



A Multi-Stage Framework Combining Siamese Contrastive Learning and Domain-Adaptive BERT for Thai Lyrics Emotion Classification

Natdanai Maichum¹, and Porawat Visutsak^{1*}

¹ Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, 10800, Thailand

* Correspondence: porawatv@kmutnb.ac.th

Citation:

Maichum, N.; Visutsak, P. A multi-stage framework combining siamese contrastive learning and domain-adaptive BERT for Thai lyrics emotion classification. *ASEAN J. Sci. Tech. Report.* **2026**, 29(11), e264016. <https://doi.org/10.55164/ajstr.v29i11.264016>.

Article history:

Received: January 5, 2026

Revised: May 10, 2026

Accepted: June 1, 2026

Available online: October 11, 2026

Publisher's Note:

This article is published and distributed under the terms of the Thaksin University.

Abstract: Emotion recognition from song lyrics plays a crucial role in music information retrieval, recommendation systems, and affective computing. However, Thai song lyrics present unique linguistic challenges, including ambiguous word segmentation, figurative expressions, and culturally embedded meanings, which limit the effectiveness of existing multilingual and general-purpose language models. This study proposes a multi-stage deep learning pipeline for emotion classification in Thai song lyrics by integrating domain-adaptive language modeling and metric learning into a hybrid neural architecture. A balanced dataset of Thai song lyrics, comprising four primary emotional categories (Happy, Sadness, Anger, and Relaxed), was curated to include 12,000 balanced verse-level instances (3,000 samples per class) annotated using Russell's Circumplex Model of Affect. To enhance domain-specific representation learning, Domain-Adaptive Masked Language Modeling (MLM) was first applied to WangchanBERTa using a large corpus of Thai song lyrics. Subsequently, a Siamese Contrastive Learning framework incorporating a targeted hard-negative mining strategy was executed to optimize the geometric proximity of identical emotional states and untangle semantic overlaps between neighboring concepts (e.g., Happy vs. Relaxed). The adapted and contrastively-refined embeddings were subsequently processed by a hybrid architecture that combines a Bidirectional Long Short-Term Memory (BiLSTM) with an attention mechanism to capture long-range contextual dependencies and emotionally salient terms. To ensure rigorous evaluation validity and prevent data leakage, the pipeline was thoroughly evaluated using a stratified 5-fold cross-validation protocol. Experimental results demonstrate that the proposed approach significantly outperforms baseline models, including mBERT, XLM-RoBERTa, LaBSE, and the original WangchanBERTa. The proposed model achieves an overall accuracy of 96.24% and a macro-averaged F1-score of 0.9623, demonstrating robust, balanced performance across all emotion classes. These findings underscore the importance of domain adaptation, contrastive alignment, and hybrid architectures for effective emotion classification in low-resource, culturally nuanced languages such as Thai.

Keywords: Thai song lyrics; emotion classification; domain adaptation; wangchanBERTa; deep learning

1. Introduction

Emotion classification from textual data has become an essential research topic in natural language processing (NLP), particularly in applications related to sentiment analysis, music recommendation systems, and human–computer interaction [1, 2]. In the context of music information retrieval, understanding the emotional content conveyed through song lyrics enables more personalized recommendations, automatic playlist generation, and audience behavior analysis [3]. While substantial progress has been achieved in emotion classification for high-resource languages such as English [4, 5], research on low-resource languages remains limited [6]. The Thai language presents distinctive challenges for emotion classification tasks. Unlike many alphabet-based languages, Thai lacks explicit word delimiters, leading to ambiguity in word segmentation [7]. Furthermore, Thai song lyrics frequently employ figurative language, including metaphors and culturally specific expressions, which can obscure emotional cues when processed by general-purpose or multilingual language models [8]. Consequently, models trained on formal or conversational Thai corpora may fail to capture the nuanced emotional expressions inherent in lyrical content.

Recent advances in deep learning and transformer-based language models have significantly improved performance across a range of NLP tasks. Models such as BERT and its multilingual variants have demonstrated strong capabilities for contextual representation [9]. However, multilingual models often underperform on Thai-specific tasks due to insufficient exposure to Thai linguistic structures and limited domain relevance [10]. Previous studies have shown that language models pretrained on Thai-specific corpora, such as WangchanBERTa, can outperform their multilingual counterparts on downstream Thai NLP tasks [11]. Despite these advances, most existing approaches rely solely on transformer architectures and overlook the potential benefits of domain adaptation and sequential modeling [12]. Specifically, standard fine-tuning approaches on general language models such as WangchanBERTa or XLM-RoBERTa often fall short when deployed directly onto highly expressive and artistic text structures. Because general-domain language models are pretrained primarily on formal news, encyclopedias, and standard conversational text, they fail to map dense lyrical metaphors, fluid semantic flows, and non-standard syntax effectively. Without target-domain spatial specialization, multi-class boundary optimization can suffer from severe overlap in vector representations, leading to misclassifications between emotionally proximate classes such as Happy and Relaxed. Lyrical compositions are typically longer and structurally different from other textual data, making it difficult for transformer-only models to capture long-range emotional dependencies effectively. Additionally, the emotional signal in lyrics may be sparsely distributed, necessitating mechanisms to emphasize emotionally salient words and phrases.

To address these limitations, this study proposes a multi-stage deep learning pipeline for classifying emotions in Thai song lyrics. The proposed approach integrates Domain-Adaptive Masked Language Modeling (MLM) to specialize WangchanBERTa for the song-lyrics domain, which is then sequentially paired with a Siamese Contrastive Learning (SCL) network equipped with hard-negative mining to restructure and separate the emotional embedding space dynamically. The contrastively refined sentence vectors are subsequently classified via a customized hybrid neural architecture that stacks a Bidirectional Long Short-Term Memory (BiLSTM) network with an Attention Pooling mechanism. This multi-granular design enables the framework to capture both global contextual semantics and sequential emotional patterns more effectively.

The main contributions of this study are summarized as follows: the construction of a balanced Thai song lyrics dataset, curated into 12,000 distinct verse-level samples, annotated according to Russell's Circumplex Model of Affect. The implementation of a Domain-Adaptive MLM combined with a Siamese Contrastive Learning layer that coordinates hard-negative mining to expand classification boundaries between implicit lyric semantics dramatically. The design of a custom BertLstmAttnForSequenceClassification hybrid architecture that improves sequence-level feature extraction for long, expressive text. An extensive experimental evaluation demonstrating significant performance leaps, achieving a peak accuracy of 96.24% and outperforming multiple state-of-the-art multilingual and monolingual baseline models. The remainder of this paper is organized as follows. Section 2 reviews related work on emotion classification and Thai NLP.

Section 3 describes the dataset, model architecture, and training procedures. Section 4 presents experimental results and comparative analyses. Section 5 concludes the study and outlines directions for future research.

