Vol. 8, No. 3, June 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

An Empirical Evaluation of Confidence Miscalibration in Vanilla BERT-Based Stress Detection on Social Media

Rizaldi[1]*, Kusrini[2], Ema Utami[3], I Made Artha Agastya[4]

[1,2,3]Doctoral Program in Informatics, Postgraduate Program, Universitas AMIKOM Yogyakarta, Indonesia

[4] Master Program in Informatics, Postgraduate Program, Universitas AMIKOM Yogyakarta, Indonesia

Received:
November 1, 2025
Revised:
May 10, 2026
Accepted:
May 27 2026
Published:
June 22, 2026

Corresponding Author:
Author Name*:
Rizaldi
Email*:
rizaldi@students.amikom.ac.id

DOI:
10.63158/journalisi.v8i3.1634

Abstract. This study evaluates the reliability of confidence estimates produced by a Vanilla BERT classifier for stress detection using the Dreaddit benchmark. BERT-base-uncased was fine-tuned on 3,553 labeled text segments, following the standard split of 2,838 training samples and 715 test samples. The model was assessed as a single diagnostic baseline without additional linguistic features, label smoothing, post-hoc calibration, or other calibration interventions. Evaluation was conducted using discriminative performance metrics, including accuracy, precision, recall, and F1-score, as well as probabilistic reliability metrics, including Brier Score, Expected Calibration Error, Adaptive Calibration Error, and a reliability diagram. The Vanilla BERT model achieved 79.02% accuracy, 78.00% precision, 82.65% recall, and 80.26% F1-score, indicating competitive classification performance for stress detection. However, the calibration results revealed noticeable miscalibration, with a Brier Score of 0.1565, Expected Calibration Error of 0.0847, and Adaptive Calibration Error of 0.0880. The most prominent confidence mismatch occurred in the 0.8–0.9 confidence interval, while the 0.9–1.0 interval contributed the most to Expected Calibration Error due to its larger sample proportion. These findings show that although Vanilla BERT performs reasonably well in distinguishing stressed from non-stressed text, its confidence estimates are not fully reliable. Therefore, this study positions Vanilla BERT as a diagnostic reliability baseline and emphasizes the importance of evaluating stress detection models using both classification performance and probabilistic calibration criteria.

Keywords: Stress detection, Vanilla BERT, expected calibration error, reliability diagram, uncertainty estimation

1. INTRODUCTION

Stress is one of the mental health problems increasingly expressed through social media. Platforms such as Reddit provide a space for users to write personal experiences, emotional pressure, relational conflicts, financial problems, trauma, and anxiety in textual narratives. Social media can function as a coping space where individuals express psychological distress and seek emotional support indirectly [1]. This condition creates opportunities for developing Natural Language Processing models to automatically detect stress indications from social media texts. However, the error distribution shown in Figure 3 also indicates that discriminative performance alone is insufficient to assess the reliability of the model in sensitive digital mental health contexts. Reddit-based stress detection research was strengthened by the introduction of Dreaddit, a benchmark dataset for stress analysis in social media constructed from several stressor domains, including abuse, anxiety, financial, post-traumatic stress disorder, and social domains [2]. Turcan and McKeown showed that machine learning and deep learning approaches can be used for text-based stress classification, with BERT-base achieving an F1-score of 0.8065 and the best non-neural model, Logistic Regression with domain-specific Word2Vec and additional features, achieving an F1-score of 0.7980 [2]. These results position Dreaddit as an important benchmark for evaluating social media-based stress detection models.

Along with the development of Transformer models, mental health text classification has shifted from classical approaches to BERT-, RoBERTa-, and other Transformer-based architectures. Prior studies have shown that Hugging Face Transformer models, MIRoBERTa, and hybrid Transformer architectures can improve contextual representation and classification performance in mental health prediction tasks using social media texts [3]-[5]. In the specific context of Dreaddit, Ilias et al. demonstrated that Transformer models can be improved by integrating linguistic features and label smoothing, with their M-BERT + LIWC + Label Smoothing model achieving an F1-score of 0.8310 [6]. Classical machine learning approaches also remain relevant, as Oryngozha et al. showed that a Bag of Words model with Logistic Regression could still achieve an F1-score of around 0.79 in Reddit-based stress detection [7]. These studies indicate that both Transformer-based and lexical models can produce competitive classification performance, but classification

performance alone does not necessarily indicate that the model's probability estimates are reliable.

A major limitation of Transformer-based models is their tendency to produce overly confident predictions. Modern deep neural networks often generate prediction probabilities that are not aligned with actual accuracy, especially when trained with hard labels and loss functions oriented toward accuracy optimization [8]. This issue is particularly important in stress detection because social media texts are often ambiguous, informal, contextual, and influenced by users' writing styles. A model may predict a text as Stress with high confidence simply because it contains negative lexical cues, even though the broader context does not necessarily indicate stress. Conversely, stress expressions may be missed when the text is written in a polite, constructive, or transactional tone.

