Journal of Current Science and Technology, January-March 2026 Vol. 16 No. 1, Article 149

ISSN 2630-0656 (Online)

Cite this article: Takaew, S., & Romsaiyud, W. (2026). An explainable approach to sentiment analysis of Thai hotel reviews using a fine-tuned language model and SHAP. *Journal of Current Science and Technology*, *16*(1), Article 149. https://doi.org/10.59796/jcst.V15N4.2025.149

Journal of Current Science and Technology

Journal homepage: https://jcst.rsu.ac.th

An Explainable Approach to Sentiment Analysis of Thai Hotel Reviews Using a Fine-Tuned Language Model and SHAP

Sabsathit Takaew*, and Walisa Romsaiyud

School of Science and Technology, Sukhothai Thammathirat Open University, Nonthaburi 11120, Thailand

*Corresponding author; E-mail: 2669600047@stou.ac.th

Received 4 June 2025; Revised 11 July 2025; Accepted 7 August 2025, Published online 20 December 2025

Abstract

Sentiment analysis plays a pivotal role in the hotel industry, where user-generated reviews significantly influence customer decisions. However, traditional machine learning (ML) methods often struggle with the linguistic nuances of languages such as Thai. This study investigates the effectiveness of fine-tuning WangchanBERTa, a monolingual Thai large language model (LLM), for sentiment classification of hotel reviews from Bangkok. The model's performance was compared with ML algorithms, including extreme gradient boosting (XGBoost), support vector machines (SVM), logistic regression (LR), and multinomial naïve Bayes (MNB). The comparison highlights the advantages of deep contextual understanding enabled by transformer-based architecture. To improve interpretability, Shapley Additive Explanation (SHAP) was applied to the best-performing model to analyze feature importance. The results show that the fine-tuned LLM outperformed all ML models, achieving over 92% across all evaluation metrics (accuracy, precision, recall, and F1-score). SHAP analysis enhanced transparency by revealing sentiment drivers relevant to the hotel domain. This study contributes to the advancement of natural language processing (NLP) for low-resource languages by demonstrating the effectiveness of domain-specific fine-tuning combined with explainable artificial intelligence (XAI) in practical applications.

Keywords: *WangchanBERTa; Thai language model; sentiment analysis; SHAP; hotel reviews; low-resource languages*

1. Introduction

Recent advancements in natural language processing (NLP), particularly with the emergence of large language models (LLMs), have revolutionized industrial applications, opened new avenues for research, and reduced dependence on traditional NLP techniques (Kastrati et al., 2025). LLMs trained on extensive datasets have demonstrated exceptional linguistic capabilities across a wide range of NLP tasks (Ozkaya, 2023), including sentiment analysis (e.g., Chaisen et al., 2024; Indira Kumar et al., 2025; Thiengburanathum & Charoenkwan, 2023), named entity recognition (e.g., Saetia et al., 2024), question answering (e.g., Han et al., 2024), text classification (e.g., Gatchalee et al., 2023; Indira Kumar et al., 2025), text summarization (e.g., Büyük, 2024; Sanchan, 2024), sentence augmentation (e.g., Bimagambetova et al., 2023), and machine translation (e.g., Khoboko et al., 2025).

However, the success of LLMs has been largely concentrated in high-resource languages such as English, Chinese, and German (Bimagambetova et al., 2023). In contrast, Thai remains significantly underrepresented in language technology research, despite being spoken by nearly 70 million people worldwide (Arreerard et al., 2022). Thai is classified as a low-resource language (Arreerard et al., 2022; Lowphansirikul et al., 2021), facing notable challenges in the development of robust language models due to limited data availability and domain-specific resources.

Among the foundational LLMs that have driven recent progress in NLP are BERT (bidirectional encoder representations from transformers) (Devlin et al., 2019) and RoBERTa (robustly optimized BERT pretraining approach) (Liu et al., 2019). BERT introduced a bidirectional transformer architecture trained with masked language modeling (MLM) and next sentence prediction, significantly improving performance across various NLP benchmarks. RoBERTa refined this approach by eliminating the next sentence prediction objective and enhancing training strategies through larger batch sizes, longer training durations, and more extensive corpora.

To overcome Thai NLP limitations, researchers have explored two primary approaches. First, multilingual models such as mBERT (Devlin, 2019) and XLM-R with its extended versions XLM-RX and XLM-RXXL (Conneau et al., 2020; Goyal et al., 2021). These models support Thai alongside other languages but show suboptimal performance on Thai-specific tasks (e.g., Gatchalee et al., 2023; Saetia et al., 2024). Second, dedicated Thai language models such as BERT-th, which was trained on 515 MB of Thai Wikipedia data. It offers language-specific optimization but is constrained by limited training corpora and processing capabilities (ThAIKeras, 2018). WangchanBERTa was developed to bridge these gaps as a monolingual Thai language model based on the RoBERTa architecture. It was trained on 78 GB of diverse, cleaned Thai text with support for 416-token sequences (Lowphansirikul et al., 2021).