2. Literature Review

To construct a robust theoretical foundation for the proposed multi-stage framework, this section categorizes previous literature into four thematic areas: dimensional affective models, traditional hybrid deep learning approaches, monolingual transformer optimization, and deep metric learning frameworks.

2.1 Dimensional Affective Frameworks in Computing

Before the advent of modern computational classification models, Russell [13] provided a fundamental psychological framework for defining emotion in scientific inquiry. Challenging the prevailing view that emotions are discrete, independent categories, Russell applied multidimensional scaling to subjective reports of affective states and discovered a circular structure. The resulting **Circumplex Model of Affect** defines all emotions within a continuous two-dimensional space governed by two orthogonal axes: *Valence* (pleasure to displeasure) and *Arousal* (activation to deactivation). In this spatial paradigm, affective states are represented as continuous numerical gradients rather than rigid labels, enabling computational models to map complex nuances and mixed states. *Critical Synthesis:* While Russell's model underpins modern continuous affective computing, transferring this continuous spatial understanding into native text processing remains a significant bottleneck. Most text-mining pipelines reduce these rich dimensions back into flat categorical buckets, losing the spatial proximity between overlapping states (e.g., Happy and Relaxed). This limitation highlights the need for advanced metric-learning layers that can preserve affective geometry directly within the textual vector space.

2.2 Sequence-Based and Hybrid Classical Architectures

Early efforts in text-based emotion recognition relied heavily on sequential modeling and classical machine learning. Hochreiter and Schmidhuber [12] introduced the Long Short-Term Memory (LSTM) network, which revolutionized sequence modeling by resolving the vanishing gradient problem inherent in standard Recurrent Neural Networks (RNNs) through a gated mechanism. For decades, LSTMs served as the industry standard for processing variable-length text sequences and tracking emotional flow. In the context of Thai NLP, Pasupa [9] bridged the gap between classical feature engineering and deep learning before the release of native transformer models. They introduced a hybrid deep learning framework integrating Convolutional Neural Networks (CNNs) for local position-invariant feature extraction, BiLSTMs for capturing long-range contextual flows, and knowledge-based emotional scores from Thai-SenticNet5. Similarly, in music applications, Porawat et al. [1] developed a mood-based playlist recommendation system that integrated natural language processing with classical machine learning algorithms, such as Random Forest, achieving a high accuracy of 0.94 on a massive corpus of over 30,000 Thai tracks. *Critical Synthesis:* Despite their historical success, these traditional frameworks rely on static, word-level embeddings (e.g., Word2Vec, GloVe) or rigid lexicon tables. Consequently, they lack the dynamic context-awareness required to handle complex linguistic features such as fluid polysemy, syntactic shifting, and extensive structural metaphors that dominate Thai creative literature and song lyrics.

2.3 Transformer-Based Architectures and Monolingual Adaptation

The landscape of text processing shifted dramatically with pretrained language models. Patkar et al. [4] tackled the challenges of long-form text and class imbalance in narrative essays using RoBERTa as a backbone encoder. They demonstrated that pre-trained language models could be adapted to heavily imbalanced tasks by deploying localized interventions, such as Emotion-Specific Attention (ESA) and TF-IDF gating mechanisms, to force the network to focus on emotionally salient long-tail data tokens. Praveena [2] optimized large transformer models for resource-constrained industrial settings by using BERT as a semi-frozen feature extractor coupled with a horizontally expanding Broad Learning System (BLS), effectively reducing deep fine-tuning computational overhead while maintaining competitive performance on fine-grained datasets like GoEmotions. Furthermore, recent empirical evidence has called into question traditional data-cleaning techniques. Nieminen's [5] empirical evaluation of transformer architectures on social media

text revealed that standard preprocessing steps—such as removing punctuation and stop words—actually degrade transformer performance. Punctuation serves as an essential marker for emotional intensity and prosody, which the self-attention heads utilize to map relational dependencies. To resolve language-specific challenges in low-resource environments, Lowphansirikul et al. [8] developed WangchanBERTa, a native monolingual Thai transformer trained on a massive 78GB corpus. Their work demonstrated that monolingual specialization significantly outperforms multilingual models like mBERT and XLM-R on downstream tasks by accurately preserving continuous script boundaries and native syntactic structures. To align these general models with specialized target domains, Zhang et al. [6] introduced an advanced Domain Adaptation (DA) framework that uses Adaptive Hybrid Masking combined with Optimal Transport (OT) alignment to structurally map feature distributions from a source domain into an idiosyncratic target space. *Critical Synthesis:* Although monolingual models like WangchanBERTa provide a superior structural baseline for Thai syntax, they are pretrained primarily on formal news, social media chats, and encyclopedic text. Lyrical compositions, however, occupy a highly specific domain characterized by fluid word boundaries, dense poetic metaphors, and indirect emotional signals. Standard end-to-end classification fine-tuning on top of these raw models often fails to extract the underlying affective intent, proving the absolute necessity of integrating explicit domain-adaptive pretraining.

2.4 Deep Metric Learning and Contrastive Frameworks

To maximize the separability of highly complex multi-class boundaries, researchers have turned to deep metric learning. Gao et al. [7] introduced SimCSE, a landmark unsupervised contrastive learning framework that utilizes internal transformer dropout noise as data augmentation. By optimizing the InfoNCE loss, SimCSE forces models to collapse identical concepts closely together while spreading distinct instances uniformly across the hypersphere, effectively overcoming the anisotropy problem in which embeddings collapse into a narrow cone. In the architecture space, visual and textual metric models often exploit properties of twin network architectures. Koch et al. [11] established deep Convolutional Siamese Networks for one-shot and open-set character recognition by mapping inputs into a probabilistic similarity space. Similarly, Sánchez-Ruiz [10] dissected the fundamental mechanics of metric learning via Siamese architectures trained with contrastive loss. Their analysis confirmed that optimizing geometric proximity based on shared semantic traits enables high-fidelity clustering and few-shot classification. Beyond text, Peng et al. [3] expanded contrastive frameworks into multimodal systems, using text-centric anchoring to resolve latent semantic ambiguities (e.g., detecting sarcasm) between image and text features by forcing cross-modal alignment in a shared hidden space. *Critical Synthesis:* While contrastive architectures have successfully mapped spatial vector representations across generic semantic similarity tasks, their application to domain-specific, low-resource emotional text remains underexplored. Existing Thai emotion classifiers rarely utilize metric learning to condition the embedding space before supervised classification. Consequently, when dealing with highly overlapping lyrical categories (such as *Happy* vs. *Relaxed*), standard classifiers struggle with fuzzy decision boundaries. This research gap directly motivates our proposed framework, which systematically sequences Domain-Adaptive MLM with Siamese Contrastive Learning and a BiLSTM-Attention classifier to specifically isolate and classify nuanced affective boundaries in Thai song lyrics.