Recent studies on trustworthy artificial intelligence emphasize that language models should not be evaluated solely through accuracy or F1-score. Confidence estimation, calibration, and uncertainty estimation are important for safer model use in sensitive domains [9], [10]. Moreover, miscalibration is not always visible through global metrics because it may be hidden within certain input groups [11]. Therefore, calibration evaluation should consider semantic characteristics of the data, not only aggregate probability distributions. This direction is consistent with embedding-based topic representation methods such as Top2Vec, which can be used to map the semantic structure of social media texts [12].

Despite these advances, prior studies still leave several related gaps. First, Dreaddit-based stress detection studies have mainly emphasized discriminative performance, such as accuracy, precision, recall, and F1-score, while the reliability of model confidence remains less examined [2]-[7]. More broadly, uncertainty estimation and confidence-calibration studies have shown that neural networks may produce over-confident or under-confident predictions, and that calibration concerns the alignment between estimated confidence and actual accuracy [8], [9], [13]. Second, calibration-oriented Transformer studies have generally reported global calibration metrics but have not sufficiently connected miscalibration with confidence-bin behavior and systematic error-pattern interpretation [6], [11]. Third, limited work has examined high-confidence errors in

Dreaddit, particularly cases where a model produces incorrect predictions with strong confidence despite the ambiguity of social media narratives.

Therefore, this study presents a single-baseline empirical evaluation of Vanilla BERT for social media-based stress detection. The novelty lies in combining reproducible Vanilla BERT baseline replication, probabilistic calibration evaluation, confidence-bin analysis, and systematic error-pattern interpretation. The objective is to evaluate whether a competitive Vanilla BERT classifier on Dreaddit also produces confidence estimates aligned with empirical correctness. This study treats Dreaddit labels as text-based stress indications rather than clinical diagnoses and positions the proposed analysis as a diagnostic reliability baseline for future calibration, uncertainty estimation, and thematic diagnostic studies.

2. METHODS

2.1. Research Design

This study used a computational experimental design to evaluate the classification performance and probabilistic reliability of a single Vanilla BERT baseline for text-based stress detection on social media. The main focus was not only to measure the model's ability to predict Stress and Not Stress labels, but also to examine whether the confidence scores produced by the model were aligned with its actual accuracy. Therefore, the experiment was positioned as a reproducible baseline evaluation protocol and an initial diagnostic stage before future calibration or uncertainty-aware model development.

The experimental workflow consisted of five stages: dataset preparation, text tokenization, Vanilla BERT fine-tuning, classification and calibration evaluation, and qualitative error analysis. Figure 1 illustrates the complete research workflow used in this study, from Dreaddit dataset preparation to qualitative error-pattern interpretation.

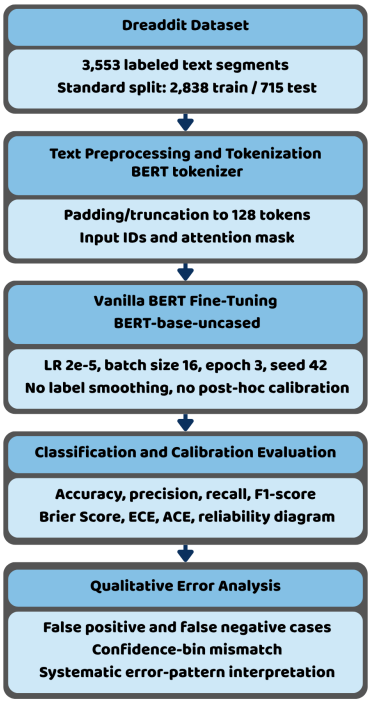

Figure 1. Research Workflow of the Vanilla BERT Baseline Evaluation Protocol

2.2. Dataset

The dataset used in this study was Dreaddit, a benchmark dataset for stress detection on Reddit introduced by Turcan and McKeown [2]. Dreaddit consists of 3,553 text segments with binary labels: Stress and Not Stress. The labels were obtained through a crowdsourcing annotation process using a majority voting mechanism. In this study, Dreaddit labels were treated as text-based stress indications, not as clinical diagnoses.

This study preserved the standard Dreaddit train-test split to maintain fair comparability with previous studies. The training set consisted of 2,838 segments, while the test set consisted of 715 segments. The test set was used for evaluating the classification performance and probabilistic calibration of the Vanilla BERT baseline. The class distribution of the Dreaddit dataset used in this study is presented in Table 1.

Table 1. Class Distribution of the Dreaddit Dataset

Split	Not Stress	Stress	Total
Training set	1,350 (47.57%)	1,488 (52.43%)	2,838
Test set	346 (48.39%)	369 (51.61%)	715
Total	1,696 (47.73%)	1,857 (52.27%)	3,553

This distribution shows that the Stress and Not Stress classes are relatively balanced in both the training and test sets.

