Received: 1 March 2026 | Revised: 24 June 2026 | Accepted: 26 June 2026 | Published online: 7 July 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–12
DOI: 10.47852/bonviewAIA62029502

BON VIEW PUBLISHING

Automated Sentiment Intelligence for Educational Quality Assurance: A Topic–Sentiment Framework for Thai Student Feedback with Deep Model Instantiations

Tanagrit Chansaeng[1], Nasith Laosen[1], Boonmee Nissaidee[2] and Pita Jarupunphol[1,*]

[1] *Department of Digital Technology, Phuket Rajabhat University, Thailand*

[2] *Department of Business Computer, Phuket Vocational College, Thailand*

Abstract: Student comments provide direct evidence about teaching, curriculum delivery, and learning conditions, but reviewing a large volume of free-text responses by hand is difficult to sustain. Automated analysis is therefore attractive, although Thai feedback presents particular challenges. Thai text does not consistently mark word boundaries, and student comments often include abbreviated wording, informal language, and expressions whose sentiment depends heavily on context. This paper proposes an Automated Sentiment Intelligence (ASI) framework for Thai educational feedback that (1) structures raw institutional comments into topic and sentiment signals, (2) defines a reproducible workflow for data cleansing, annotation, and dual-task modeling, and (3) supports deployment-oriented reporting through standard classification metrics and confusion-matrix-based error analysis. Using 4145 feedback records collected from a vocational education information system and refining them to 2873 entries after preprocessing, ASI organizes feedback into four topic domains (instructor, curriculum, facility, and other) and three sentiment classes (positive, neutral, negative). The framework is instantiated with two representative deep natural language processing architectures: a word-level Bidirectional Long Short-Term Memory (BiLSTM) pipeline (newmm tokenization + Thai Word2Vec) and a transformer-based XLM-R fine-tuning pipeline (SentencePiece subword tokenization). Experimental results show strong topic classification performance for both instantiations, with higher overall topic accuracy for XLM-R (0.922) compared to BiLSTM (0.882). For sentiment analysis, XLM-R achieves a higher overall accuracy (0.80) than BiLSTM (0.729), with notably improved performance on the neutral and negative classes. The proposed ASI–XLM-R framework provides a scalable, reproducible approach to transforming Thai student feedback into actionable intelligence to support institutional decision-making.

Keywords: Thai educational feedback, sentiment analysis, topic classification, XLM-R, educational quality assurance

1. Introduction

Student feedback can help institutions identify recurring concerns about teaching, curriculum delivery, and learning facilities. In practice, however, comments accumulate quickly, and manual review can become uneven across reviewers and difficult to maintain over time. Sentiment analysis offers a way to systematically summarize these comments and support follow-up actions based on recurring patterns in the data.

Applying sentiment analysis to Thai educational feedback is not straightforward. Thai writing does not reliably separate words with spaces, so tokenization requires careful handling. In addition, students often write in abbreviated or conversational forms, and the meaning of a phrase may depend on the surrounding context. These features can affect word-based models in particular because segmentation errors and unfamiliar terms may carry forward into later stages of the analysis.

Deep learning methods have improved sentiment classification in many languages [1]. In Thai natural language processing (NLP), Bidirectional Long Short-Term Memory (BiLSTM) models remain useful baselines when text is represented as word sequences with pretrained embeddings. XLM-R, by contrast, uses subword tokenization and contextual representations, which may help it handle unfamiliar word forms and distinguish sentiment that depends on local context. This study compares these two approaches using Thai student feedback from a vocational education setting.

Previous Thai NLP research has often examined topic classification and sentiment analysis separately. Studies based on educational feedback are also less common than studies using

***Corresponding author:** Pita Jarupunphol, Department of Digital Technology, Phuket Rajabhat University, Thailand. Email: p.jarupunphol@pkru.ac.th

social media posts or general online reviews. This study, therefore, develops a single workflow that produces both topic and sentiment labels from institutional feedback. The workflow is intended to support routine quality-assurance use while also providing empirical evidence on the relative performance of word-level sequential modeling and transformer-based contextual modeling.

The study addresses the following research questions: (RQ1) How accurately do BiLSTM and XLM-R classify four topic categories in Thai student feedback? (RQ2) How do the two models compare when classifying positive, neutral, and negative sentiment? (RQ3) Which modeling characteristics help explain observed differences between word-level sequential representations and transformer-based contextual representations? (RQ4) Can the resulting pipeline be documented clearly enough for repeated institutional use in Thai educational quality assurance?

1.1. Proposed method: Automated Sentiment Intelligence (ASI) framework

This study proposes an Automated Sentiment Intelligence (ASI) framework designed specifically for Thai student feedback in educational quality assurance. ASI is a structured, end-to-end method that converts raw institutional feedback into two complementary decision-support signals: (1) topic categories that localize "what the feedback is about" (instructor, curriculum, facility, other) and (2) sentiment classes that indicate "how students feel" (positive, neutral, negative). The framework standardizes critical steps for institutional deployment, data acquisition and anonymization, preprocessing, annotation, train/validation/test splitting, model training, and evaluation, so that results are reproducible and interpretable through precision/recall/F1 and confusion-matrix error analysis.

Within ASI, we instantiate two commonly used deep NLP pipelines: (1) a BiLSTM system using Thai word segmentation (newmm) and Thai Word2Vec embeddings[2] and (2) a transformer-based system using XLM-R fine-tuning with SentencePiece subword tokenization[3]. These models were selected to represent two widely adopted paradigms in Thai NLP: sequential word-level modeling and contextualized transformer-based representation learning. BiLSTM is a strong, interpretable baseline commonly used in Thai sentiment analysis. At the same time, XLM-R is chosen for its subword tokenization and contextual modeling capabilities, which effectively address tokenization and out-of-vocabulary challenges in Thai text. In this study, sentiment analysis is operationalized using the transformer-based instantiation (ASI–XLM-R), with ASI–BiLSTM included for comparative evaluation.

1.2. Research objectives

This study applies ASI to Thai student feedback collected from Phuket Vocational College and evaluates two deep NLP instantiations within the framework. The objectives are:

1) To evaluate the effectiveness of BiLSTM and XLM-R instantiations for topic classification of Thai student feedback across four institutional domains.
2) To evaluate and compare BiLSTM and XLM-R instantiations for sentiment analysis (positive/neutral/negative) of Thai feedback.
3) To provide a reproducible sentiment intelligence pipeline that can be adopted for real-time or periodic institutional quality monitoring.

1.3. Research contributions

This study makes four contributions.