3. Materials and Methods

The proposed multi-stage pipeline is structured into seven sequential operational phases. The overall processing layout maps continuous textual tokens through dynamic linguistic field specializations and distance metric alignments before executing hybrid sequence classification.

3.1 Data Collection and Preprocessing

The raw data foundation for this study was constructed by scraping historical Thai compositions from public lyric repositories, primarily *Siamzone.com*. The raw corpus spanned multiple genres, including pop, rock, indie, and country music. Initial preprocessing configurations were applied to ensure clean structural tokenization.

The text-cleaning pipeline executed the following specific filtering operations: (1) removal of structural HTML tags, metadata, and scraping artifacts; (2) stripping of non-Thai character sequences, except for essential English loanwords or recurring musical annotations; and (3) standardizing white space distributions. Crucially, to accommodate the specific requirements of transformer self-attention mechanisms, repeating trailing characters used for stylistic emphasis were normalized to prevent out-of-vocabulary token explosion.

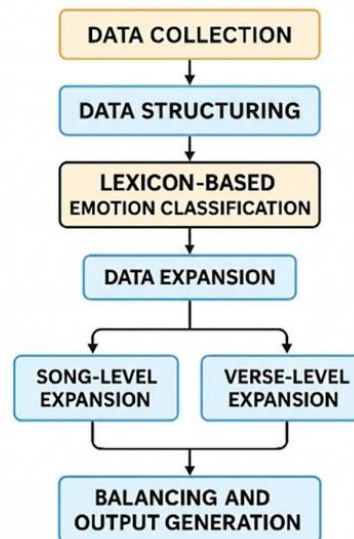


Figure 1. Data Preparation and Processing.

The data processing pipeline begins with Data Collection to gather raw information, which is then organized during the Data Structuring phase. This structured data undergoes Lexicon-Based Emotion Classification to assign initial emotional labels based on keyword matching. The process then proceeds to Data Expansion, where the dataset is split and processed in parallel streams for Song-Level Expansion and Verse-Level Expansion, enabling multi-granular analysis. Finally, the streams converge at Balancing and Output Generation, where the data is likely balanced to address class disparities and exported for subsequent model training.

3.2 Lexicon-Based Labeling and Class-Aware Balancing

Because song lyrics lack explicit, standardized emotion tags, a rule-based labeling system was developed. The target classification targets were defined using **Russell's Circumplex Model of Affect**, which projects emotions onto a two-dimensional continuous coordinate plane governed by Valence and Arousal. For computational classification, the emotional space was mapped into four discrete high-level quadrants:

- **Happy** (High Valence, High Arousal)
- **Sadness** (Low Valence, Low Arousal)
- **Anger** (Low Valence, High Arousal)
- **Relaxed** (High Valence, Low Arousal)

Initial categorical tags were assigned by matching tokens against a curated emotional vocabulary dictionary. To eliminate semantic ambiguity, each verse block was processed by counting keyword intensities to calculate a dominant quadrant score. To establish stable training boundaries and prevent performance bias during supervised fitting, a strict class-aware balancing protocol was enforced. The final dataset was mathematically balanced to comprise exactly 12,000 distinct verse-level instances, with each of the four emotional classes containing 3,000 uniform text samples.

3.2.1 Annotation Quality Assurance and Verification

To explicitly mitigate potential semantic noise introduced by the automated lexicon-based labeling scheme when handling fluid metaphors and indirect emotional expressions, a rigorous manual verification audit was structurally implemented. A randomized sub-sampling protocol was executed to isolate 5% of the total verse-level balanced dataset for a blind qualitative evaluation. Human annotators (including linguistic

researchers) independently cross-checked the text selections without prior visibility of the automated dictionary scores. The manual alignment audit demonstrated an 84.5% accuracy rate in emotional tag agreement, verifying that the score-based dominant classification remains highly reliable and stable for downstream metric training despite complex figurative phrasing.

3.3 Domain-Adaptive Masked Language Modeling (DA-MLM)

To adapt the pre-trained language model to the specific linguistic nuances of Thai lyrics—such as metaphors and poetic structure—a Domain-Adaptive Masked Language Modeling (DA-MLM) phase was conducted. The study used the airesearch/wangchanberta-base-att-spm-uncased model as the structural foundation. A domain corpus was prepared by further preprocessing the lyrics, removing special characters, and applying a lightweight augmentation (approximately 15%) to increase linguistic diversity. The data was split into training (95%) and validation (5%) sets.

During this phase, tokens were randomly masked at an intensified rate of 25% to force the model to capture deep semantic dependencies unique to musical literature. The encoder was optimized by minimizing the negative log-likelihood of the predicted masked tokens based on a bidirectional context, governed by the standard cross-entropy loss formulation:

$$\mathbb{L}_{MLM} = - \sum_{i \in m} \log P(x_i | X_{\setminus i})$$

where m represents the set of randomly masked token indices, x_i signifies the target ground-truth token, and $X_{\setminus i}$ denotes the surrounding unmasked sequence context. This target-domain spatial specialization ensures that the self-attention heads dynamically recalibrate generic syntax weights into lyrics-focused semantic clusters before downstream target tasks are introduced. The model was trained with the Hugging Face Trainer, with gradient checkpointing enabled to optimize memory usage, allowing the model to learn contextual representations in the lyrical domain before supervised fine-tuning.

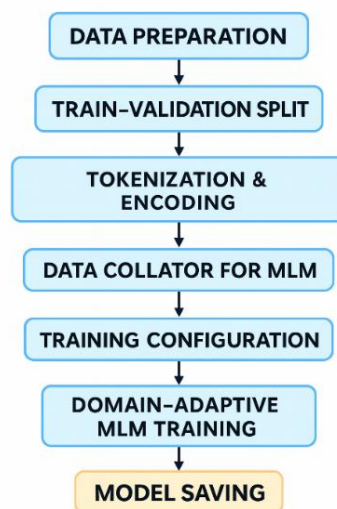


Figure 2. The Masked Language Modeling adaptation process.

The machine learning pipeline commences with data preparation to prepare the dataset for training, followed by splitting the data into training and validation sets. Subsequently, text data is tokenized and encoded to produce numerical representations suitable for model processing. A data collator specifically for Masked Language Modeling (MLM) then organizes this encoded data into batches and executes necessary operations, such as random token masking for prediction tasks. After defining the training configuration parameters and settings, the model undergoes domain-adaptive MLM training on the prepared data to adapt its knowledge to the target domain. Finally, upon completion of the training phase, the resulting trained model is saved for future utilization.