2.3. Text Preprocessing and Tokenization

The texts were processed using the BERT-base-uncased tokenizer. Tokenization was performed using truncation and padding with a maximum sequence length of 128 tokens. Each text was converted into three main input components: input_ids, attention_mask, and labels. The uncased variant was selected because social media texts often contain inconsistent capitalization. Using this tokenizer ensured that all texts were processed in a format compatible with the BERT architecture [14]. An example of the preprocessing and tokenization representation is shown in Table 2. The example is presented as a shortened excerpt to illustrate how a raw text segment was converted into BERT-compatible input representation.

Table 2. Example of Text Preprocessing and Tokenization Representation

Stage	Example
Raw text excerpt	"I get pissed… she ends up breaking up with me…"
Tokenized sequence	[CLS] i get pissed … she ends up breaking up with me … [SEP]
Input representation	Converted into input_ids and attention_mask
Sequence handling	Truncated or padded to 128 tokens
Label representation	Stress = 1, Not Stress = 0

This representation shows that the raw text was not manually transformed into handcrafted linguistic features but was encoded using the BERT tokenizer and represented as token IDs and attention masks for model fine-tuning.

2.4. Baseline Model and Training Configuration

The baseline model used in this study was BERT-base-uncased. The model was fine-tuned for binary classification using the Dreaddit training set. The selection of BERT-base-uncased and the fine-tuning setting follows the general BERT fine-tuning paradigm introduced by Devlin et al. and subsequent BERT text-classification fine-tuning studies [14], [15]. The training configuration is presented in Table 3.

Table 3. Training Configuration of the Vanilla BERT Baseline

Parameter	Value
Model	BERT-base-uncased
Maximum sequence length	128 tokens
Learning rate	2e-5
Training batch size	16
Evaluation batch size	32
Weight decay	0.01
Number of epochs	3
Loss function	Cross-Entropy
Random seed	42

This experiment was not intended as a full hyperparameter optimization process. The configuration was used as a controlled baseline replication so that the evaluation results could be reproduced and compared with previous Dreaddit studies. The model was trained for 3 epochs using a fixed configuration. The standard Dreaddit split was preserved for comparability with previous studies. Since no additional validation split or repeated-seed model selection was used, the results are interpreted as a controlled diagnostic baseline rather than a fully optimized or multi-run generalization estimate.

2.5. Classification Performance Evaluation

Classification performance was evaluated using accuracy, precision, recall, and F1-score. These metrics were calculated from the confusion matrix components, namely true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Accuracy measures the proportion of correct predictions among all test samples, precision measures the correctness of samples predicted as Stress, recall measures the model's ability to detect actual Stress samples, and F1-score balances precision and recall. Equations (1)–(4) define the standard classification metrics used in this study.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

2.6. Probabilistic Calibration Evaluation

In addition to classification metrics, this study evaluated the probabilistic reliability of the model using Brier Score, Expected Calibration Error (ECE), Adaptive Calibration Error (ACE), and a reliability diagram. This evaluation was necessary because a model with a high F1-score does not necessarily produce trustworthy confidence estimates [6], [8], [9]. The use of ECE and reliability diagrams also follows common evaluation protocols in neural network calibration studies [13]. Equations (5)–(10) define the softmax-based confidence score, ECE, ACE, and Brier Score used to evaluate probabilistic reliability.

For each test sample x_i, the model produced class probabilities using the softmax function:

$$P(y = c \mid x_i) = \frac{\exp(z_{ic})}{\sum_{j=1}^{C} \exp(z_{ij})} \quad (5)$$

where z_{ic} is the logit for class $c$for sample x_i, and C is the total number of classes.
The predicted label was assigned to the class with the highest predicted probability:

$$\hat{y}_i = \arg\max_c P(y = c \mid x_i) \quad (6)$$

The prediction confidence was defined as the maximum predicted class probability:

$$\hat{p}_i = \max_c P(y = c \mid x_i) \quad (7)$$

where $\hat{p}_i$ denotes the prediction confidence for sample x_i. In this study, ECE, ACE, and the reliability diagram were computed based on prediction confidence, whereas the Brier Score was computed based on the predicted probability of the positive class, namely the Stress class.

Expected Calibration Error (ECE) measures the weighted average deviation between model confidence and empirical accuracy across confidence bins. In this study, ECE was computed using 10 equal-width confidence bins. Ten bins were selected to balance interpretability and sample adequacy in the Dreaddit test set. With 715 test samples, this bin setting provides sufficiently interpretable confidence intervals while avoiding overly sparse calibration bins. If B_m denotes the set of samples in the m-th bin, ECE is defined as follows:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|\ B_m\ |}{n} |\ \mathrm{acc}(B_m) - \mathrm{conf}(B_m)\ | \tag{8}$$

where $M = 10$, n is the total number of test samples, and $|\ B_m\ |$ is the number of samples in bin B_m. The empirical accuracy and average confidence in each bin are computed as:

$$\mathrm{acc}(B_m) = \frac{1}{|\ B_m\ |} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i)$$

$$\mathrm{conf}(B_m) = \frac{1}{|\ B_m\ |} \sum_{i \in B_m} \hat{p}_i$$

where $\mathbf{1}(\hat{y}_i = y_i)$ is an indicator function that equals 1 if the prediction is correct and 0 otherwise. A lower ECE indicates better calibration.