Despite these methodological advancements, relatively few studies have investigated domain-specific applications of Thai LLMs that incorporate model interpretability, particularly within the hospitality industry. Previous research on Thai sentiment analysis in this domain, such as the work conducted by Khamphakdee, & Seresangtakul (2021b), predominantly employed traditional machine learning (ML) methods, relying on various feature extraction and embedding techniques, including Delta term frequency-inverse document frequency (Delta TF-IDF), term frequency-inverse document frequency (TF-IDF), N-gram models, and Word2Vec embeddings, to represent textual data for sentiment analysis. In a related study, Khamphakdee, & Seresangtakul (2021a) proposed a framework for constructing a Thai sentiment corpus using hotel reviews. Their approach employed cosine similarity to cluster reviews by sentiment and used TF-IDF for feature representation. Khamphakdee, & Seresangtakul (2023) recently conducted a comprehensive comparison of traditional ML, deep learning, hybrid models, and LLMs, incorporating various feature representations such as Word2Vec embeddings and Delta TF-IDF. However, their study did not explore model interpretability, leaving a gap in explainable artificial intelligence (XAI) for this domain. While these approaches have laid a valuable foundation, they are inherently limited in their ability to capture contextual semantics and lack mechanisms for providing transparent explanations of model predictions.

This highlights a critical need for models that not only improve accuracy but also enhance interpretability in domain-specific contexts. Traditional ML methods are foundational but limited because they rely on static features and struggle to capture complex linguistic nuances, especially in morphologically rich languages such as Thai. This need is particularly pronounced in the hotel industry, where user-generated reviews significantly influence customer decisions. To bridge this gap, the present study introduces TRUE (Thai Review Understanding with Explainable Sentiment), a novel framework that combines WangchanBERTa's contextual capabilities with Shapley additive explanation (SHAP). TRUE not only achieves high accuracy but also offers transparent, actionable insights into sentiment drivers. This enables informed decision-making for hospitality stakeholders and advances the state of the art in low-resource language NLP.

The main contributions and results of this study are summarized as follows:

- A monolingual Thai language model was fine-tuned for sentiment classification of hotel reviews.
- A novel framework named TRUE is proposed, which combines the predictive power of a transformer-based model with interpretable outputs, demonstrating the effectiveness of domain-specific fine-tuning and XAI for low-resource languages in practical applications.
- The fine-tuned LLM achieved over 92% across all evaluation metrics, outperforming traditional ML models that achieved below 90%.
- SHAP was used to analyze feature importance and interpret sentiment drivers relevant to the hotel domain.

The remainder of this paper is organized as follows. The materials and methods section outlines the data collection, preprocessing steps, and methodologies employed for sentiment classification. The results section presents performance evaluations and interpretability analyses. Finally, key findings are discussed, and potential directions for future research are proposed.

2. Objectives

This study aims to develop an explainable approach for sentiment analysis of Thai hotel reviews by fine-tuning a Thai language model and integrating SHAP for model interpretation. The objective is to improve sentiment classification accuracy while providing transparent explanations that support practical decision-making in the hotel domain.

3. Materials and Methods

This section outlines the research procedures employed in this study. The overall framework, named TRUE (Thai Review Understanding with Explainable Sentiment), is illustrated in Figure 1. The acronym *TRUE* reflects the framework's objective of achieving authentic understanding by combining the predictive strength of a black-box model with the transparency of interpretable outputs. This framework details the systematic process for developing sentiment classification models using traditional ML methods and a fine-tuning of a monolingual Thai language model. The best-performing model is subsequently interpreted using an XAI technique, specifically SHAP, to enhance model transparency and provide comprehensive insights into sentiment patterns in the Thai hotel industry.