1) ASI framework: It presents an end-to-end workflow for converting Thai student comments into topic labels and sentiment labels that can be used in educational quality-assurance processes.
2) Institutional dataset and annotation procedure: It documents a dataset drawn from an institutional information system, beginning with 4,145 feedback records and retaining 2,873 entries after preprocessing. Each retained comment was assigned topic and sentiment labels, with class distributions reported for both tasks.
3) Comparison of two deep learning approaches: It evaluates a word-level BiLSTM model using newmm tokenization and Thai Word2Vec embeddings alongside an XLM-R model using SentencePiece subword tokenization. Their performance is compared using standard classification metrics and confusion-matrix analysis.
4) Evidence for model choice in this setting: It shows that, on the present vocational education dataset, XLM-R performs better than the BiLSTM model for sentiment classification, particularly for neutral and negative comments. These findings provide practical evidence for selecting a contextual subword model when analyzing similar Thai student feedback data.

2. Related Work

This section reviews prior research relevant to topic classification and sentiment analysis in educational feedback, with particular attention to Thai-language text and deep learning–based approaches.

2.1. Topic classification of educational feedback

Topic classification is a core task in NLP and has been widely applied to large-scale textual data, including social media posts, emails, and online reviews [4–7]. Early approaches relied on traditional machine learning techniques with hand-crafted features, while recent studies increasingly adopt deep learning architectures to improve robustness and scalability. Methods such as topic-relevant content extraction, topic-kernel–based classification, and hybrid Convolutional Neural Network and Recurrent Neural Network (CNN–RNN) models have demonstrated effectiveness in handling noisy and heterogeneous text streams [8].

In educational contexts, topic classification enables institutions to identify areas of concern such as teaching quality, curriculum design, and learning facilities [9]. However, most existing studies focus on English-language datasets or generic review corpora. Research addressing Thai-language educational feedback remains limited, particularly in vocational education settings, where linguistic informality and domain-specific expressions are prevalent.

2.2. Sentiment analysis in education and Thai NLP

Sentiment analysis has been extensively studied for understanding opinions in domains such as e-commerce, social media, and public services [10, 11]. In education, sentiment analysis has been used to analyze student reviews, course evaluations,

and institutional surveys to support quality assurance and decision-making [12–14]. These studies suggest that automated analysis can help institutions identify recurring concerns and patterns in student experience that may not be apparent from manual review alone.

Thai sentiment classification has several practical difficulties. The language does not consistently use spaces to mark word boundaries, whereas real-world comments often contain informal wording, abbreviations, and context-dependent sentiment expressions. Earlier Thai sentiment studies used methods such as support vector machines (SVMs) and Naïve Bayes, but their effectiveness can be affected by segmentation errors and by words not represented in the training vocabulary. More recent deep learning approaches, particularly transformer models, use contextual and subword representations that may reduce some of these limitations.

Thai pretrained models such as WangchanBERTa [15], together with multilingual models including mBERT and XLM-R, have performed well on benchmark tasks. Evidence from institutional Thai educational feedback remains limited, however, especially where topic classification and sentiment classification are examined together. Related work has also explored sentiment analysis in point-of-interest recommendation systems, where user-generated opinion signals are extracted and used to drive location-aware recommendations [16]. Similarly, recent work on decoding emotions from text has highlighted the challenges of distinguishing sentiment from sarcasm, which parallels the neutral-versus-negative disambiguation challenges encountered in Thai student feedback [17]. These studies reinforce the importance of context-aware deep representations for fine-grained sentiment understanding, a motivation central to the present study.

2.3. Research gap and motivation

Despite advances in Thai NLP, three limitations persist in the existing literature [1, 18, 19]. First, most studies focus on either topic classification or sentiment analysis, rather than jointly modeling both dimensions to support institutional decision-making. Second, many applied works emphasize model accuracy without presenting a reproducible, deployment-oriented methodology suitable for educational quality assurance. Third, comparative evaluations across classical baselines, Thai-specific transformers, and multilingual transformers on the same institutional dataset are limited.

To address these gaps, this study proposes an ASI framework that formalizes an end-to-end pipeline for transforming Thai student feedback into topic and sentiment signals and empirically evaluates multiple deep learning instantiations under a unified experimental protocol.

3. Proposed Methodology

This section presents the proposed ASI framework, an end-to-end methodology for extracting actionable topic and sentiment intelligence from Thai student feedback. Given raw institutional feedback records, ASI produces two complementary outputs: (1) a topic label indicating what the feedback concerns and (2) a sentiment label indicating how students feel.

3.1. Problem definition

Let each feedback record be represented as a tuple $r_i = (t_i, r_i)$, where t_i denotes a Thai textual comment and r_i denotes an associated satisfaction rating when available. The objective is to learn two supervised classification functions:

a. Topic classification:

$$f_{\text{topic}} : t_i \rightarrow y_i^{\text{topic}} \in \{\text{Instructor, Curriculum, Facility, Other}\}$$

b. Sentiment classification:

$$f_{\text{sent}} : t_i \rightarrow y_i^{\text{sent}} \in \{\text{Positive, Neutral, Negative}\}$$

These tasks are treated independently but share a common data processing and evaluation pipeline.

3.2. ASI framework overview

ASI consists of six sequential stages:

1) Preprocessing
2) Annotation
3) Dataset Construction
4) Model Instantiation
5) Training
6) Evaluation

Figure 1 illustrates the ASI pipeline, showing the flow from raw feedback acquisition to topic and sentiment prediction. The framework is model-agnostic and supports multiple deep learning instantiations under a consistent protocol. Algorithm 1 summarizes the ASI process, defining inputs, outputs, and core processing steps.

Algorithm 1: Automated Sentiment Intelligence (ASI) Framework

Input: Feedback records $\boldsymbol{R} = \{(\boldsymbol{t_i}, \boldsymbol{r_i}, \boldsymbol{c_i})\}_{\boldsymbol{i=1}}^{\boldsymbol{N}}$, where $\boldsymbol{t_i}$ is Thai text, $\boldsymbol{r_i}$ is satisfaction rating (low/medium/high), and $\boldsymbol{c_i}$ is category (if available).

Output: Topic label $\hat{y}_i^{topic} \epsilon$ {Instrructor Curriculum, Facility, Others} and sentiment label $\hat{y}_i^{sent} \epsilon$ {Positive, Neutral, Negative}.