Adaptive Calibration Error (ACE) follows a similar principle to ECE but uses adaptive bins instead of fixed-width bins. In this study, ACE was computed using 10 adaptive bins with relatively balanced sample counts, using the same bin count as ECE for consistency. The adaptive bins were formed by sorting test samples according to prediction confidence and partitioning them into 10 equal-frequency bins. ACE is defined as follows:

$$\mathrm{ACE} = \frac{1}{M} \sum_{m=1}^{M} |\ \mathrm{acc}(B_m^{\mathrm{adaptive}}) - \mathrm{conf}(B_m^{\mathrm{adaptive}})\ | \tag{9}$$

where B_m^{adaptive} denotes the set of samples in the m-th adaptive bin. ACE complements ECE by reducing the potential bias caused by uneven confidence distributions across fixed-width bins. A lower ACE indicates better calibration.

The Brier Score was used to measure the mean squared error between the predicted probability of the positive class and the actual binary label. In this study, the positive class was Stress. The Brier Score is defined as follows:

$$\mathrm{Brier} = \frac{1}{n} \sum_{i=1}^{n} (p_i - y_i)^2 \tag{10}$$

where p_i is the predicted probability of the Stress class for sample i, and y_i is the actual binary label, with 1 representing Stress and 0 representing Not Stress. A lower Brier Score indicates better probabilistic prediction quality.

The reliability diagram was used to visualize the relationship between model confidence and empirical accuracy across the 10 equal-width confidence bins. A model is considered well calibrated when its reliability curve approaches the diagonal line of perfect calibration. Points below the diagonal indicate overconfidence, meaning that the model's confidence is higher than its empirical accuracy, while points above the diagonal indicate underconfidence.

2.7. Error Analysis

Error analysis was conducted to understand the model's failure patterns, particularly in cases where the model made incorrect predictions with high confidence. False positive and false negative samples were extracted from the test-set prediction output and sorted by model confidence. To make the analysis more systematic, all 86 false positives and 64 false negatives were screened using dominant linguistic-pattern categories derived from the observed error samples.

The analysis focused on three broad conditions: Not Stress texts containing negative lexical cues, Stress texts written in polite or constructive tones, and ambiguous texts that could lead to different human interpretations. The resulting pattern counts were used to complement illustrative examples and to explain how lexical shortcuts and narrative ambiguity may contribute to model overconfidence. Therefore, error analysis was not merely used as a complement to quantitative evaluation, but also as an interpretive basis for explaining the sources of model miscalibration.

2.8. Reproducibility Protocol

To ensure reproducibility, the experiment was conducted in a Python-based computational environment using Hugging Face Transformers and PyTorch. The random seed was fixed at 42 for the random module, NumPy, PyTorch, CUDA, and HuggingFace Trainer. All main experimental configurations, including model architecture, maximum sequence length, learning rate, batch size, weight decay, number of epochs, and evaluation metrics, were explicitly reported. Because this study used the standard Dreaddit train-test split and a single random seed, the reported values should be interpreted as controlled baseline diagnostic results rather than estimates of multi-run variability.

3. RESULTS AND DISCUSSION

3.1. Classification Performance of Vanilla Bert

Before reporting the final test-set classification performance, the training dynamics of the Vanilla BERT baseline were examined across three epochs. Figure 2 presents the changes in training loss, evaluation loss, and F1-score during fine-tuning. Figure 2 shows that training loss decreased from 0.4406 to 0.2047 across three epochs, while evaluation loss increased after the first epoch. The F1-score improved substantially from 0.7612 in epoch 1 to approximately 0.8026 in epoch 2 and remained relatively stable in epoch 3. This pattern indicates that the baseline achieved competitive discriminative performance within a short fine-tuning schedule, but the increasing evaluation loss also supports the need for calibration analysis rather than relying solely on F1-score.

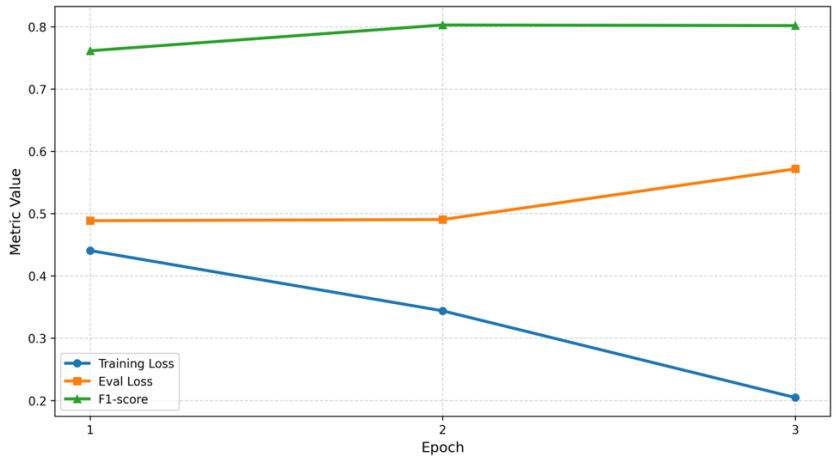

Figure 2. Training Dynamics of the Vanilla BERT Baseline Across Three Epochs

The first experiment was conducted to evaluate the discriminative ability of the Vanilla BERT model in classifying text into Stress and Not Stress. The evaluation results on the Dreaddit test set show that the model achieved an accuracy of 79.02%, precision of 78.00%, recall of 82.65%, and F1-score of 80.26%, as presented in Table 4.