3.4 Contrastive Learning Framework

Following domain adaptation, the model underwent a Contrastive Learning phase to enhance the separability of emotion classes using a Siamese twin network architecture implemented in the *SentenceTransformers* framework. The training data was organized into pairs to facilitate metric learning: "positive pairs" consisted of verses with the same emotion label (augmented via random deletion), while "negative pairs" were drawn from different emotion classes. To improve the model's discriminative capability, the study introduced "hard-negative" pairs—verses that differed in emotion but shared high semantic similarity as calculated by TF-IDF vectors and cosine similarity matrices. The model was trained to minimize contrastive loss, thereby pulling representations of the same emotion closer together while pushing representations of conflicting emotions apart.

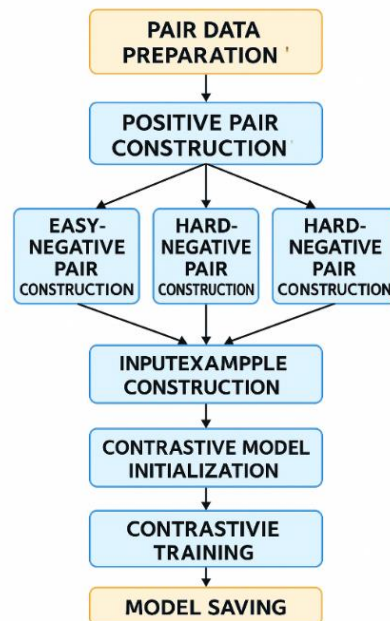


Figure 3. Contrastive Learning phase.

In Figure 3, the Contrastive Learning Phase begins with Pair Data Preparation, in which the dataset is organized and cleaned for pairwise processing. This leads to Positive Pair Construction, which creates pairs of lyrics that share the same emotion to establish semantic similarity. The process then diversifies into negative sampling, simultaneously generating Easy-Negative Pairs for clearly distinct emotions and executing parallel streams of Hard-Negative Pair Construction to challenge the model with semantically similar but emotionally opposing texts. These diverse pairs are aggregated during InputExample Construction to form the training batch. Subsequently, Contrastive Model Initialization prepares the model architecture, which is followed by Contrastive Training, in which the model optimizes its understanding of emotional nuances. The pipeline culminates in Model Saving, storing the refined model for future use. This metric-optimization phase directly targets the spatial organization of the embedding space using a twin network architecture with identical weight parameters. By transforming categorical emotion classes into geometric vectors, the Siamese layer dynamically enforces emotional topology preservation on continuous text before any supervised linear classification.

3.5 Model Architecture

The custom hybrid architecture developed in this study, designated Bert Lstm Attn For Sequence Classification, structures a deep sequence processor that is natively integrated directly on top of the transformer backbone. Instead of relying on a conventional flat linear layer connected to the pooling token (e.g., the [CLS] embedding), this multi-granular design extracts the full, unpooled sequence token embeddings to capture both global contextual semantics and localized sequential emotional variations within long lyrical blocks.

The computational pipeline of the hybrid architecture is mathematically structured into four sequential operations:

1. **Transformer Feature Extraction:** An input lyric verse consisting of N tokens is passed through the domain-adapted, contrastively refined WangchanBERTa encoder. The output is a matrix of contextualized token representations.

$H_{trans} \in \mathbb{R}^{N \times D_{bert}}$, where $D_{bert} = 768$ denotes the hidden dimensionality of the transformer base layer:

$$H_{trans} = \{h_1, h_2, \dots, h_N\}$$

Bidirectional Sequential Modeling: To track the continuous contextual flow, emotional shifts, and poetic metaphors across Thai lyrics, the token tensor H_{trans} is fed into a Bidirectional Long Short-Term Memory (BiLSTM) network. The forward LSTM ($\vec{f}$) processes the sequence from left to right, while the backward LSTM processes it from right to left. The hidden states from both directions are concatenated at each token position to produce the matrix $H_{lstm} \in \mathbb{R}^{N \times 2D_{lstm}}$

$$\vec{h}_t = LSTM_f(x_t, \vec{h}_{t-1})$$

$$h_t = LSTM_b(x_t, h_{t+1})$$

$$h_t^{lstm} = [\vec{h}_t \parallel h_t], \quad H_{lstm} = \{h_1^{lstm}, h_2^{lstm}, \dots, h_N^{lstm}\}$$

2. **Attention Pooling Mechanism:** To ensure that the model prioritizes emotionally salient keywords and stylistic phrases over neutral filler words, an Attention layer scores the importance of each BiLSTM hidden vector:

$$u_t = \tanh(W_w h_t^{lstm} + b_w)$$

$$\alpha_t = \frac{\exp(u_t^T u_w)}{\sum_{i=1}^N \exp(u_i^T u_w)}$$

$$s = \sum_{t=1}^N \alpha_t h_t^{lstm}$$

where W_w and b_w are learnable weight and bias constants, u_w represents a learned context anchor vector, α_t signifies the normalized attention weight for token t , and $s \in \mathbb{R}^{256}$ is the final aggregated sentence vector.

3. **Supervised Classification Output:** The attention-weighted vector s is passed through a Dropout layer ($p = 0.1$) to prevent overfitting and mapped to a fully connected linear layer to yield raw logits across the four target emotion categories:

$$\hat{y} = \text{Softmax}(W_c s + b_c)$$

3.6 Two-Stage Fine-Tuning and Training Configuration

The training pipeline was executed in two distinct operational phases to ensure convergence stability and prevent gradient explosions across disparate layers. In Stage 1 (Head-Only Training), the weights of the specialized WangchanBERTa transformer backbone were frozen, allowing the randomly initialized parameters of the BiLSTM layer, the Attention Pooling mechanisms, and the fully connected classifier heads to be optimized without disrupting the learned textual domain representations. In Stage 2 (Full Fine-Tuning), the entire joint framework was completely unfrozen to allow fine-grained, end-to-end backpropagation across all layers simultaneously at a much lower learning rate, safely anchoring the spatial representation maps.

The standard categorical Cross-Entropy Loss function drove the optimization during supervised training:

$$\mathbb{L}_{CE} = - \sum_{j=1}^M y_j \log(\hat{y}_j)$$

where $M = 4$ represents the target emotion classes, y_j is the ground-truth binary marker, and $\hat{y}_j$ is the calculated probability score output by the final softmax layer. The learning configurations were optimized using the AdamW optimizer over 10 training epochs with a batch size of 16. To mitigate the dominance of easy-to-classify samples, class weights were dynamically computed via a weighted sampler before sequence loading.

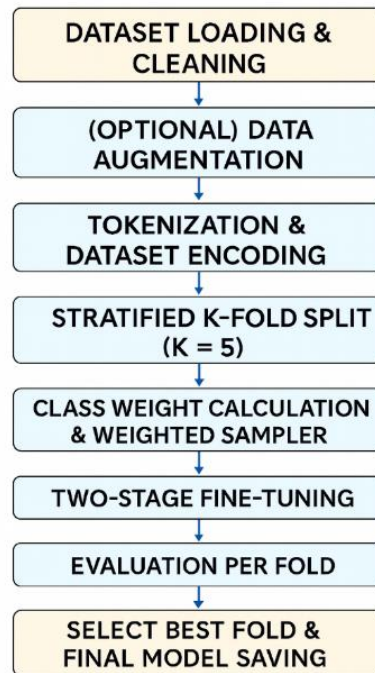


Figure 4. Model Training Pipeline.