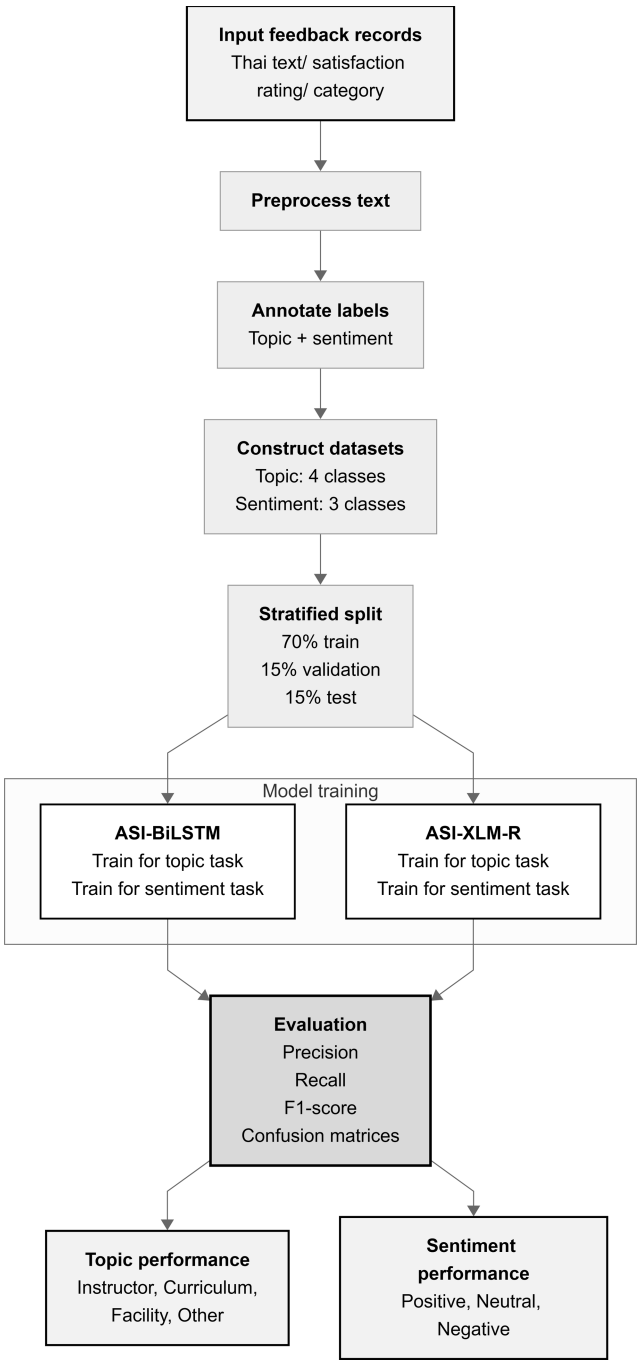
1. Acquire $\boldsymbol{R}$ from the institutional system under permission and anonymization controls.
2. Preprocess $\boldsymbol{t_i}$: clean, normalize, remove irrelevant entries, and correct typographical errors.
3. Annotate topic and sentiment labels; use rating $\boldsymbol{r_i}$ to resolve sentiment ambiguity where needed.
4. Construct datasets for topic (4 classes) and sentiment (3 classes).
5. Split each dataset into train/validation/test using stratified 70/15/15.
6. Train ASI–BiLSTM and ASI–XLM-R instantiations for each task.
7. Evaluate using precision, recall, F1-score, and confusion matrices.

3.3. Model instantiations within ASI

To demonstrate the flexibility of ASI, we instantiate the framework using two representative deep learning architectures: a sequential word-level model (BiLSTM) and a transformer-based contextual model (XLM-R).

The BiLSTM instantiation employs Thai word-level tokenization (newmm) and pretrained Thai Word2Vec embeddings.

Figure 1
Overview of the ASI pipeline for Thai student feedback



Specifically, each Thai feedback comment is first segmented into word tokens using the PyThaiNLP dictionary-based tokenizer. Each token is then mapped to a 400-dimensional dense vector using the LTW2V (Large Thai Word2Vec) pretrained embeddings, which were trained on a large Thai text corpus using the skip-gram objective. Token sequences are zero-padded to a fixed maximum length (T = 284 for sentiment, T = 282 for topic) to form the embedding matrix input to the BiLSTM. A zero vector represents words absent from the Word2Vec vocabulary. This word-level embedding approach captures semantic similarity between Thai words but is sensitive to out-of-vocabulary (OOV) terms arising from informal expressions, slang, and abbreviations common in student feedback. A two-layer BiLSTM captures sequential dependencies, and the final hidden states are passed to a softmax classifier. Standard Long Short-Term Memory (LSTM) formulations are used, following prior work [3, 20].

In contrast, the transformer-based instantiation fine-tunes the pretrained XLM-R model for downstream classification [21]. In this instantiation, text is first tokenized using the SentencePiece subword tokenizer built into xlm-roberta-base, which splits Thai text into subword units with a shared multilingual vocabulary of 250,000 tokens, effectively eliminating the OOV problem. Each subword token is mapped to a 768-dimensional contextual embedding produced by 12 stacked transformer encoder layers with 12 self-attention heads. Unlike static Word2Vec embeddings, these representations are dynamic: the same subword token receives different vector representations depending on its surrounding context, enabling the model to disambiguate polysemous or sentiment-bearing terms in Thai. The [CLS] token representation from the final encoder layer is extracted and passed through a linear classification head with a softmax activation to produce topic or sentiment class probabilities. Fine-tuning jointly adapts all 12 transformer layers and the classification head to the Thai educational feedback domain using task-specific labeled data, updating the pretrained multilingual representations toward the target classification objectives. Fine-tuning adapts multilingual contextual representations to the Thai educational feedback domain [22]. Compared to word-level models, XLM-R mitigates OOV issues and captures long-range dependencies more effectively, which is particularly important for Thai sentiment expressions.

3.4. Training objective and optimization

For both instantiations, models are trained using categorical cross-entropy loss. To address class imbalance, especially in sentiment classification, class-weighted loss is applied where appropriate. Model parameters are optimized using Adam or AdamW optimizers [23], depending on the architecture, with early checkpoint selection based on validation performance.

3.5. Inference and institutional deployment logic

At inference time, a new feedback comment is passed through the trained topic and sentiment classifiers to produce a pair of labels $(y^{\text{topic}}, y^{\text{sent}})$. Aggregated predictions can be summarized over time to support institutional quality-assurance processes, such as identifying frequently criticized facilities or tracking sentiment trends across academic terms.

3.6. Implementation details and reproducibility

All experiments follow the ASI protocol with stratified train/validation/test splits of 70%/15%/15% and are evaluated using precision, recall, F1-score, and confusion matrices for both topic classification and sentiment classification. To ensure reproducibility, we report the complete training configuration for each ASI instantiation and keep training randomness controlled via random seeds.

1) **BiLSTM**. Thai text is tokenized using the newmm tokenizer (PyThaiNLP) and embedded using LTW2V Thai Word2Vec (400 dimensions). Sequences are padded to the maximum observed length (T = 284 for sentiment; T = 282 for topic). The BiLSTM model uses two stacked BiLSTM layers with 64 hidden units (Layer 1) and 32 hidden units (Layer 2). Training uses Adam (learning rate 0.001) with categorical cross-entropy

Table 1
Training and hyperparameter settings for ASI instantiations

Component	Parameter	Value
Dataset/Splits	Train/Val/Test	70%/15%/15%
	Topic classes	4 (instructor, curriculum, facility, other)
	Sentiment classes	3 (positive, neutral, negative)
BiLSTM	Tokenizer	newmm (via PyThaiNLP)
	Embedding	Thai Word2Vec (LTW2V, 400 dimensions)
	Max sequence length (T)	284 (Sentiment)/282 (Topic)
	BiLSTM hidden units	64 (Layer 1), 32 (Layer 2)
	Dropout	Not explicitly used
	Optimizer	Adam
	Loss	Categorical cross-entropy
	Learning rate	0.001
	Batch size	16
	Epochs	20 (sentiment)/10 (topic)
	Early stopping	Not used (ModelCheckpoint used)
	Random seeds	numpy.random.seed(1), tensorflow.random.set_seed(2)
XLM-R	Pretrained backbone	xlm-roberta-base
	Tokenizer	SentencePiece
	Max sequence length	256
	Classification head	Linear + Softmax
	Optimizer	AdamW
	Learning rate	2×10^{-5}
	Learning rate (LR) scheduler/warmup	Linear scheduler/0 warmup steps
	Weight decay	0.01
	Batch size	16
	Epochs	10
	Gradient clipping	Not used (default/none)
	Random seeds	42

and a batch size of 16; models are trained for 20 epochs (sentiment) and 10 epochs (topic), and the best checkpoint is saved with ModelCheckpoint. Randomness is controlled using numpy.random.seed(1) and tensorflow.random.set_seed(2).