Table 4. Classification Performance of Vanilla BERT on the Dreaddit Test Set

Metric	Value	Interpretation
Accuracy	79.02%	Overall proportion of correct predictions
Precision	78.00%	Correctness of predictions classified as Stress

Metric	Value	Interpretation
Recall	82.65%	Ability to detect actual Stress samples
F1-score	80.26%	Harmonic balance between precision and recall

The results in Table 4 show that Vanilla BERT achieved competitive classification performance on the Dreaddit dataset. The higher recall compared to precision indicates that the model was relatively sensitive in detecting texts labeled as Stress. This sensitivity is useful because the model was able to capture most text-based stress indications. However, the lower precision also indicates that this sensitivity was accompanied by an increase in false positives, where Not Stress texts were incorrectly predicted as Stress. This error pattern is further clarified through the confusion matrix in Figure 3. The model correctly classified 260 Not Stress samples and 305 Stress samples. However, it also produced 86 false positives and 64 false negatives.

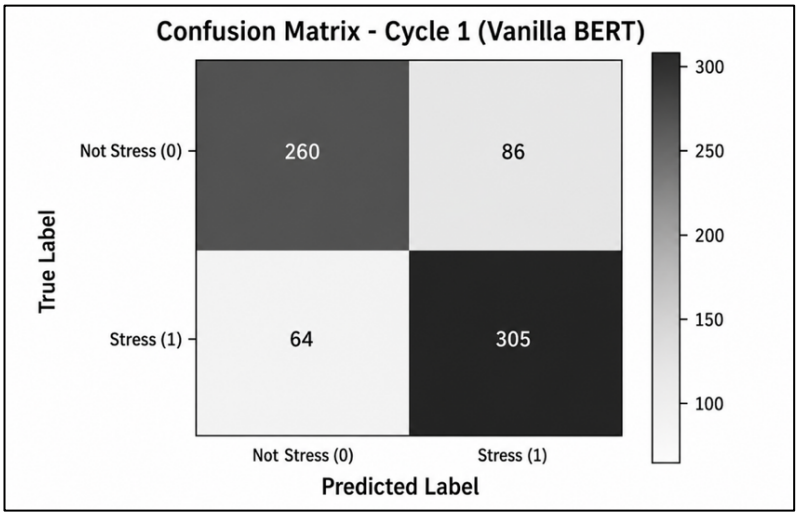

Figure 3. Confusion Matrix of Vanilla BERT on the Dreaddit Test Set

This finding answers the first objective of the study, namely, to establish a Vanilla BERT baseline that can be compared with previous studies. The F1-score of 80.26% indicates that Vanilla BERT remains a strong baseline for text-based stress classification. However, the error distribution shown in Figure 3 also indicates that discriminative performance alone is insufficient to assess the readiness of the model for digital mental health monitoring support systems.

3.2. Comparison With Previous Dreaddit Baselines

To assess the empirical position of the model against previous studies, the Vanilla BERT results were compared with several relevant baselines on the Dreaddit dataset. This comparison is presented in Table 5.

Table 5. Comparison With Previous Baselines on the Dreaddit Dataset

Model	Precision	Recall	F1-score	Source
Logistic Regression + domain Word2Vec + features	74.33%	83.20%	79.80%	[2]
BERT-base	75.18%	86.99%	80.65%	[2]
BoW + Logistic Regression	N/A	N/A	79.00%	[7]
BERT-base reported baseline	N/A	N/A	80.65%	[6]
M-BERT + LIWC + Label Smoothing	N/A	N/A	83.10%	[6]
Vanilla BERT	78.00%	82.65%	80.26%	This study

Based on Table 5, the Vanilla BERT model in this study achieved an F1-score of 80.26%, which is close to the BERT-base baseline reported in previous Dreaddit studies. Its F1-score is also slightly higher than the classical Logistic Regression-based and BoW + Logistic Regression baselines. This indicates that the Vanilla BERT baseline used in this study is reasonably competitive from a discriminative performance perspective and is suitable as a diagnostic baseline for subsequent probabilistic reliability analysis.

However, the result remains below the M-BERT + LIWC + Label Smoothing model. This difference is expected because the present study evaluates a Vanilla BERT model without additional linguistic features, label smoothing, or calibration intervention. Therefore, this comparison should not be interpreted as a claim of superiority over previous calibrated or feature-enhanced models. Instead, the purpose of this comparison is to show that a BERT baseline with competitive F1-score may still suffer from confidence reliability problems.