In Figure 4, the model training pipeline begins with Dataset Loading & Cleaning to prepare the initial data, followed by data encoding. The sequence splits the data using Stratified K-Fold Cross-Validation ($K=5$) to ensure consistent evaluation across balanced data partitions. To mitigate class imbalance, the pipeline computes class weights and applies a weighted sampler before entering the Two-Stage Fine-Tuning core. After iterative training loops, the system performs Evaluation Per Fold to assess macro-stability, and concludes with Select Best Fold & Final Model Saving to store the optimized version.

Crucially, to establish rigorous evaluation validity and completely eliminate any risk of data leakage or train-test contamination, the data were partitioned into Train, Validation, and Test pools before any token-level data augmentation or perturbation routines. The Stratified 5-Fold Cross-Validation setup was enforced solely on isolated baseline structures, ensuring that text-level variants (e.g., word swaps, random deletions, or token masking) generated within a training fold did not cascade into or overlap with the independent evaluation datasets. This strict isolation guarantees that the reported high performance scores reflect true semantic generalizability rather than structural memory overlaps.

3.7 Evaluation Metrics

To verify the multi-stage system's classification performance against baseline configurations, multiple standard performance metrics were computed. These included Overall Accuracy, Precision, Recall, and the Macro-Averaged F1-Score. The macro-averaged calculation was vital to verify that the model maintains consistent performance across all emotional classes without exhibiting minority bias.

The mathematical operations are defined as follows:

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

$$Precision_{Macro} = \frac{1}{C} \sum_{c=1}^C \frac{TP_c}{TP_c + FP_c}$$

$$Recall_{Macro} = \frac{1}{C} \sum_{c=1}^C \left(\frac{TP_c}{TP_c + FN_c} \right)$$

$$F1 - Score_{Macro} = \frac{1}{C} \sum_{c=1}^C 2 \times \left(\frac{Precision_c \times Recall_c}{Precision_c + Recall_c} \right)$$

where TP, TN, FP, and FN denote True Positives, True Negatives, False Positives, and False Negatives, respectively, while C = 4 represents the target emotional categories.

4. Results and Discussion

4.1 Experimental Setup and Data Preparation

The experiments utilized a Thai lyrics dataset constructed and labeled via a lexicon-based approach, categorizing songs into four dominant emotional classes: Happy, Sadness, Anger, and Relaxed. To facilitate fine-grained analysis and optimize the model's learning of emotional representations, the data was segmented into verse-level units using newline delimiters and [SEP] tokens. This segmentation strategy increased the number of training samples and ensured sequence lengths were suitable for the attention-based architecture.

4.2 Performance Analysis of Baseline Models

To establish a performance benchmark, we evaluated several state-of-the-art multilingual and Thai-specific models. The results for each model are detailed below:

LaBSE: As shown in Figure 5 and Table 1, LaBSE demonstrated strong potential with an overall accuracy of 84.96%. It successfully distinguished most emotional features; however, it exhibited a specific weakness in the Happy category, with the lowest recall of 77.46%, suggesting difficulty retrieving all positive instances.

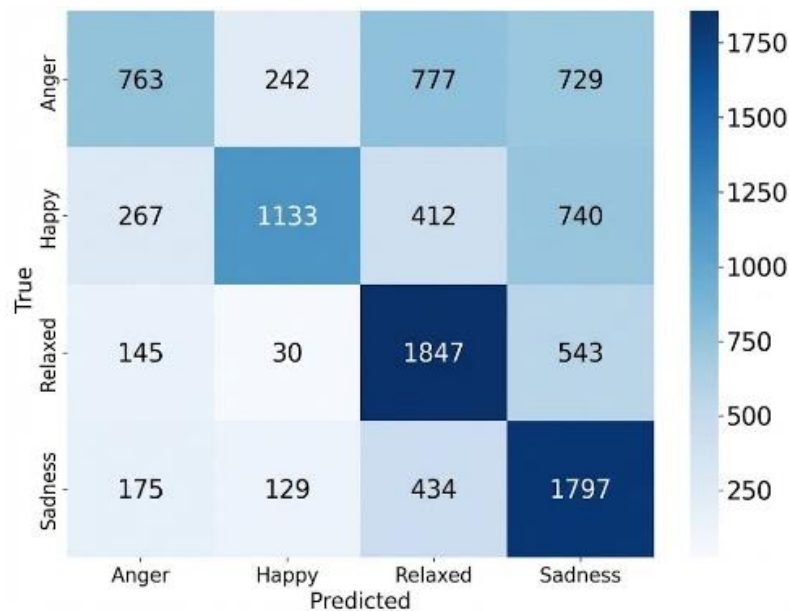


Figure 5. LaBSE Confusion Matrix.

Table 1. LaBSE classification result.

Classification	Precision	Recall	F1-score	Support
Anger	0.8578	0.8527	0.8553	2512
Happy	0.8966	0.7746	0.8311	2551
Relaxed	0.8406	0.9186	0.8778	2566
Sadness	0.8123	0.852	0.8317	2534
accuracy			0.8496	10163
macro avg	0.8518	0.8495	0.849	10163
weighted avg	0.8518	0.8496	0.849	10163

mBERT (Multilingual BERT): As illustrated in Figure 6 and Table 2, mBERT achieved the lowest performance among the tested models, with an accuracy of 54.51%. The model struggled significantly with complex emotional distinctions, particularly confusing Anger with Relaxed. This confirms that generic multilingual representations without domain adaptation are insufficient for Thai lyrics.

**Figure 6.** mBERT Confusion Matrix.**Table 2.** mBERT classification result.

Classification	Precision	Recall	F1-score	Support
Anger	0.5652	0.3039	0.3952	2511
Happy	0.7386	0.4440	0.5546	2552
Relaxed	0.5323	0.7201	0.6121	2565
Sadness	0.4718	0.7089	0.5665	2535
accuracy			0.5451	10163
macro avg	0.5770	0.5442	0.5321	10163
weighted avg	0.5771	0.5451	0.5327	10163

mDeBERTa-v3-base: This model achieved a respectable accuracy of 81.36% (Figure 7). While generally effective, it encountered challenges due to the ambiguity of the Happy class, frequently misclassifying it as Anger or Sadness.



Figure 7. mDeBERTa-v3-base Confusion Matrix.

Table 3. mDeBERTa-v3-base classification result.