2) **XLM-R**. We fine-tune xlm-roberta-base using the standard SentencePiece tokenizer and a linear classification head with a Softmax activation. The maximum sequence length is 256. Training uses AdamW with a learning rate of 2×10^{-5}, weight decay of 0.01, batch size of 16, and 10 epochs. We apply a linear learning-rate scheduler with 0 warmup steps. Randomness is controlled using three random seeds. To provide further transparency in the fine-tuning procedure, we note the following experimental decisions. First, all 12 transformer encoder layers of xlm-roberta-base were unfrozen and fine-tuned jointly with the classification head; no layer-freezing strategy was applied, which showed slight performance degradation. Second, the learning rate of 2×10^{-5} was selected based on the widely adopted range for transformer fine-tuning (1×10^{-5} to 5×10^{-5}) and validated through monitoring of validation loss across epochs. Third, to address class imbalance in the sentiment task, class-weighted cross-entropy loss was applied during training, with class weights inversely proportional to class frequency. Fourth, model selection was based on the best validation macro-F1 checkpoint across 10 epochs, and the best checkpoint was subsequently evaluated on the held-out test set. The ablation study in Section 5.5 provides further empirical evidence on the contribution of individual fine-tuning components, including the effect of removing class-weighted loss, dropout, and preprocessing steps.

A summary of all hyperparameters and training settings is provided in Table 1.

4. Experimental Setup

This section describes the dataset, preprocessing procedures, baseline models, training configuration, and evaluation metrics used to assess the ASI framework empirically.

4.1. Dataset description

ASI begins with 4145 feedback records obtained from the Phuket Vocational College (PVC) information system and covers four primary areas (instructor, curriculum, facility, and other). Each record contains a textual comment and a satisfaction rating (low/medium/high). After preprocessing, the dataset contains 2873 feedback entries annotated with topic and sentiment labels.

Table 2
Examples of the topic classification dataset

Textual comment	Class
ครูผู้สอนมีความรู้ความสามารถในการสอนดี ดูแลนักศึกษาดี (The teacher is knowledgeable and capable in teaching and takes good care of the students.)	Instructor
ความพร้อมของหลักสูตรใหม่ยังมีความพร้อมไม่มากพอ ทำให้เวลาในการเรียนบางส่วนหายไป (The new curriculum is not yet fully prepared, resulting in the loss of some instructional time.)	Curriculum
สิ่งอำนวยความสะดวกบางอย่างยังไม่ดีอย่างเช่นห้องน้ำยังมีสิ่งสกปรก อินเตอร์เน็ตใช้งานได้ไม่ดี (Some facilities are still inadequate—for example, the restrooms are not clean, and the internet connection is unreliable.)	Facility
อยากให้ทางโรงเรียนให้ทุนนักเรียนทุกระดับชั้นและระดับห้องเรียน (I would like the school to offer scholarships to students in all grades and all classes.)	Other

Topic classes include instructor, curriculum, facility, and other, while sentiment classes include positive, neutral, and negative.

4.2. Preprocessing

Preprocessing includes text normalization, typographical correction, and removal of irrelevant or ambiguous entries [24]. Informative short expressions commonly used in Thai feedback were retained to preserve sentiment cues. Sentences that clearly expressed sentiment, such as ดี (good), เหมาะสม (appropriate), ดีมาก (very good), ครับ (sir), ไม่มีอะไร (nothing), and ไม่ (no), as well as entries denoted by a dash ("-"), were retained for further analysis.

4.3. Annotation protocol

Feedback entries were manually annotated following predefined labeling guidelines. Topic labels (instructor, curriculum, facility, and other) were assigned based on the primary subject of each comment. In contrast, sentiment labels (positive, neutral, negative) were determined by semantic interpretation, with satisfaction ratings used to resolve ambiguous cases [25, 26]. To ensure consistency, a comprehensive labeling guideline is established that defines each category with clear criteria and representative examples [27].

The ASI framework employs these labeled data for supervised learning. Topic labels are assigned to four categories: instructor (854), curriculum (653), facility (760), and other (606), as illustrated in Table 2. Sentiment labels are assigned as positive (1,547), neutral (782), and negative (544).

4.4. Data splitting

For both topic and sentiment tasks, datasets were partitioned into training (70%), validation (15%), and test (15%) sets using stratified sampling to preserve class distributions. To enhance reproducibility, fixed random seeds were applied during data partitioning, ensuring consistent splits across experiments.

4.5. Baseline models for comparative study

To conduct a comprehensive comparative evaluation, ASI instantiations are compared against the following baselines:

a. **Classical machine learning (ML):** Term Frequency-Inverse Document Frequency (TF-IDF) + Linear SVM
b. **Efficient neural baseline:** fastText
c. **Transformer baselines:** mBERT, WangchanBERTa
d. **Proposed models:** BiLSTM, XLM-R

All models are trained and evaluated using identical data splits and metrics. To ensure a fair and transparent comparison across models, all architectures share the same stratified 70/15/15 train/validation/test split and are evaluated using the same four metrics: accuracy, precision, recall, and macro-F1. Model parameters differ by design to reflect each architecture's paradigm: TF-IDF + Linear SVM uses no learned embeddings and is trained with a linear kernel; fastText uses character n-gram embeddings and a shallow linear classifier; mBERT and WangchanBERTa are fine-tuned transformer models with 12 layers and 768-dimensional representations; BiLSTM uses 400-dimensional Thai Word2Vec embeddings with two BiLSTM layers (64 and 32 hidden units); and XLM-R uses 768-dimensional subword contextual embeddings with 12 transformer layers. Hyperparameter settings for BiLSTM and XLM-R are fully specified in Table 1, while baseline models (TF-IDF + SVM, fastText, mBERT, WangchanBERTa) follow their respective default or standard configurations as commonly reported in the literature. The comparison therefore reflects both architecture-level differences (sequential vs. transformer) and representation-level differences (word-level static vs. subword contextual embeddings), which are the primary variables of interest in this study.