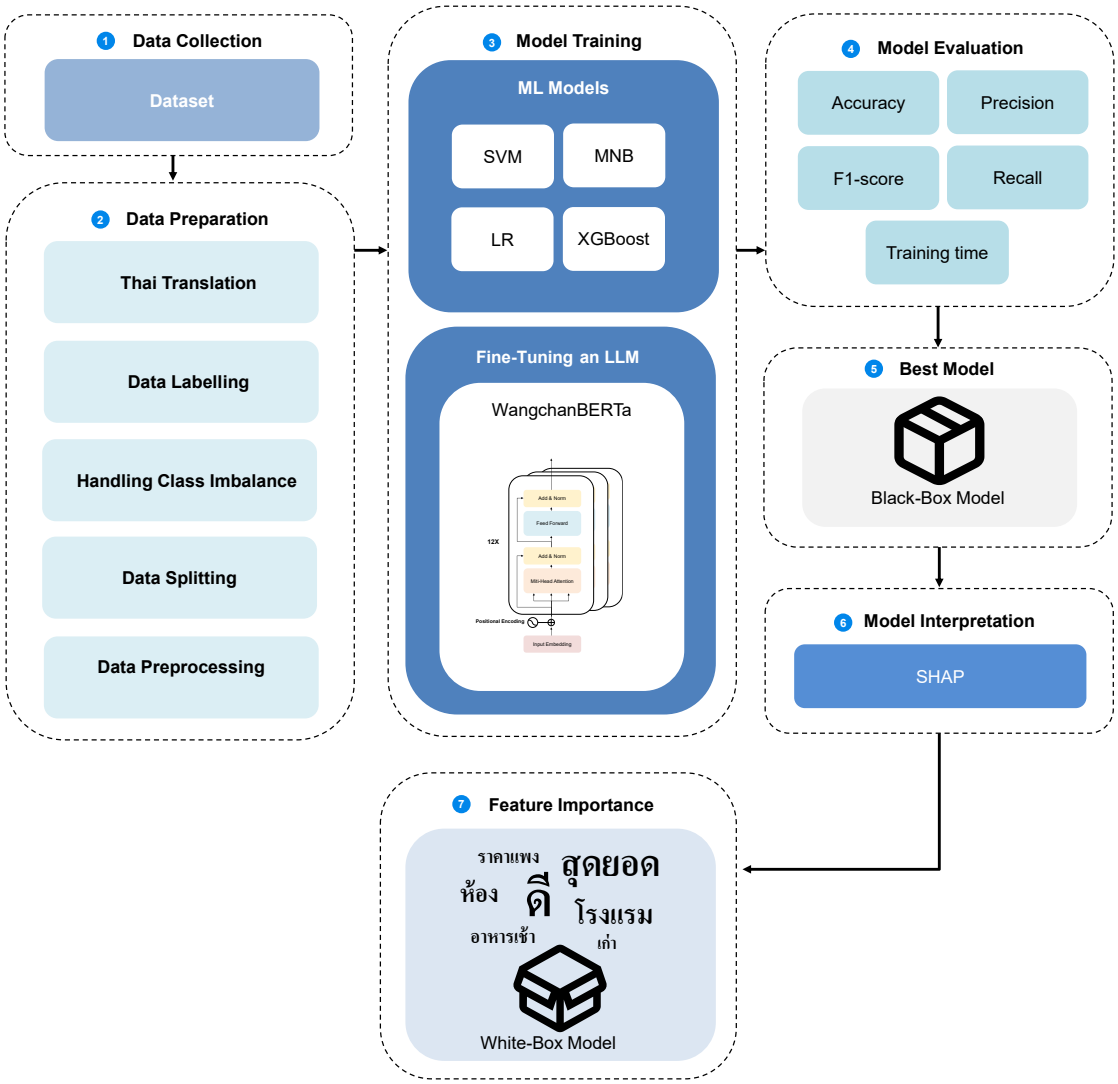

Figure 1 Overview of the TRUE framework showing the complete data pipeline, including data preprocessing, model training, and SHAP-based interpretability.

3.1 Data Collection

The dataset used in this study was sourced from Booking.com by extracting customer reviews from 100 hotels located across various districts in Bangkok, Thailand. Web scraping was performed using the BeautifulSoup library to automate and streamline the data collection process. Key review attributes were collected, including review title, rating, reviewer's country of origin, hotel district, and other relevant details. In total, 33,789 unlabeled hotel reviews were collected between January 2022 and December 2023.

The reviews reflect contemporary Thai language usage, including informal expressions and social media-style phrasing commonly found in user-generated content. This linguistic diversity captures the natural variation in how customers express their opinions about hotel experiences, providing a representative sample of real-world Thai language patterns in the hospitality domain. An overview of the dataset variables is presented in Table 1.

3.2 Thai Translation

The collected reviews were originally written in various languages, including Chinese, Spanish, English, German, and others. However, this study focuses exclusively on Thai-language reviews. To ensure consistency in language processing, all non-Thai reviews were translated into Thai using the GoogleTranslator function from the deep_translator library. The translation process was configured with 'auto' as the source language and 'th' as the target language, enabling automatic detection of the original language and seamless translation into Thai. This step ensured a unified linguistic dataset suitable for training a monolingual Thai sentiment analysis model.