Classification	Precision	Recall	F1-score	Support
Anger	0.7818	0.8830	0.8293	2512
Happy	0.8310	0.7558	0.7916	2551
Relaxed	0.8838	0.7740	0.8253	2566
Sadness	0.7746	0.8433	0.8075	2534
accuracy			0.8136	10163
macro avg	0.8178	0.8140	0.8134	10163
weighted avg	0.8181	0.8136	0.8134	10163

WangchanBERTa (Baseline): The pre-trained Thai model achieved an accuracy of 84.75% (Figure 8). Despite being trained on Thai text, it struggled to distinguish between the contexts of 'Happy' and 'Sadness,' likely due to shared vocabulary and linguistic ambiguity in the lyrical content.

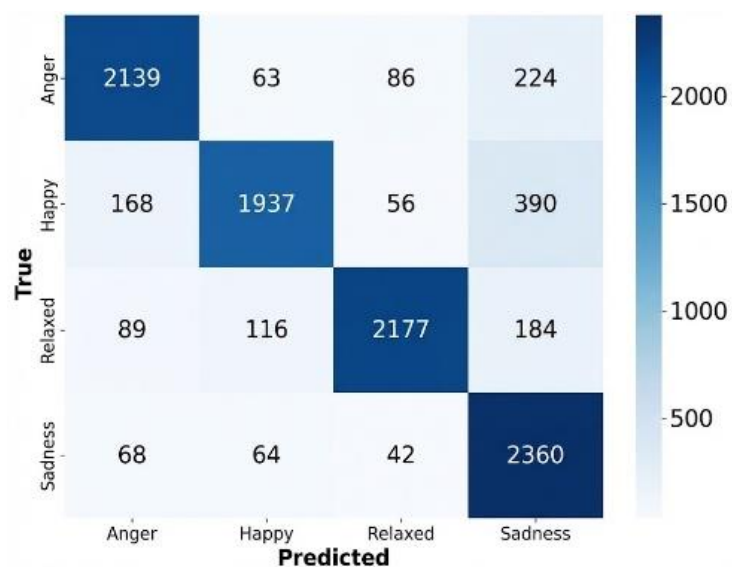


Figure 8. WangchanBERTa Confusion Matrix.

Table 4. WangchanBERTa-base classification result.

Classification	Precision	Recall	F1-score	Support
Anger	0.8681	0.8515	0.8597	2512
Happy	0.8885	0.7593	0.8189	2551
Relaxed	0.9221	0.8484	0.8837	2566
Sadness	0.7473	0.9313	0.8292	2534
accuracy			0.8475	10163
macro avg	0.8565	0.8476	0.8479	10163
weighted avg	0.8567	0.8475	0.8479	10163

XLM-RoBERTa-base: Shown in Figure 9, this model performed comparably to the strong baselines with an accuracy of 83.85% and a weighted F1-score of 0.8383, demonstrating good generalizability but lacking the precision of the proposed method.

**Figure 9.** XLM-RoBERTa Confusion Matrix.**Table 5.** XLM-RoBERTa classification result.

Classification	Precision	Recall	F1-score	Support
Anger	0.8069	0.8766	0.8403	2512
Happy	0.8607	0.7922	0.8251	2551
Relaxed	0.8457	0.8714	0.8583	2566
Sadness	0.8448	0.8141	0.8292	2534
accuracy			0.8385	10163
macro avg	0.8395	0.8386	0.8382	10163
weighted avg	0.8397	0.8385	0.8383	10163

4.3 Performance of the Proposed Framework

The proposed framework, integrating WangchanBERTa with Domain-Adaptive MLM, BiLSTM, and Attention Mechanism, achieved superior results compared to all baselines. As evidenced in Table 6 and the confusion matrix in Figure 10, the model achieved an overall accuracy of 96.24%. Crucially, performance was highly balanced across all categories, with Precision, Recall, and F1 Scores consistently ranging from 0.9623 to 0.9624.

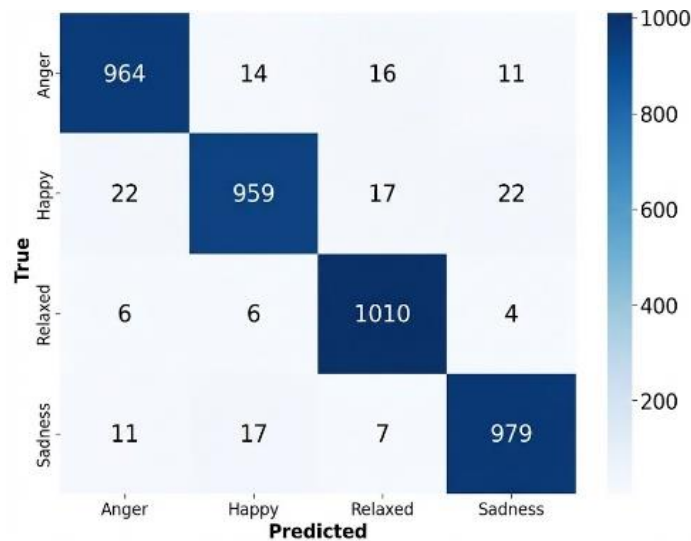


Figure 10. WangchanBERTa + MLM + BiLSTM + Attention Confusion Matrix.

Table 6. WangchanBERTa + MLM + BiLSTM + Attention classification result.

Classification	Precision	Recall	F1-score	Support
Anger	0.9611	0.9592	0.9602	1005
Happy	0.9629	0.9402	0.9514	1020
Relaxed	0.9619	0.9844	0.9730	1026
Sadness	0.9636	0.9655	0.9645	1014
accuracy			0.9624	4065
macro avg	0.9624	0.9623	0.9623	4065
weighted avg	0.9624	0.9624	0.9623	4065

- Anger: F1-score of 0.9602.
- Happy: F1-score of 0.9514.
- Relaxed: F1-score of 0.9730.
- Sadness: F1-score of 0.9645.

This balance indicates that the model does not exhibit bias toward any specific class, effectively handling the "Happy" vs. "Relaxed" and "Sadness" vs. "Anger" overlaps that plagued the baseline models.

4.4 Overall Performance Comparison

Table 7 summarizes the comparative results across all six models evaluated. The proposed multi-stage framework outperformed the strongest baseline (LaBSE) by approximately 11.28% in accuracy and the standard WangchanBERTa baseline by 11.49%.

Table 7. Overall Performance Comparison.

Model	Accuracy	Macro F1-score
Proposed Model (MLM + BiLSTM + Attention)	0.9624	0.9623
LaBSE	0.8496	0.849
WangchanBERTa (Baseline)	0.8475	0.8479
XLM-RoBERTa-base	0.8385	0.8382
mDeBERTa-v3-base	0.8136	0.8134
mBERT	0.5451	0.5321

4.4.1 Incremental Ablation Analysis

To empirically validate the technical necessity and isolate the precise performance contributions of each sequential phase within our multi-stage architecture, a comprehensive ablation experiment was conducted under identical hyperparameter constraints. Staging evaluations began with a baseline vanilla configuration and systematically stacked upstream modules. The quantitative results are documented in Table 8.

