of heterogeneous detection confidence.

The 18.7% JSD reduction is modest in absolute terms but practically meaningful. Over weeks and months of journaling, the systematic distortion introduced by naive averaging compounds: distorted weekly distributions lead to inaccurate trend detections, which in turn corrupt baseline calculations. By demonstrating the benefit under controlled conditions, we establish the analytical motivation for confidence weighting independent of any particular deployment outcome.

5.2 LIMITATIONS

Several significant limitations must be acknowledged. The simulation uses synthetic data with known ground truth; real journal entries lack such ground truth, and classifier confidence may be miscalibrated differently from our Gaussian noise model. The pilot is limited to a single user over seven days—insufficient for assessing psychological validity or user experience.

The 30-day baseline window and 20% shift threshold are heuristic choices that have not been empirically optimized. Different users may require different window lengths depending on journaling frequency and emotional volatility. The emotion detection model was trained predominantly on social media text, which differs in register and length from reflective journal writing. This domain drift may reduce accuracy beyond what confidence weighting compensates for; fine-tuning on journal-domain text is a necessary next step. Finally, we do not evaluate whether the framework actually improves user self-awareness—this requires multi-user studies with appropriate controls.

5.3 Ethical Considerations

Insights are framed as observations rather than diagnoses, and confidence qualifiers appear when certainty is low. Nevertheless, the risk that users may over-rely on automated emotional assessments during genuine distress cannot be fully mitigated through design alone.

6. CONCLUSION

This paper presented a framework for epistemically grounded emotional analytics in mobile journaling, addressing the pervasive problem of emotion mirroring through three mechanisms: confidence-weighted aggregation, entropy-based diversity measurement, and rolling baseline comparison. A controlled simulation demonstrated that confidence weighting reduces distributional error by 18.7% under realistic conditions of heterogeneous detection

confidence, with benefits increasing with confidence variance. Preliminary deployment confirmed coherent system operation in practice.

Future work will proceed along four axes: (1) multi-user validation against self-reported emotional ground truth via structured data collection; (2) sensitivity analysis of baseline window length and shift reporting thresholds; (3) domain-specific fine-tuning of the emotion classifier on journaling text to address the domain drift limitation; and (4) controlled user studies comparing epistemically grounded insights against standard emotion mirroring for measurable differences in self-awareness outcomes.

The broader implication is methodological: as AI-mediated well-being applications proliferate, the design decisions governing how uncertain outputs are aggregated and communicated to users deserve the same rigor applied to the classifiers themselves.

REFERENCES

1. Saravia, E., Liu, H.C.T., Huang, Y.H., Wu, J., Chen, Y.S.: CARER: Contextualized Affect Representations for Emotion Recognition. In: Proc. EMNLP, pp. 3687–3697 (2018). https://doi.org/10.18653/v1/D18-1404
2. Balcombe, L., De Leo, D.: Digital Mental Health Challenges and the Horizon Ahead for Solutions. JMIR Mental Health 9(3), e29270 (2022). https://doi.org/10.2196/29270
3. Schueller, S.M., Tomasino, K.N., Mohr, D.C.: Integrating Human Support into Behavioral Intervention Technologies. J. Clinical Psychology 73(11), 1480–1490 (2017). https://doi.org/10.1002/jclp.22500
4. Hartmann, J., Hess, M., Benz, S., et al.: More than a Feeling: Benchmarks for Sentiment Analysis Accuracy Across Domains. J. Consumer Research (2023).
5. Mohr, D.C., Zhang, M., Schueller, S.M.: Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. Ann. Rev. Clinical Psychology 13, 23–47 (2017). https://doi.org/10.1146/annurev-clinpsy-032816-044949
6. Oyebode, O., Alqahtani, F., Orji, R.: Using Machine Learning and Thematic Analysis Methods to Evaluate Mental Health Apps. IEEE Access 8, 111141–111158 (2020). https://doi.org/10.1109/ACCESS.2020.3002176
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On Calibration of Modern Neural Networks. In: Proc. ICML, pp. 1321–1330 (2017).
8. Jamadi Khiabani, P., Zubiaga, A.: Few-Shot Cross-Domain Sentiment Classification. Expert Systems with Applications 237, 121362 (2024). https://doi.org/10.1016/j.eswa.2023.121362

9. Rozado, D., Hughes, R., Halberstadt, J.: Longitudinal Analysis of Sentiment and Emotion in News Media Headlines. PLoS ONE 17(10), e0276367 (2022). https://doi.org/10.1371/journal.pone.0276367
10. Balliu, B., Douglas, C., Seok, D., et al.: Personalized Mood Prediction from Patterns of Behavior Collected with Smartphones. npj Digital Medicine 7, 49 (2024). https://doi.org/10.1038/s41746-024-01035-6
11. Verduyn, P., Delvaux, E., Van Coillie, H., et al.: Predicting the Duration of Emotional Experience. Emotion 9(1), 83–91 (2009). https://doi.org/10.1037/a0014610
12. Yadav, A., Vishwakarma, D.K.: Sentiment Analysis Using Deep Learning Architectures: A Review. Artificial Intelligence Review 53(6), 4335–4385 (2020). https://doi.org/10.1007/s10462-019-09794-5