3.3 Data Labeling

To prepare the dataset for supervised learning, sentiment labels were assigned based on the review ratings. Reviews were categorized into two sentiment classes: positive (1) and negative (0). Specifically, reviews with a rating of 7 or higher were labeled as positive, while those with a rating below 7 were classified as negative. This binary classification scheme enables a clear distinction between satisfactory and unsatisfactory hotel experiences and is well-suited for evaluating sentiment analysis performance. This labeled dataset served as the foundation for training and evaluating classification models.

While this rating-based labeling approach enables efficient large-scale annotation, it may introduce noise when ratings do not align with the expressed sentiment. To assess its reliability, a manual review of randomly selected samples was conducted by a native Thai speaker with relevant expertise. The majority of labels were found to be consistent with textual sentiment. However, some residual noise may persist, particularly in reviews with ambiguous tone or mixed opinions. Table 2 shows examples of labeled reviews and their assigned sentiment classes.

3.4 Handling Data Imbalance

The original dataset exhibited a significant imbalance between positive and negative reviews. Specifically, there were 27,774 positive reviews and only 6,015 negative reviews. This imbalance posed a risk of biasing the model toward the majority class, potentially impairing its generalization capability and reducing performance on the minority class.

To address this issue, four resampling techniques were considered, as previously investigated by Simmachan & Boonkrong (2025): (1) retaining the original imbalanced dataset, (2) under-sampling the majority class, (3) over-sampling the minority class, and (4) a hybrid approach combining both under-sampling and over-sampling. Their study showed that these techniques significantly improved classification performance, particularly in terms of F1-score and balanced accuracy. These findings highlight the importance of selecting appropriate resampling strategies when working with imbalanced datasets in classification tasks.

In this study, a random under-sampling (RUS) technique was adopted to balance the dataset while minimizing the introduction of noise and preserving data reliability. As shown in Figure 2, this technique randomly removed samples from the majority (positive) class until its count matched that of the minority (negative) class. According to Almuayqil et al. (2022), such strategies are particularly valuable when handling large datasets, as they also help reduce computational complexity during training.

Following the application of RUS, the dataset was balanced to consist of 12,000 reviews, with 6,000 positive and 6,000 negative reviews. The resulting balanced dataset, illustrated in Figure 3, provided a more robust foundation for model training and reduced the risk of majority-class bias. However, a key limitation of RUS is that it may discard informative samples from the majority class, potentially resulting in the loss of valuable contextual patterns that could enhance the model's generalization performance. The trade-off between simplicity and information retention must be considered when applying RUS to text classification tasks.

Table 1 Overview of the variables included in the hotel review dataset.

Variable	Description	Data type
review_id	A unique identifier assigned to each hotel review.	Integer
review_title	The headline provided by the reviewer.	String
review_rating	The overall rating out of 10 reflecting the hotel experience.	Integer
review_date	The date when the review was posted.	Date
reviewer_country	The country of origin of the reviewer.	String
review	The detailed feedback provided by the reviewer about their hotel experience.	String
hotel_district	The district in Bangkok where the hotel is located.	String
hotel_name	The official name of the hotel reviewed.	String

Table 2 Examples of hotel reviews with assigned binary sentiment labels.

Review	Review rating	Label (sentiment)
"ที่พักเก่าในกรุงเทพมหานครรายล้อมด้วยสวน แต่สภาพโดยรวมไม่คุ้มกับราคา ห้องพักดูเก่าและไม่สะอาด เฟอร์นิเจอร์มีร่องรอยการใช้งานและบางส่วนชำรุด ทำให้ประสบการณ์เข้าพักไม่เป็นไปตามที่คาดหวัง" *("An old accommodation in Bangkok surrounded by gardens, but overall, not worth the price. The room appeared outdated and was not very clean, with furniture showing signs of wear and some damage, leading to a stay that did not meet expectations.")*	3	0 (negative)
"เป็นปาร์ตี้ส่งท้ายปีเก่าที่ยอดเยี่ยมมาก พนักงานมืออาชีพ อาหารเช้าอร่อย ทำเลดีใกล้เซ็นทรัลเวิลด์ สะอาดมากทุกอย่างสมบูรณ์แบบสำหรับเรา" *("It was a fantastic New Year's Eve party. The staff were professional, the breakfast was delicious, and the location was great, close to CentralWorld. The place was very clean, and everything was perfect for us.")*	10	1 (positive)