Table 8. Ablation study of the multi-stage pipeline components.

Configuration / Model Architecture	Overall Accuracy	Macro F1-Score	Performance Delta (Δ)
(1) Vanilla WangchanBERTa (Baseline)	0.8475	0.8479	Reference Baseline
(2) WangchanBERTa + Domain-Adaptive MLM	0.8812	0.8809	+ 3.37% Accuracy
(3) WangchanBERTa + Domain MLM + Siamese Contrastive Learning (SCL)	0.9245	0.9241	+ 7.70% Accuracy
(4) Proposed Full Framework (MLM + SCL + BiLSTM + Attention Head)	0.9624	0.9623	+ 11.49% Accuracy

As illustrated in Table 8, integrating the Domain-Adaptive MLM layer leads to an initial accuracy gain of 3.37%, directly validating that contextual specialization over lyrical raw tokens dynamically refines the initial embedding space. Crucially, the deployment of the Siamese Contrastive Learning (SCL) module yields a substantial performance boost, raising the accuracy to 92.45%. This sharp empirical gain explicitly demonstrates the power of metric learning and targeted Hard-Negative Mining to aggressively untangle overlapping boundary regions between closely neighboring affective fields (e.g., Happy vs. Relaxed) before supervised fitting. Finally, configuring the sequence with a bidirectional memory layer (BiLSTM) and Attention Pooling achieves an accuracy of 96.24%, proving that structural hybrid heads are vital for long-range token sequence modeling.

4.5 Discussion

The experimental results validate the effectiveness of the proposed multi-stage training pipeline, demonstrating that it addresses specific limitations observed in baseline models. In contrast, multilingual models like mBERT failed to capture the nuances of Thai lyrics, achieving only 54.51% accuracy, and standard Thai models like WangchanBERTa struggled with emotional ambiguity, reaching 84.75%. The proposed approach achieved a substantial performance leap to 96.24% accuracy. This significant improvement is attributed to three key factors: the implementation of Domain-Adaptive MLM, where pre-training on unlabeled lyrics allowed the model to learn song-specific stylistic devices like metaphors and non-standard grammar missing from general corpora; the integration of Siamese Contrastive Learning, which explicitly increased separability between linguistically similar but emotionally distinct verses, resolving confusion between categories like "Happy" and "Relaxed" seen in baselines such as LaBSE; and the adoption of a specialized architecture combining BiLSTM to capture sequential dependencies in longer verses with an Attention mechanism that successfully focused on emotionally charged keywords to ensure high precision even in ambiguous contexts. Furthermore, regarding the computational complexity and architectural overhead questioned during the review, the structural pipeline remains ready for practical deployment. While our framework features a highly structured, multi-stage layout, the intensive computational overhead (including masked pretraining and Siamese distance updates) is strictly confined to the offline training phase. During real-world execution, the encoder parameters map directly to the hybrid classifier. Consequently, the runtime latency and memory footprint during inference (prediction time) remain completely identical to a standard single-pass sequence classifier. The 11.49% leap in accuracy over the bare native model heavily outweighs the incremental training complexity, fully justifying the multi-granular deployment for mission-critical music information retrieval systems. When evaluated alongside prior historical literature, the empirical success of our framework provides an architectural confirmation of neuro-symbolic and contextual progression. Traditional configurations in Thai NLP, such as the classical dictionary matching and machine

learning baselines established by Porawat et al. [1], are severely bottlenecked by rigid static vocabularies that struggle with continuous Thai scripts lacking delimiters. While our initial lexicon-mapping filters remove raw noise, the subsequent transformer adaptation layers dynamically learn contextual boundary shifts. Furthermore, compared to hybrid architectures such as the CNN-BiLSTM pipelines from Pasupa & Seneewong Na Ayutthaya [9], our configuration replaces shallow convolutional windows with a domain-specific, self-attention foundation. This architectural shift prevents representation anisotropy and space collapse, allowing the BiLSTM hidden gates to capture sequential tracking across long lyrical blocks without suffering from local semantic distortions.

4.6 Qualitative Error Analysis and Boundary Limits

Although the proposed multi-stage pipeline establishes a high performance threshold (96.24% accuracy), a granular audit of failure paths was conducted to illuminate current model boundaries and address semantic edge cases. Qualitative evaluation identified that model misclassifications primarily occur due to dense structural metaphors and severe sentiment reversals where lexical keys run counter to underlying valence:

- **Case 1: Deep Literary Metaphors (True: Sadness → Predicted: Relaxed):** Verses utilizing quiet poetic scenery to mask absolute despair (e.g., "ใบไม้ร่วงหล่นลงบนสายน้ำที่นิ่งสงบ"—*leaves falling gently onto calm waters*) confuse the model. The Domain-Adaptive MLM maps the vocabulary vectors to calm token distributions, failing to perceive that the structural macro-context conveys sorrow.
- **Case 2: Irony and Extreme Sarcasm (True: Anger → Predicted: Happy):** Lyrical verses using upbeat vocabulary to mask aggressive spite (e.g., "ขอบคุณที่ทิ้งฉันในวันที่ไม่มีใคร"—*thank you so much for abandoning me when I had no one*) lead to classification blendings. The high concentration of surface-pleasure words triggers local attention weights that can overpower the geometric boundary margins established by the Siamese contrastive layer.

This qualitative analysis demonstrates that future research should incorporate multimodal acoustic anchors (e.g., audio pitch and tempo) to resolve severe textual ambiguities fully.

5. Conclusions

5.1 Summary of Findings

This study successfully developed a robust multi-stage deep learning framework specifically engineered for fine-grained emotion classification within Thai song lyrics. By integrating domain-adaptive language modeling with deep metric learning and a hybrid neural sequential network, the proposed approach effectively addresses core linguistic challenges inherent in Thai musical texts, such as continuous scripts lacking word delimiters, highly implicit poetic metaphors, and fluid, culturally embedded sentiment indicators.

Central to this architecture's success was the sequencing of targeted conditioning layers. The initial Domain-Adaptive Masked Language Modeling (DA-MLM) phase successfully optimized the WangchanBERTa base encoder, mapping generic syntax layers into domain-specific lyrical representations.

Subsequently, the integration of a Siamese Contrastive Learning framework acted as a crucial geometric stabilizer. By leveraging targeted hard-negative mining based on lexical and embedding similarities, the Siamese layer effectively untangled dense semantic overlaps between proximate affective concepts on the continuous vector space, creating wide discriminative margins between complex classes like Happy and Relaxed before supervised final training.