Furthermore, Table 5 should be interpreted only as a discriminative performance comparison because not all previous studies reported probabilistic calibration metrics such as Brier Score, Expected Calibration Error, Adaptive Calibration Error, or reliability

diagrams. This supports the main argument of this study that good classification performance does not automatically guarantee trustworthy prediction probabilities.

3.3. Probabilistic Calibration Results

The second objective of this study was to measure the probabilistic reliability of the model. The evaluation was conducted using Brier Score, Expected Calibration Error, Adaptive Calibration Error, and a reliability diagram. The evaluation results are presented in Table 6.

Table 6. Calibration Performance of Vanilla BERT

Calibration Metric	Value	Interpretation
Brier Score	0.1565	Squared error between Stress-class probability and the true binary label
Expected Calibration Error	0.0847	Average deviation of 8.47 percentage points between confidence and empirical accuracy
Adaptive Calibration Error	0.0880	Adaptive-bin calibration deviation of 8.80 percentage points

Table 6 shows that the model has a non-negligible reliability gap. The Expected Calibration Error value of 0.0847 means that there is an average deviation of approximately 8.47 percentage points between model confidence and empirical accuracy. The Adaptive Calibration Error value of 0.0880 shows that the miscalibration problem remains visible even when the data are divided using an adaptive binning scheme. Meanwhile, the Brier Score of 0.1565 indicates that the predicted probability for the Stress class still contains a considerable squared error.

These findings reveal a duality between classification performance and probabilistic reliability. The model may appear strong in terms of F1-score, but the confidence it produces does not fully reflect the actual likelihood of correctness. This behavior is visualized through the confidence-based reliability diagram in Figure 4.

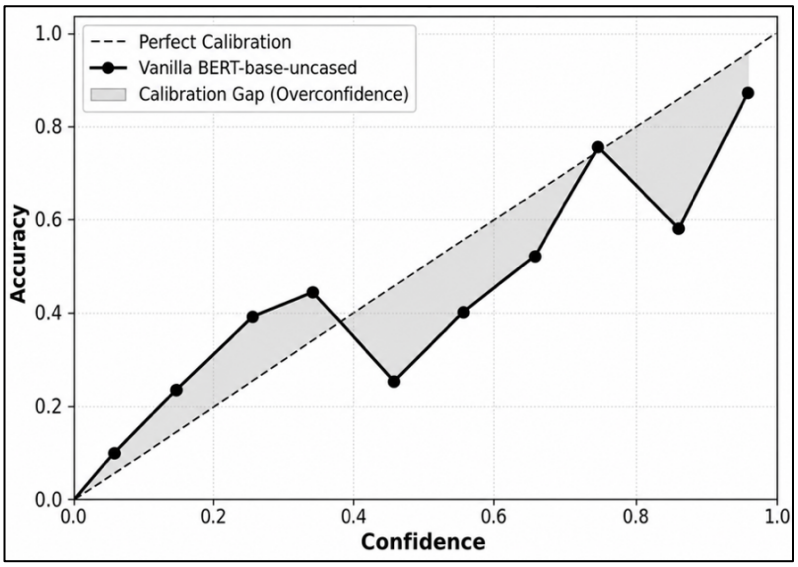

Figure 4. Confidence-Based Reliability Diagram of Vanilla BERT on the Dreaddit Test Set

Figure 4 shows a gap between the model curve and the perfect calibration line, especially in the high-confidence region. To make this interpretation more explicit, Table 7 reports the non-empty confidence bins used to compute the reliability diagram and ECE.

Table 7. Confidence-Bin Reliability Analysis of Vanilla BERT

Confidence Range	Count	Mean Confidence	Accuracy	Gap	ECE Contribution
0.5-0.6	45	0.5478	0.5556	+0.0078	0.0005
0.6-0.7	50	0.6564	0.5400	-0.1164	0.0081
0.7-0.8	61	0.7451	0.6721	-0.0730	0.0062
0.8-0.9	109	0.8554	0.6881	-0.1673	0.0255
0.9-1.0	450	0.9526	0.8822	-0.0704	0.0443

Table 7 indicates that the largest absolute mismatch occurred in the 0.8-0.9 confidence range, where mean confidence was 0.8554 but empirical accuracy was only 0.6881, producing an absolute gap of 0.1673. However, the largest contribution to ECE came from the 0.9-1.0 confidence range because this bin contained 450 of 715 test samples. This shows that overconfidence was not only present in isolated cases but was concentrated in the high-confidence region where the model assigned very strong confidence to a large portion of predictions.

3.4. Granular Error Analysis

The third objective of this study was to analyze the model's error patterns, particularly in cases where the model made incorrect predictions with high confidence. This analysis is important to explain why the model becomes overconfident and when the model fails. Based on the confusion matrix, the model produced 86 false positives and 64 false negatives. Among the false positives, 35 cases or 40.7% had confidence of at least 0.90, while among the false negatives, 18 cases or 28.1% had confidence of at least 0.90. This indicates that a meaningful portion of the errors were not low-confidence borderline mistakes, but high-confidence failures that require closer interpretation. A dominant-pattern screening of all false positive and false negative cases is summarized in Table 8.