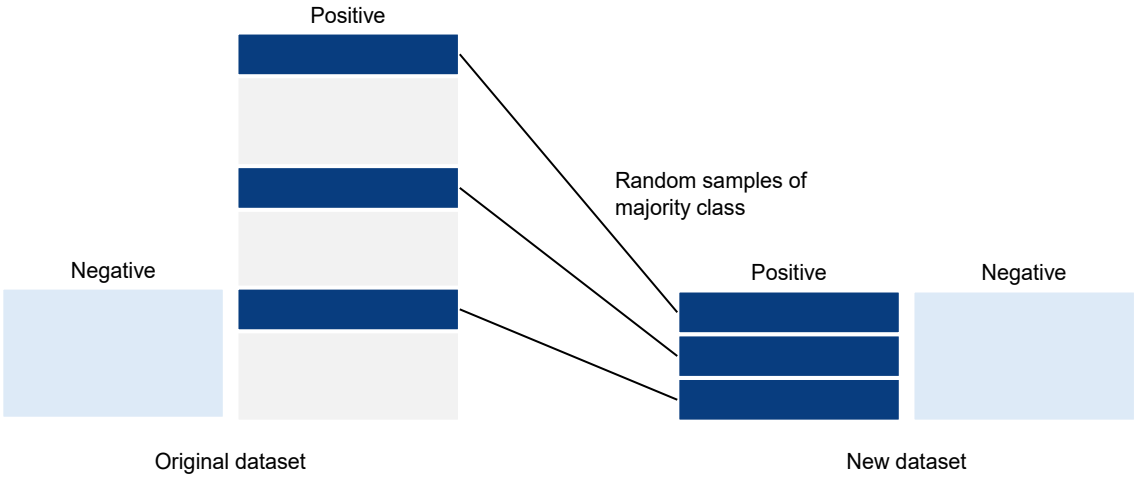

Figure 2 Illustration of the random under-sampling (RUS) technique used to balance the dataset by reducing majority-class samples.

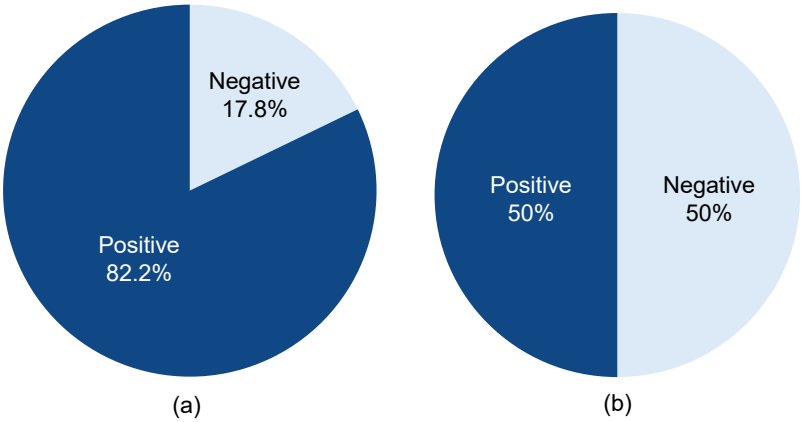

Figure 3 Class distribution of the dataset before (a) and after (b) applying random under-sampling (RUS).

3.5 Data Splitting

The final balanced dataset was divided using an 80:10:10 split for training, validation, and testing, respectively. This stratified split ensured that each subset maintained an equal distribution of positive and negative reviews, supporting reliable model development and fair evaluation. The choice of the 80:10:10 split was informed by prior research, such as Golchubian et al. (2021), which demonstrated that this ratio offers a better trade-off between training robustness and model generalization compared to more conservative splits such as 60:20:20. Specifically, it enables the model to learn effectively from a large training set while reserving sufficient data for unbiased validation and testing.

Although k-fold cross-validation is generally considered more robust for assessing model generalizability, it was not used for this study due to the significant computational cost and time required for training. Additionally, the hold-out method was selected to maintain a consistent evaluation set across all model comparisons, ensuring fairness and reproducibility. This approach represents a practical compromise, reducing training time while maintaining separate datasets for model tuning and evaluation.

3.6 Data Preprocessing

To ensure the quality and consistency of textual input, a comprehensive preprocessing pipeline was applied prior to model training. This procedure followed the methodologies established by Lowphansirikul et al. (2021) and incorporated additional best practices from related studies.

First, HTML character entities were replaced with their corresponding characters, line breaks were converted into spaces, and multiple consecutive spaces were consolidated into a single space. The process also involved removing empty brackets such as (), {}, and [], which often appear as artifacts from text extraction, particularly from sources such as Wikipedia (Lowphansirikul et al., 2021).

Additionally, repetitive characters were handled effectively (Howard & Ruder, 2018; Lowphansirikul et al., 2021) by reducing sequences of more than three identical characters or emojis to enhance text consistency. For example, expressions like "ดีมากกกก" (very gooddd) were normalized to "ดีมาก" (very good), and repeated emojis were reduced to a single instance.

For tokenization, a word-level method was used, employing the newmm dictionary-based maximal matching tokenizer (Phatthiyaphaibun et al., 2020), which is well-suited for Thai text segmentation. Since Thai does not use spaces to separate words (Lowphansirikul et al., 2021), tokenization was a critical preprocessing step. Repeated words were addressed after tokenization rather than before, due to the absence of explicit word boundaries in Thai, unlike English.