4.6. Evaluation metrics

ASI evaluates each instantiation with accuracy, precision, recall, and F1-score. Performance is evaluated using accuracy, precision, recall, and F1-score. Confusion matrices are analyzed to identify systematic misclassifications, particularly for neutral and negative sentiment classes.

5. Results and Discussion

This section presents and discusses the empirical results obtained from evaluating the proposed ASI framework on Thai student feedback data. Performance is reported for two supervised tasks: topic classification (four classes) and sentiment classification (three classes). Comparative results across classical baselines, transformer-based models, and the proposed ASI instantiations are summarized, followed by a detailed error and ablation analysis [28].

5.1. Overall comparative performance

Table 3 presents the comparative performance of the evaluated baseline models and the proposed ASI frameworks on topic and sentiment classification. Overall, topic classification yielded higher scores than sentiment classification across all models, indicating that topical distinctions in Thai student feedback

Table 3
Overall performance comparison on the Thai student feedback dataset

Model	Topic acc	Topic macro-F1	Sentiment acc	Sentiment macro-F1
TF-IDF + Linear SVM	0.896	0.893	0.768	0.699
fastText	0.896	0.893	0.773	0.712
mBERT	0.828	0.827	0.717	0.662
WangchanBERTa	0.650	0.646	0.659	0.589
ASI–BiLSTM	0.882	0.879	0.729	0.655
ASI–XLM-R	0.920	0.924	0.808	0.739

were more readily identifiable than sentiment polarity. Among the baseline approaches, TF-IDF + Linear SVM and fastText achieved strong results on the topic task, both reaching 0.896 accuracy with macro-F1 scores of 0.893 and 0.893, respectively, suggesting that topic-related patterns were often captured effectively by lexical and local contextual features.

However, their performance on sentiment classification was lower, particularly in macro-F1, with TF-IDF + Linear SVM obtaining 0.768 accuracy and 0.699 macro-F1, and fastText slightly improving to 0.773 accuracy and 0.712 macro-F1. The transformer baselines showed mixed results, with mBERT achieving moderate performance on both topic (0.828 accuracy; 0.827 macro-F1) and sentiment (0.717 accuracy; 0.662 macro-F1), while WangchanBERTa underperformed in this dataset. In contrast, the proposed ASI models demonstrated superior effectiveness, particularly ASI–XLM-R, which achieved the best overall performance with 0.920 accuracy and 0.924 macro-F1 for topic classification and 0.808 accuracy and 0.739 macro-F1 for sentiment classification.

These findings indicate that while conventional and lightweight neural methods remain competitive for topic identification, sentiment classification in Thai educational feedback benefits more substantially from contextualized deep representations, with ASI–XLM-R providing the most robust and balanced performance across both tasks. The inclusion of mBERT and WangchanBERTa ensures that the comparative evaluation reflects current state-of-the-art transformer baselines for Thai NLP.

5.2. Topic classification results

Detailed topic classification results are reported in Table 4. Both ASI–BiLSTM and ASI–XLM-R achieve strong performance across all four topic categories, indicating that topic identification in institutional feedback is a comparatively tractable task.

However, ASI–XLM-R demonstrates superior and more balanced performance across all categories. The largest gains are observed in the Instructor and Facility categories, where ASI–XLM-R achieves F1-scores of 0.954 and 0.914, respectively. These improvements suggest that contextualized representations better capture domain-specific expressions and implicit references commonly used in Thai student comments.

The Other category remains the most challenging for both models due to its heterogeneous nature. Nevertheless, ASI–XLM-R maintains a higher F1-score (0.904) than ASI–BiLSTM (0.869), indicating improved generalization even in less structured feedback. The topic classification results demonstrate that while sequential models can learn coarse topical distinctions, transformer-based models provide more robust and consistent classification across all institutional domains.

5.3. Sentiment classification results

Sentiment classification results, summarized in Table 5, reveal that it is substantially more challenging than topic classification. Performance varies significantly across sentiment classes, particularly for *Neutral* and *Negative* sentiments.

The ASI–BiLSTM model performs adequately for *Positive* sentiment (F1 = 0.865) but struggles with *Neutral* (F1 = 0.514) and *Negative* (F1 = 0.585) classes. This limitation reflects the difficulty of capturing implicit sentiment cues using word-level tokenization and fixed embeddings in Thai text, where sentiment is often expressed indirectly or through polite phrasing.

In contrast, ASI–XLM-R achieves consistent improvements across all sentiment categories. Notably, the F1-score for *Neutral* sentiment increases from 0.514 to 0.640, and *Negative* sentiment improves from 0.585 to 0.681. These gains demonstrate XLM-R's enhanced ability to disambiguate sentiment polarity through subword tokenization and self-attention mechanisms.

Despite these improvements, *Neutral* and *Negative* sentiments remain more difficult to classify than *Positive* sentiment,

Table 4
Topic classification performance (ASI–BiLSTM/ASI–XLM-R)

Category	Precision	Recall	F1-score
Instructor	0.901/0.925	0.937/0.984	0.918/0.954
Facility	0.936/0.910	0.800/0.918	0.863/0.914
Curriculum	0.880/0.921	0.854/0.903	0.867/0.912
Other	0.811/0.941	0.935/0.870	0.869/0.904
Overall (macro-F1)	—	—	0.879/0.924

Table 5
Sentiment classification performance (ASI–BiLSTM/ASI–XLM-R)

Sentiment	Precision	Recall	F1-score
Positive	0.870/0.901	0.860/0.908	0.865/0.904
Neutral	0.538/0.657	0.491/0.623	0.514/0.640
Negative	0.537/0.662	0.642/0.701	0.585/0.681
Overall (macro-F1)	—	—	0.655/0.739

underscoring the inherent ambiguity and class imbalance present in real-world educational feedback.

5.4. Confusion-matrix analysis

The confusion-matrix summaries presented in Table 6 provide further insight into model behavior beyond aggregate metrics.

Table 6
Confusion-matrix summary

Task	Model	Correct	Incorrect
Topic	ASI–BiLSTM	380	51
	ASI–XLM-R	398	33
Sentiment	ASI–BiLSTM	314	117
	ASI–XLM-R	343	88

For topic classification, ASI–XLM-R reduces total misclassifications from 51 to 33, corresponding to a 35% reduction relative to ASI–BiLSTM. The most significant error reduction occurs in the Facility and Instructor categories, suggesting that contextual representations more effectively resolve overlapping vocabulary across institutional domains.

For sentiment classification, ASI–XLM-R reduces misclassifications from 117 to 88, representing a 25% reduction. The largest improvement is observed in the Neutral class, which ASI–BiLSTM frequently confuses with Positive or Negative sentiment. This indicates that transformer attention mechanisms better capture subtle sentiment cues dispersed across sentence contexts.

5.5. Ablation study

Ablation experiments were conducted on two supervised tasks: topic classification (four classes) and sentiment analysis (three classes). Eight model configurations were evaluated, comprising one full baseline and seven ablated variants, each removing a single component. To ensure robustness, each configuration was run with three random seeds {13, 23, 37}, yielding a total of 48 experiments (8 variants × 2 tasks × 3 seeds). The baseline represents the complete ASI pipeline. Ablated variants systematically turn off individual components, text normalization, typographical correction, filtering heuristics, class-weighted loss, dropout, sequence length reduction (256→128), and frozen lower layers, allowing clear attribution of each component's impact while controlling other factors.