Table 8. Systematic Breakdown of Error Patterns

Error Type	Dominant Pattern	Frequency	Percentage
False Positive	Relationship/family context	42 of 86	48.8%
False Positive	Financial/problem-solving cues	22 of 86	25.6%
False Positive	Negative emotional/conflict cues	16 of 86	18.6%
False Positive	Other ambiguous cues	6 of 86	7.0%
False Negative	Relationship/trauma narrative	33 of 64	51.6%
False Negative	Polite/constructive advice tone	18 of 64	28.1%
False Negative	Implicit/indirect distress expression	5 of 64	7.8%
False Negative	Other ambiguous cues	8 of 64	12.5%

Table 8 shows that false positives were most frequently associated with relationship or family narratives and financial or problem-solving cues, while false negatives were dominated by relationship or trauma narratives and polite or constructive advice-seeking tones. These results support the qualitative interpretation that the model often relied on surface lexical cues and could miss stress when it was expressed indirectly or in a calm, advice-oriented style.

Table 9 shows that overconfidence is not limited to one error direction. In false positive cases, the model is overly confident in predicting Stress when Not Stress texts contain negative words or narratives that resemble stress-related contexts. In false negative

cases, the model is overly confident in predicting Not Stress when Stress texts are written in a polite, constructive, or advice-seeking tone.

Table 9. Examples of High-Confidence Errors Produced by Vanilla BERT

Text Sample	True Label	Predicted Label	Confidence	Error Interpretation
"...I get pissed; my SO gets angry... she ends up breaking up with me. I am weary of this whole emotional drama..."	Not Stress	Stress	97.1%	False positive. The model was overly influenced by negative lexical cues such as "angry," "breaking up," and "emotional drama."
"...I WANT to pay back every penny... I just feel like this is too big of a loan... stuck in such a rut."	Not Stress	Stress	96.6%	False positive. The model overemphasized financial stress-related terms while overlooking the transactional and problem-solving context.
"...I failed here. She's talking about wanting to divorce me... looking into therapy. I really appreciate advice."	Stress	Not Stress	95.5%	False negative. The model failed to detect stress because polite and constructive expressions masked the underlying stress context.

These findings have important scientific implications. First, Transformer models may learn shortcuts based on emotional keywords rather than understanding the narrative context as a whole. Second, training with hard labels on a crowdsourced dataset may encourage the model to produce extreme probabilities even when the human labels themselves contain ambiguity. Third, global evaluations such as F1-score and ECE are not sufficient to explain the source of model errors. Granular analysis is required and, in the next stage, thematic diagnostics based on semantic groups will be needed to determine whether miscalibration is concentrated in specific topics.

The error-pattern counts are interpretive categories derived from the Dreaddit test set and the Vanilla BERT predictions in this study. Therefore, these findings should be interpreted within the scope of this dataset, model architecture, and experimental protocol. They should not be generalized as universal failure mechanisms of Transformer-based stress detection models without further validation on other datasets, model variants, and repeated experimental runs.

3.5. Discussion

This section discusses the implications of the classification, calibration, and error-analysis results. The findings should be interpreted within the scope of one Vanilla BERT baseline, one benchmark dataset, the standard Dreaddit split, and a single fixed random seed. Overall, the results of this study answer the main question formulated in the Introduction: whether a BERT model that is competitive in classification is also probabilistically reliable. The findings show that the answer is not entirely. Vanilla BERT achieved a competitive F1-score, but it still showed miscalibration with an ECE of 0.0847 and an ACE of 0.0880. Therefore, discriminative performance and probabilistic reliability should be treated as two different evaluation dimensions.

Compared with previous Dreaddit baselines, this study not only replicates BERT classification performance but also adds probabilistic reliability evaluation, confidence-bin analysis, and systematic error-pattern interpretation. However, the comparison with prior studies should be interpreted carefully because earlier baselines were not evaluated using a harmonized calibration protocol. In this sense, Table 5 supports the discriminative validity of the baseline, while Tables 6–9 and Figures 3–4 provide the reliability-oriented and error-oriented diagnosis contributed by this study.

The findings also indicate that global metrics such as F1-score, ECE, and ACE are not sufficient to explain why the model fails. The confidence-bin results show where the model becomes overconfident, while the error-pattern analysis explains how certain linguistic patterns, such as negative lexical cues, relationship or family narratives, financial cues, and polite or constructive tones, may contribute to false positive and false negative predictions. However, these patterns should be interpreted as findings from the Dreaddit test set and Vanilla BERT predictions in this study, not as universal failure mechanisms of all Transformer-based stress detection models.