Lastly, for the fine-tuning method, spaces were replaced with <_> to explicitly denote word breaks, as they function as punctuation in Thai, indicating sentence boundaries similar to periods in English. This ensured that the SentencePiece tokenizer could correctly process the text for tasks such as word segmentation and sentence breaking (Lowphansirikul et al., 2021). These preprocessing steps collectively ensured that the dataset was of high quality and well-suited for subsequent stages of model development.

3.7 Model Training

This study evaluates Thai sentiment classification using two approaches: traditional ML methods and fine-tuning an LLM. Despite the strong performance of transformer-based models in NLP tasks, traditional ML models were included as baselines to highlight the performance gains from fine-tuning and to reflect their continued relevance in low-resource or time-sensitive settings. All experiments were conducted in Google Colab to ensure consistency; traditional ML models were trained on a central processing unit (CPU), while fine-tuning the LLM was performed using a graphics processing unit (GPU) due to its greater computational requirements. The following subsections outline the training procedures for both approaches.

3.7.1 Traditional ML Methods

The four selected classifiers represented a diverse set of ML algorithms: (1) extreme gradient boosting (XGBoost), (2) support vector machine (SVM), (3) logistic regression (LR), and (4) multinomial naïve Bayes (MNB). These models were chosen to enable a comprehensive evaluation across different algorithmic paradigms commonly used in text classification. XGBoost was included for its powerful ensemble learning capabilities and ability to model complex such as non-linear relationships. SVM is known for its strong generalization performance in high-dimensional settings. LR provides a simple yet effective linear baseline that is easy to interpret, while MNB, a probabilistic classifier, is particularly well-suited for handling word frequency features typical of text data. This diverse selection facilitated a balanced comparison between advanced and interpretable models, as well as between deterministic and probabilistic approaches. Each classifier's performance was optimized through hyperparameter tuning using GridSearchCV, which systematically explored a predefined set of parameter values to identify the optimal configuration. The results of this tuning are summarized in Table 3.

To enhance computational efficiency and reduce the dimensionality of the feature space, several preprocessing optimizations were applied to the TF-IDF representation. Specifically, the parameters ngram_range, min_df, and max_df were adjusted to limit the number of features while preserving essential textual patterns. A unigram model was selected, with min_df = 5 and max_df = 0.9, resulting in a final feature set comprising 4,065 features.

These configurations enabled the models to learn effectively from the data while minimizing overfitting and reducing computational overhead. The combination of systematic feature reduction and hyperparameter optimization ensured that each model was tuned for maximum performance in the sentiment classification task.

3.7.2 Fine-tuning an LLM

WangchanBERTa, developed by Lowphansirikul et al. (2021), is a Thai-language model based on the RoBERTa-base architecture proposed by Liu et al. (2019). While BERT initially introduced a groundbreaking transformer-based framework capable of capturing bidirectional context through self-attention mechanisms, RoBERTa refined this approach by improving key training strategies, resulting in enhanced performance on downstream NLP tasks. Building on these advancements, WangchanBERTa was specifically designed to handle the linguistic complexities of Thai, including its lack of word boundaries, tonal structure, and rich morphology.

To comprehensively capture the grammatical, semantic, and contextual nuances of Thai, WangchanBERTa was pre-trained on a large and diverse corpus spanning various domains and genres. The pre-training task followed the MLM objective, wherein certain tokens are randomly masked and the model learns to predict them using contextual cues from both directions. This training strategy enabled the development of deep contextual representations tailored for Thai.

Table 3 Hyperparameter configurations and optimal parameter selections for each machine learning model.

Model	Hyperparameter	Parameter list	Best parameter
LR	C	[0.1, 1, 10]	1
	solver	['liblinear', 'lbfgs']	lbfgs
MNB	alpha	[0.01, 0.1, 0.5, 1, 2, 5]	5
SVM	C	[0.1, 1, 10, 100]	1
	kernel	['rbf']	rbf
	gamma	['scale']	scale
XGBoost	subsample	[0.5, 0.7, 1]	0.5
	learning_rate	[0.1, 0.01, 0.001]	0.1
	max_depth	[3, 5, 7]	7

When applying pre-trained language models to downstream tasks, two primary strategies are commonly used: feature-based and fine-tuning approaches (Devlin et al., 2019). The feature-based approach, exemplified by ELMo (Peters et al., 2018), incorporates pre-trained representations as additional input features into task-specific architectures. In contrast, the fine-tuning approach, as demonstrated by models such as OpenAI's GPT (Radford et al., 2018), involves updating pre-trained parameters during task-specific training with minimal architectural modification. Fine-tuning has shown strong effectiveness in adapting pre-trained models to a wide range of NLP tasks, even with limited labeled data (Howard & Ruder, 2018).