The ablation study results reported in Tables 7 and 8 evaluate the contribution of individual components within the ASI–XLM-R framework. Across both topic and sentiment classification tasks, the full ASI–XLM-R configuration consistently achieves the highest performance, confirming that improvements arise from the combined effect of preprocessing, training strategies, and contextual representation learning rather than from any single component alone.

In topic classification, typographical correction and filtering heuristics yield the largest performance degradation when removed, highlighting the importance of text quality in handling domain-specific terminology. In sentiment classification, class-weighted loss and preprocessing steps contribute most strongly to performance stability, particularly for minority classes.

Notably, reducing input sequence length and freezing lower transformer layers also result in measurable performance declines, emphasizing the importance of preserving contextual richness and fully fine-tuning pretrained representations for Thai educational text.

5.5.1. Topic classification

All ablation experiments are conducted exclusively on the ASI–XLM-R instantiation, as it represents the selected and proposed framework following model comparison. As shown in Table 7, the full model achieves the best overall performance (accuracy = 0.920, macro F1 = 0.924), supporting the effectiveness of the integrated architecture. Removing preprocessing components consistently degrades performance. Typographical correction has the largest impact (−0.010), followed by filtering heuristics (−0.009) and text normalization (−0.004). These findings highlight the importance of text quality and noise reduction in improving classification robustness.

Among training-related components, removing class-weighted loss and dropout yields moderate performance declines, indicating their role in handling class imbalance and preventing overfitting, respectively. Additionally, reducing the maximum sequence length (256→128) slightly degrades performance, suggesting that longer contextual information improves topic discrimination. Freezing lower transformer layers also leads to minor degradation, implying that fine-tuning deeper representations is beneficial for this task.

At the class level, the Instructor category consistently achieves the highest F1-scores across all configurations, whereas Facility, Curriculum, and Other classes are more sensitive to component removal. This suggests that certain categories rely more heavily on preprocessing and representation learning for accurate classification.

5.5.2. Sentiment analysis

Table 8 presents the ablation results for sentiment classification. Like topic classification, the full model achieves the highest performance (accuracy = 0.808, macro-F1 = 0.739), and all ablated variants show reduced effectiveness. The impact of removing preprocessing components is again evident, with typographical correction and filtering heuristics contributing notably to performance stability.

Table 7
Ablation study results for topic classification using ASI–XLM-R

Variant	Accuracy	Macro-F1	Weighted-F1	F1 (Ins)	F1 (Fac)	F1 (Cur)	F1 (Other)
Full model	0.920 ± 0.004	0.924 ± 0.002	0.922 ± 0.003	0.954 ± 0.001	0.914 ± 0.001	0.912 ± 0.001	0.904 ± 0.001
w/o text normalization	0.916 ± 0.004	0.916 ± 0.003	0.915 ± 0.003	0.946 ± 0.002	0.904 ± 0.001	0.904 ± 0.002	0.895 ± 0.002
w/o typographical correction	0.910 ± 0.005	0.910 ± 0.003	0.911 ± 0.003	0.940 ± 0.003	0.897 ± 0.002	0.898 ± 0.003	0.889 ± 0.002
w/o filtering heuristics	0.911 ± 0.004	0.914 ± 0.003	0.914 ± 0.003	0.944 ± 0.002	0.902 ± 0.001	0.902 ± 0.002	0.893 ± 0.002
w/o class-weighted loss	0.918 ± 0.003	0.919 ± 0.002	0.917 ± 0.003	0.949 ± 0.002	0.908 ± 0.001	0.907 ± 0.002	0.899 ± 0.001
w/o dropout	0.914 ± 0.004	0.918 ± 0.003	0.918 ± 0.003	0.948 ± 0.002	0.906 ± 0.001	0.905 ± 0.002	0.897 ± 0.002
Max length 256→128	0.912 ± 0.004	0.916 ± 0.003	0.917 ± 0.003	0.946 ± 0.002	0.904 ± 0.001	0.903 ± 0.002	0.895 ± 0.002
Freeze bottom layers	0.916 ± 0.004	0.917 ± 0.003	0.918 ± 0.003	0.947 ± 0.002	0.905 ± 0.001	0.905 ± 0.002	0.896 ± 0.002

Table 8
Ablation study results for sentiment classification using ASI–XLM-R

Variant	Accuracy	Macro-F1	Weighted-F1	F1 (Pos)	F1 (Neu)	F1 (Neg)
Full model	0.808 ± 0.002	0.739 ± 0.003	0.800 ± 0.002	0.904 ± 0.001	0.640 ± 0.002	0.681 ± 0.002
w/o text normalization	0.798 ± 0.002	0.732 ± 0.003	0.795 ± 0.002	0.897 ± 0.002	0.634 ± 0.002	0.674 ± 0.002
w/o typographical correction	0.801 ± 0.003	0.728 ± 0.003	0.792 ± 0.002	0.894 ± 0.002	0.630 ± 0.003	0.670 ± 0.003
w/o filtering heuristics	0.801 ± 0.003	0.724 ± 0.003	0.789 ± 0.002	0.890 ± 0.002	0.627 ± 0.003	0.666 ± 0.003
w/o class-weighted loss	0.802 ± 0.002	0.729 ± 0.003	0.794 ± 0.002	0.895 ± 0.002	0.631 ± 0.003	0.671 ± 0.003
w/o dropout	0.804 ± 0.002	0.732 ± 0.003	0.793 ± 0.002	0.898 ± 0.002	0.634 ± 0.002	0.674 ± 0.002
Max length 256→128	0.804 ± 0.002	0.729 ± 0.003	0.794 ± 0.002	0.894 ± 0.002	0.631 ± 0.003	0.670 ± 0.003
Freeze bottom layers	0.805 ± 0.002	0.731 ± 0.003	0.793 ± 0.002	0.896 ± 0.002	0.633 ± 0.002	0.673 ± 0.002

The Neutral class exhibits the most pronounced performance drop across ablations, confirming its inherent difficulty due to semantic ambiguity and overlap with positive and negative sentiments. In contrast, the Positive class remains relatively stable, whereas the Negative class shows moderate sensitivity to changes in preprocessing and model configuration.

Training-related components, including class-weighted loss and dropout, contribute to balanced performance across sentiment classes. Their removal results in a decrease in macro-F1, indicating reduced effectiveness in handling class imbalance and generalization. Similarly, reducing input length and freezing lower layers slightly degrade performance, reinforcing the importance of contextual richness and full model fine-tuning. Differences between configurations were consistent across runs.