The main contribution of these findings is to strengthen the shift from accuracy-oriented stress detection toward reliability-aware stress detection. Social media-based stress detection models need to be evaluated not only based on whether their predictions are correct or incorrect but also based on how faithfully their confidence represents the actual likelihood of correctness. Because this experiment used one Vanilla BERT architecture, the standard Dreaddit split, and a single fixed random seed, the results should be interpreted as a diagnostic reliability baseline rather than a comprehensive calibration benchmark. Future work should estimate run-to-run variability through repeated seeds and extend the analysis to calibrated, uncertainty-aware, or topic-aware model variants.

4. CONCLUSION

This study concludes that Vanilla BERT provides a competitive diagnostic baseline for text-based stress detection on the Dreaddit dataset, but its classification performance does not fully guarantee probabilistic reliability. The model achieved competitive discriminative performance; however, calibration evaluation, confidence-bin analysis, and error-pattern interpretation showed that its confidence estimates were not always aligned with empirical correctness. The findings support the need to evaluate stress detection models using both discriminative and probabilistic criteria, particularly when the labels represent text-based stress indications rather than clinical diagnoses. This study should therefore be interpreted as a diagnostic reliability baseline, not as a comprehensive calibration benchmark or a model ready for real-world clinical decision support. Future work should evaluate repeated-seed stability, apply calibration or uncertainty-aware methods, and extend thematic diagnostics to identify whether miscalibration is concentrated in specific semantic groups.

REFERENCES

[1] X. Sun, B. J. Li, H. Zhang, and G. Zhang, "Social media use for coping with stress and psychological adjustment: A transactional model of stress and coping perspective," *Front. Psychol.*, vol. 14, 2023, doi: 10.3389/fpsyg.2023.1140312.

[2] E. Turcan and K. McKeown, "Dreaddit: A Reddit Dataset for Stress Analysis in Social Media," in *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, Hong Kong: Association for Computational Linguistics, Oct. 2019, pp. 97–107. doi: 10.18653/v1/D19-6213.

[3] A. Pourkeyvan, R. Safa, and A. Sorourkhah, "Harnessing the Power of Hugging Face Transformers for Predicting Mental Health Disorders in Social Networks," *IEEE Access*, vol. 12, pp. 28025–28035, 2024, doi: 10.1109/ACCESS.2024.3366653.

[4] M. Sao and H. J. Lim, "MIRoBERTa: Mental Illness Text Classification With Transfer Learning on Subreddits," *IEEE Access*, vol. 12, pp. 197454–197466, 2024, doi: 10.1109/ACCESS.2024.3522465.

[5] A. Karamat, M. Imran, M. U. Yaseen, R. Bukhsh, S. Aslam, and N. Ashraf, "A Hybrid Transformer Architecture for Multiclass Mental Illness Prediction Using Social Media Text," *IEEE Access*, vol. 13, pp. 12148–12167, 2025, doi: 10.1109/ACCESS.2024.3519308.

[6] L. Ilias, S. Mouzakitis, and D. Askounis, "Calibration of Transformer-Based Models for Identifying Stress and Depression in Social Media," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 2, pp. 1979–1990, Apr. 2024, doi: 10.1109/TCSS.2023.3283009.

[7] N. Oryngozha, P. Shamoi, and A. Igali, "Detection and Analysis of Stress-Related Posts in Reddit's Acamedic Communities," *IEEE Access*, vol. 12, pp. 14932–14948, 2024, doi: 10.1109/ACCESS.2024.3357662.

[8] J. Gawlikowski *et al.*, "A survey of uncertainty in deep neural networks," *Artif. Intell. Rev.*, vol. 56, pp. 1513–1589, Oct. 2023, doi: 10.1007/s10462-023-10562-9.

[9] J. Geng, F. Cai, Y. Wang, H. Koeppl, P. Nakov, and I. Gurevych, "A Survey of Confidence Estimation and Calibration in Large Language Models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 6577–6595. doi: 10.18653/v1/2024.naacl-long.366.

[10] S. Roohi, R. Skarbez, and H. D. Nguyen, "Reliable uncertainty estimation in emotion recognition in conversation using conformal prediction framework," *Natural Language Processing*, vol. 31, no. 5, pp. 1163–1186, Sep. 2025, doi: 10.1017/nlp.2024.48.

[11] J.-Q. Yang, D.-C. Zhan, and L. Gan, "Beyond Probability Partitions: Calibrating Neural Networks with Semantic Aware Grouping Appendix," in *Advances in Neural Information Processing Systems*, New Orleans, Louisiana, USA: Neural Information Processing Systems Foundation, 2023, pp. 58448–58460. Accessed: May 01, 2026.

[12] D. Angelov, "Top2Vec: Distributed Representations of Topics," *arXiv preprint arXiv:2008.09470*, Aug. 2020, Accessed: May 17, 2026. [Online]. Available: https://arxiv.org/abs/2008.09470

[13] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, NSW, Australia: PMLR, 2017, pp. 1321–1330.

[14] J. Devlin, M.-W. Chang, and K. Lee, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[15] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?" in *Chinese Computational Linguistics*, Cham, Switzerland: Springer, 2019, pp. 194–206. doi: 10.1007/978-3-030-32381-3_16.