In this study, sentiment classification was conducted using Hugging Face's Transformers pipeline. The model used was wangchanberta-base-att-spm-uncased, which was fine-tuned on a labeled training dataset to improve sentiment detection capabilities. The SentencePiece tokenizer, retrieved from the Hugging Face model hub, was used to preprocess and tokenize the Thai input text. Key model parameters, specifically classifier.dense.bias, classifier.dense.weight, classifier.out_proj.bias, and classifier.out_proj.weight, were fine-tuned using the training set to optimize performance. To ensure efficient handling of longer sequences, the maximum token length was set to 416. A summary of the model's training configuration is provided in Table 4.

Table 4 Training parameter configuration for the fine-tuned WangchanBERTa model.

Parameter	Value
max_steps	1600
per_device_train_batch_size	2
gradient_accumulation_steps	16
learning_rate	1e-05
warmup_steps	75
weight_decay	0.1
adam_epsilon	1e-08
max_grad_norm	1.0
metric_for_best_model	f1_micro

The architecture for Thai sentiment classification, along with the fine-tuning process for the LLM, is illustrated in Figure 4. This figure outlines how the dataset is used during fine-tuning and highlights the inference process. The figure also presents an example of model inference, showing how the fine-tuned model classifies an input sentence into predefined sentiment categories. In particular, the last hidden state, labeled as H1, corresponds to the special token added at the beginning of each input sequence to represent the entire sentence for classification tasks. This hidden state is then passed as the input to the classification head to generate the final sentiment prediction.

4. Results

4.1 Model Performance

A confusion matrix was used to evaluate the classification results, consisting of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), as illustrated in Figure 5.

This evaluation method provided a clear basis for assessing each model's performance in binary sentiment classification. Accuracy, precision, recall, and F1-score were calculated using Equations 1–4 to measure classification effectiveness. To assess computational efficiency, training time was also recorded for each model. The results, including both performance metrics and training time, are summarized in Table 5.

Macro-averaging was used to ensure equal consideration of both sentiment classes. Each metric was computed independently for the positive and negative classes and then averaged, providing a fair and consistent evaluation, as shown in Equations (1)–(4).