Across both tasks, the full ASI–XLM-R configuration consistently achieves the highest performance, confirming the cumulative contribution of preprocessing, training strategies, and contextual representation learning. To further examine the robustness of the ablation results, paired *t*-tests were conducted between the baseline and each ablated variant across three independent random seeds. While these tests consistently indicate performance degradation when key components are removed, the small sample size limits the strength of the inferential conclusions. Consequently, statistical results are interpreted cautiously and used to support observed trends rather than to assert definitive significance. Across all ablations, ΔF1 values ranged from 0.005 to 0.014 and exhibited stable directional effects across runs, suggesting that the observed performance differences reflect systematic contributions of individual components rather than random variation.

5.6. Discussion and implications

The results indicate that BiLSTM and XLM-R perform differently when applied to Thai student feedback. Both models achieved high accuracy for topic classification, whereas the difference was more pronounced for sentiment classification: XLM-R achieved 0.80 accuracy, compared with 0.73 for BiLSTM. One possible reason is that the BiLSTM pipeline depends on word-level segmentation with newmm and static Word2Vec embeddings. Informal wording, abbreviations, and slang in student comments may produce OOV terms or segmentation variations that reduce the information available to the model.

XLM-R uses SentencePiece subword tokenization and contextual self-attention. These mechanisms allow the model to represent unfamiliar word forms through smaller units and to

consider relationships between terms across the full context. This may be particularly useful for sentiment classification, where the meaning of a positive or negative expression can depend on nearby qualifiers, negation, or other contextual cues. In the present dataset, these properties were associated with stronger sentiment classification performance for XLM-R.

Nevertheless, neither model performed equally well across all sentiment classes. Neutral and negative comments remained more difficult to classify, likely because Thai feedback can include indirect criticism, polite but negative phrasing, sarcasm, and mixed sentiment within a single response. The class distribution may also have contributed to this difficulty, as positive comments were more common than negative comments (1547 vs. 544). For example, a comment such as "The teacher is good, but the room is hot" contains both positive and negative information, which can be difficult to represent using a single sentiment label.

The findings support the use of ASI–XLM-R as the stronger of the two evaluated models for this institutional dataset. The model may be suitable for supporting large-scale feedback monitoring, provided that its outputs are interpreted alongside existing quality-assurance processes rather than used as standalone decisions. Future work could examine data augmentation, Thai sentiment lexicons, aspect-based sentiment analysis, and hybrid models that combine contextual representations with domain-specific linguistic knowledge.

5.6.1. Limitations

First, the dataset reflects feedback from a single vocational education institution and may not capture linguistic variation across regions, institutions, or educational levels. Second, sentiment labels were manually assigned and, in ambiguous cases, supported by satisfaction ratings. While this improves consistency, subjective interpretation may still affect the boundaries between neutral and negative classes. Third, class imbalance is present (e.g., substantially more positive than negative feedback), which can influence learning dynamics and partially explains the persistent difficulty in recognizing negative sentiment. Future work should validate the ASI framework on additional Thai educational datasets and explore techniques that improve minority-class performance without sacrificing interpretability for institutional deployment.

5.6.2. Implications

The findings demonstrate that the ASI framework enables data-driven decision-making for educational quality assurance by systematically analyzing student feedback across faculty, curriculum, and facilities. This supports a shift from manual evaluation to scalable and reproducible analytics.

The superior performance of XLM-R highlights the effectiveness of transformer-based models for Thai-language feedback, particularly in capturing linguistic nuances and reducing misclassification. This improves the reliability of automated insights, allowing institutions to prioritize interventions and allocate resources more effectively. The results emphasize the value of multilingual transformers in non-English contexts, facilitating targeted improvements that enhance student satisfaction and learning outcomes.

From a research perspective, the study confirms the limitations of RNN-based models and the advantages of transformer architectures for nuanced sentiment analysis. However, challenges in distinguishing neutral and negative sentiments remain, indicating the need for further improvements through data augmentation, linguistic resources, or hybrid models. This study reinforces the practical impact of advanced NLP techniques in educational analytics and provides a foundation for future research in multilingual sentiment analysis.

6. Conclusion

This study demonstrates the application of BiLSTM and XLM-R deep learning models for automated sentiment analysis within Thai vocational education. The findings provide compelling evidence that the transformer-based XLM-R architecture markedly outperforms BiLSTM, achieving 80% classification accuracy. This superiority reflects XLM-R's technical strengths in subword tokenization and self-attention mechanisms, which are well suited to the distinctive characteristics of the Thai language that traditional word-level embeddings often fail to capture. These results are consistent with existing research affirming the capacity of attention-based models to detect nuanced expressions and prioritize relevant keywords in student feedback [12, 13].

From a practical standpoint, the findings offer a promising framework for enhancing institutional quality assurance. The proposed ASI–XLM-R framework enables educational institutions to transition from passive data collection toward proactive, evidence-based decision-making. The framework's high precision in detecting facility- and curriculum-related issues equips PVC to allocate resources more effectively and implement targeted instructional improvements. While challenges in accurately classifying nuanced negative sentiment persist, this study establishes a solid foundation for future research, including integrating sentiment lexicons and data augmentation techniques to refine automated feedback systems further. Ultimately, this work bridges cutting-edge NLP technologies and localized educational management in non-English contexts, with AI-driven tools offering significant potential to enhance student satisfaction and sustain vocational education in an increasingly digital landscape [29].

Acknowledgment

The authors would like to express their gratitude to Phuket Vocational College for providing the student feedback dataset used in this study and to Phuket Rajabhat University for providing computational resources and technical support.

Ethical Statement

The authors declare that this study did not require formal ethical approval because Phuket Vocational College does not mandate Institutional Review Board approval for retrospective analysis of anonymized institutional feedback data collected for administrative purposes. The dataset was fully anonymized prior to analysis, and no personally identifiable information was accessed by the researchers. This exemption follows the college's research ethics policy for secondary use of administrative data.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in in GitHub at https://github.com/nasith/thai-student-feedback-topic-sentiment.

Author Contribution Statement

Tanagrit Chansaeng: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Nasith Laosen:** Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Boonmee Nissaidee:** Validation, Investigation, Resources, Data curation, Writing – review & editing, Project administration. **Pita Jarupunphol:** Conceptualization, Methodology, Software, Validation, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration.