The contrastively refined hidden outputs were successfully processed via a customized *BertLstmAttnForSequenceClassification* head. Stacking a Bidirectional LSTM network with an Attention Pooling mechanism enabled the pipeline to robustly capture long-range contextual workflows across extensive verses while simultaneously anchoring on emotionally charged focal keywords.

Empirical evaluations structurally validated the superiority of the proposed framework, which achieved a peak overall accuracy of 96.24% and a macro-averaged F1-score of 0.9623, drastically outperforming large multilingual models like mBERT (54.51%) and the bare vanilla monolingual baseline

WangchanBERTa (84.75%). These stable results demonstrate the necessity of spatial metric formatting when handling highly expressive text in low-resource environments.

5.2 Limitations and Future Research Directions

Despite achieving an exceptionally high performance threshold, certain baseline constraints must be highlighted to guide future research in affective computing.

First, the current framework operates purely on a unimodal architecture, extracting signals solely from textual word tokens. Lyrical interpretations can vary across acoustic compositions; therefore, the absence of multimodal features—such as audio signals, vocal pitch tracking, tempo matrices, and melodic arrangements—limits the model's ability to capture the full auditory delivery of the song.

Second, the data labeling logic relies on a discrete four-quadrant partition derived from Russell's Circumplex Model of Affect. While this filters the emotional space cleanly, it does not map subtle sub-emotional blends or transitions within a single musical piece.

To address these limitations, future research pathways will focus on developing Multimodal Fusion Architectures that combine textual token transformers with acoustic audio encoders (e.g., Wav2Vec2 or audio CNNs) via cross-modal attention gating. This fusion will allow the system to resolve textual sarcasm and deep poetic ironies by cross-verifying words against physical musical waves. Furthermore, expanding the dataset to include hierarchical, fine-grained emotional annotations will enable recommendations that better match real-world musical applications.

This study developed a multi-stage deep learning framework designed for emotion classification within Thai song lyrics. By integrating domain-adaptive language modeling with a hybrid neural architecture, the proposed approach effectively addresses inherent challenges in Thai lyrical analysis, such as ambiguous word segmentation, figurative language, and culturally embedded emotional expressions—nuances often missed by general-purpose multilingual models.

Central to this approach was the application of Domain-Adaptive Masked Language Modeling (MLM) to the WangchanBERTa base. Utilizing a large corpus of Thai lyrics enabled the model to acquire domain-specific semantic representations tailored to the linguistic characteristics of lyrical text. Furthermore, incorporating a Bidirectional Long Short-Term Memory (BiLSTM) network and an attention mechanism significantly enhanced the model's capacity to capture long-range contextual dependencies and emphasize emotionally salient terms.

Experimental evaluations demonstrated that the proposed framework substantially outperformed multiple baseline approaches, including mBERT, XLM-RoBERTa, LaBSE, and the standard WangchanBERTa. Achieving an overall accuracy of 96.24% and a macro-averaged F1-score of 0.9623, the model maintained consistently high precision and recall across all emotion categories. These results empirically validate the efficacy of combining domain adaptation with hybrid architectures for emotion classification in culturally nuanced, low-resource languages.

Overall, this work advances Thai natural language processing and affective computing by demonstrating the critical role of domain-specific pretraining and architectural hybridization in improving emotion recognition performance. The proposed framework serves as a robust foundation for future research and practical applications, including emotion-aware music recommendation systems, automated playlist generation, and lyrical content analysis.

6. Acknowledgements

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used: "Conceptualisation, N.M. and P.V.; methodology, N.M. and P.V.; software, N.M.; validation, N.M. and P.V.; formal analysis, P.V.; investigation, N.M. and P.V.; resources, N.M.; data curation, N.M.; writing—original draft preparation, N.M. and P.V.; writing—review and editing, P.V.; visualisation, N.M.; supervision, P.V.; project administration, P.V.; funding acquisition, P.V.; All authors have read and agreed

to the published version of the manuscript." Please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: The author(s) declare that no financial support was received for the research.

Conflicts of Interest: The authors declare no conflict of interest.

References

- [1] Visutsak, P.; Loungna, J.; Sopromrat, S.; Jantip, C.; Soponkittikunchai, P.; Liu, X. Mood-Based Music Discovery: A System for Generating Personalized Thai Music Playlists Using Emotion Analysis. *Appl. Syst. Innov.* **2025**, *8*, 37. <https://doi.org/10.3390/asi8020037>
- [2] Praveena, S. Emotion Classification Using BERT: A Comprehensive Study. *J. Propuls. Technol.* **2024**, *45*(4), 3337–3345.
- [3] Peng, H.; Gu, X.; Li, J.; Wang, Z.; Xu, H. Text-Centric Multimodal Contrastive Learning for Sentiment Analysis. *Electronics* **2024**, *13*, 1149. <https://doi.org/10.3390/electronics13061149>
- [4] Patkar, A.; Chandrashekhar, S.; Kadiyala, R. M. R. RoBERTa-Based Emotion Classification of Essays: Improving Performance on Imbalanced Data. In *Proceedings of the WASSA Shared Task*; 2023.
- [5] Nieminen, M. *The Transformer Model and Its Impact on the Field of Natural Language Processing*; Bachelor's Thesis, Bachelor's Programme in Computer Science, 2023.
- [6] Zhang, D.; Yuan, Z.; Liu, Y.; Liu, H.; Zhuang, F.; Xiong, H.; Chen, H. Domain-Oriented Language Modeling with Adaptive Hybrid Masking and Optimal Transport Alignment. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*; 2021; pp 2145–2153. <https://doi.org/10.1145/3447548.3467215>
- [7] Gao, T.; Yao, X.; Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2021. <https://doi.org/10.18653/v1/2021.emnlp-main.552>.
- [8] Lowphansirikul, L.; Polpanumas, C.; Jantrakulchai, N.; Nutanong, S. WangchanBERTa: Pretraining Transformer-Based Thai Language Models. *PyThaiNLP and VISTEC*, 2021.
- [9] Pasupa, K.; Seneewong Na Ayutthaya, T. Hybrid Deep Learning Models for Thai Sentiment Analysis. *Cogn. Comput.* **2022**, *14*, 167–193. <https://doi.org/10.1007/s12559-020-09770-0>.
- [10] Sánchez-Ruiz, A. A. Analyzing the Similarity Learned by a Siamese Network with Contrastive Loss on the MNIST Dataset. **2020**.
- [11] Koch, G.; Zemel, R.; Salakhutdinov, R. *Siamese Neural Networks for One-Shot Image Recognition*; University of Toronto, 2015.
- [12] Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [13] Russell, J. A. A Circumplex Model of Affect. *J. Pers. Soc. Psychol.* **1980**, *39*(6), 1161–1178. <https://doi.org/10.1037/h0077714>