References

[1] Khamphakdee, N., & Seresangtakul, P. (2023). An efficient deep learning for Thai sentiment analysis. *Data*, *8*(5), 1–22. https://doi.org/10.3390/data8050090

[2] Phatthiyaphaibun, W., Chaovavanich, K., Polpanumas, C., Suriyawongkul, A., Lowphansirikul, L., Chormai, P., ..., & Udomcharoenchaikit, C. (2023). PyThaiNLP: Thai natural language processing in Python. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software*, 25–36. https://doi.org/10.18653/v1/2023.nlposs-1.4

[3] Gaurav, A., Gupta, B. B., Sharma, S., Bansal, R., & Chui, K. T. (2024). XLM-RoBERTa based sentiment analysis of tweets on metaverse and 6G. *Procedia Computer Science*, *238*, 902–907. https://doi.org/10.1016/j.procs.2024.06.110

[4] Krichen, M., & Mihoub, A. (2025). Long short-term memory networks: A comprehensive survey. *AI*, *6*(9), 1–21. https://doi.org/10.3390/ai6090215

[5] Daouadi, K. E., Rebaï, R. Z., & Amous, I. (2021). Optimizing semantic deep forest for tweet topic classification. *Information Systems*, *101*, 101801. https://doi.org/10.1016/j.is.2021.101801

[6] Yin, J., Liu, X., & Yang, Z. (2024). A deep multiple-instance text binary classification for topic relevant content extraction on social media. *Journal of King Saud University-Computer and Information Sciences*, *36*(1), 1–10. https://doi.org/10.1016/j.jksuci.2023.101883

[7] Fischer, M. T., Seebacher, D., Sevastjanova, R., Keim, D. A., & El-Assady, M. (2021). CommAID: Visual analytics for communication analysis through interactive dynamics modeling. *Computer Graphics Forum*, *40*(3), 25–36. https://doi.org/10.1111/cgf.14286

[8] Thekkekara, J. P., Yongchareon, S., & Liesaputra, V. (2024). An attention-based CNN-BiLSTM model for depression detection on social media text. *Expert Systems with Applications*, *249*, 1–9. https://doi.org/10.1016/j.eswa.2024.123834

[9] Yan, C., Liu, J., Liu, W., & Liu, X. (2023). Sentiment analysis and topic mining using a novel deep attention-based parallel dual-channel model for online course reviews. *Cognitive Computation*, *15*(1), 304–322. https://doi.org/10.1007/s12559-022-10083-7

[10] Date, S. S., Shelke, M. B., Sonkamble, K. V., & Deshmukh, S. N. (2024). A systematic survey on text-based dimensional sentiment analysis: Advancements, challenges, and future directions. In *Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications*, 39–57. https://doi.org/10.1016/B978-0-443-22009-8.00014-8

[11] Alsemaree, O., Alam, A. S., Gill, S. S., & Uhlig, S. (2024). Sentiment analysis of Arabic social media texts: A machine learning approach to deciphering customer perceptions. *Heliyon*, *10*(9), 1–10. https://doi.org/10.1016/j.heliyon.2024.e27863

[12] Al-Hail, M., Zguir, M. F., & Koç, M. (2023). University students' and educators' perceptions on the use of digital and social media platforms: A sentiment analysis and a multi-country review. *IScience*, *26*(8), 1–27. https://doi.org/10.1016/j.isci.2023.107322

[13] Cuesta-Delgado, D., Barberá-Tomás, D., & Marques, P. (2025). A text-mining analysis of Latin America Universities' mission statements from a "Third Mission" perspective. *Studies in Higher Education*, *50*(7), 1420–1438. https://doi.org/10.1080/03075079.2024.2377371

[14] Grimalt-Álvaro, C., & Usart, M. (2024). Sentiment analysis for formative assessment in higher education: A systematic literature review. *Journal of Computing in Higher Education*, *36*(3), 647–682. https://doi.org/10.1007/s12528-023-09370-5

[15] Lowphansirikul, L., Polpanumas, C., Jantrakulchai, N., & Nutanong, S. (2021). *Wangchanberta: Pretraining transformer-based Thai language models*. arXiv. https://doi.org/10.48550/arXiv.2101.09635

[16] Meena, G., Indian, A., Mohbey, K. K., & Jangid, K. (2024). Point of interest recommendation system using sentiment analysis. *Journal of Information Science Theory and Practice*, *12*(2), 64–78. https://doi.org/10.1633/JISTaP.2024.12.2.5

[17] Indian, A., Manethia, P., Meena, G., & Mohbey, K. K. (2023). Decoding emotions: Unveiling sentiments and sarcasm through text analysis. In *International Conference on Deep Learning, Artificial Intelligence and Robotics*, 714–731. https://doi.org/10.1007/978-3-031-60935-0_62

[18] Arreerard, R., Mander, S., & Piao, S. S. (2022). Survey on Thai NLP language resources and tools. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 6495–6505.

[19] Chotirat, S., & Meesad, P. (2021). Part-of-speech tagging enhancement to natural language processing for Thai wh-question classification with deep learning. *Heliyon*, *7*(10), 1–13. https://doi.org/10.1016/j.heliyon.2021.e08216

[20] Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information*, *15*(9), 1–34. https://doi.org/10.3390/info15090517

[21] Vajrobol, V., Gupta, B. B., & Gaurav, A. (2024). Thai-language chatbot security: Detecting instruction attacks with XLM-RoBERTa and Bi-GRU. *Computers and Electrical Engineering*, *116*, 109186. https://doi.org/10.1016/j.compeleceng.2024.109186

[22] Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2022). XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 258–266.

[23] Reyad, M., Sarhan, A. M., & Arafa, M. (2023). A modified Adam algorithm for deep neural network optimization. *Neural Computing and Applications*, *35*(23), 17095–17112. https://doi.org/10.1007/s00521-023-08568-z

[24] Kalita, J. K., Bhattacharyya, D. K., & Roy, S. (2024). Data preparation. *Fundamentals of Data Science*, 31–46. https://doi.org/10.1016/B978-0-32-391778-0.00010-7

[25] Herrmann, M., Obaidi, M., Chazette, L., & Klünder, J. (2022). On the subjectivity of emotions in software projects: How reliable are pre-labeled data sets for sentiment analysis? *Journal of Systems and Software*, *193*, 111448. https://doi.org/10.1016/j.jss.2022.111448

[26] Li, Z., Guo, Q., Pan, Y., Ding, W., Yu, J., Zhang, Y., . . . , & Xie, Y. (2023). Multi-level correlation mining framework with self-supervised label generation for multimodal sentiment analysis. *Information Fusion*, *99*, 101891. https://doi.org/10.1016/j.inffus.2023.101891

[27] Xie, Z., Lv, Y., Song, Y., & Wang, Q. (2024). Data labeling through the centralities of co-reference networks improves the classification accuracy of scientific papers. *Journal of Informetrics*, *18*(2), 101498. https://doi.org/10.1016/j.joi.2024.101498

[28] Nalisnick, E., Smyth, P., & Tran, D. (2023). A brief tour of deep learning from a statistical perspective. *Annual Review of Statistics and Its Application*, *10*(1), 219–246. https://doi.org/10.1146/annurev-statistics-032921-013738

[29] Mahamad, S., Chin, Y. H., Zulmuksah, N. I. N., Haque, M. M., Shaheen, M., & Nisar, K. (2025). Technical review: Architecting an AI-driven decision support system for enhanced online learning and assessment. *Future Internet*, *17*(9), 1–28. https://doi.org/10.3390/fi17090383

How to Cite: Chansaeng, T., Laosen, N., Nissaidee, B., & Jarupunphol, P. (2026). Automated Sentiment Intelligence for Educational Quality Assurance: A Topic–Sentiment Framework for Thai Student Feedback with Deep Model Instantiations. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62029502