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

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

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

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

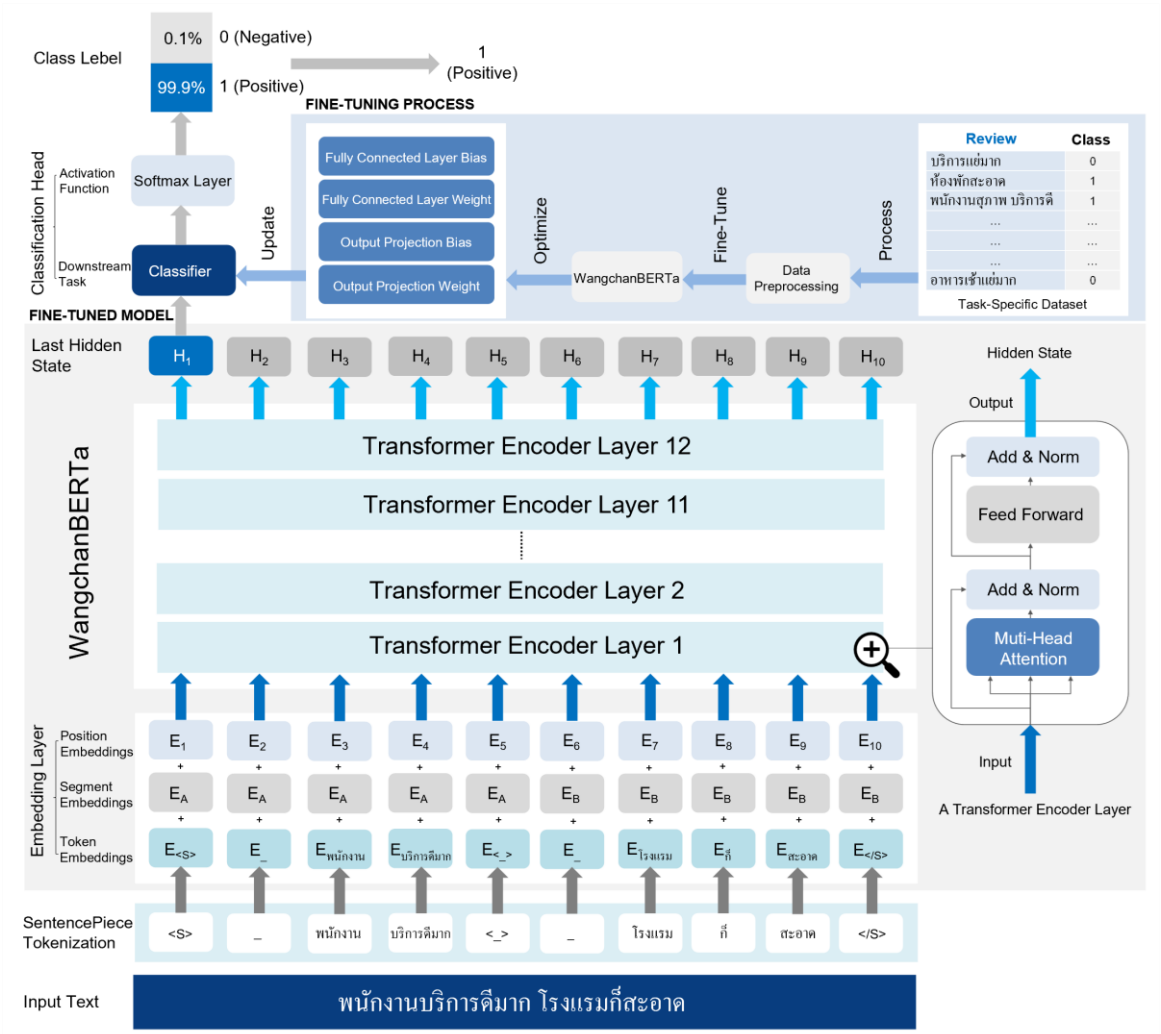

Figure 4 Architecture of the WangchanBERTa-based sentiment classification model, including the fine-tuning workflow and an example inference pipeline.

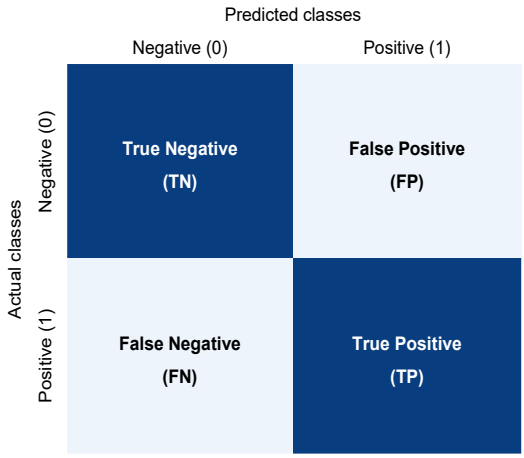

Figure 5 Structure of the confusion matrix used to evaluate binary sentiment classification (TP, FP, TN, FN).

As shown in Table 5, the WangchanBERTa model outperformed all other models across evaluation metrics for Thai sentiment classification, achieving over 92%. These results significantly exceeded those of the traditional ML models. The next best performer, SVM, achieved an accuracy of 87.50%, which is 4.58% lower than that of WangchanBERTa.

However, this superior performance came with significant computational cost. WangchanBERTa required 1,516.20 seconds for training, much longer than any traditional ML model. For comparison, LR trained in only 0.08 seconds, MNB in 0.01 seconds, and XGBoost in 15.30 seconds. Even SVM, the slowest among traditional ML models at 130.00 seconds, was over 11 times faster than WangchanBERTa. Importantly, traditional ML models were trained using the CPU, while WangchanBERTa leveraged the GPU for its training. These results show that although transformer-based models achieve state-of-the-art accuracy, they also require significantly more computational resources, particularly the parallel processing power of the GPU, for both training and inference.

Further insights are provided in Table 6, which summarizes classification metrics, including the number of correct and incorrect predictions per class, and the AUC-ROC (area under the receiver operating characteristic curve) values. WangchanBERTa again demonstrated superior performance with 567 correct predictions for class 1 (positive sentiment) and 538 correct predictions for class 0 (negative sentiment). It recorded the fewest errors with 62 false positives and 33 false negatives and achieved the highest AUC-ROC score of 0.9691, indicating strong discriminative ability between sentiment classes.

For comparison, the SVM model produced 545 correct predictions for class 1 and 505 for class 0, with an AUC-ROC of 0.9390. Other traditional ML models, such as LR and MNB, showed lower performance and higher error rates. These findings are consistent with Table 5 and further confirm WangchanBERTa's superior performance across both standard evaluation metrics and detailed diagnostic analyses.

The receiver operating characteristic (ROC) curves for traditional ML models are shown in Figure 6, while WangchanBERTa's ROC curve is displayed separately in Figure 7 to highlight its superior discriminative power. The larger area under the curve for WangchanBERTa visually supports its capability to distinguish between sentiment classes with high confidence.

While WangchanBERTa delivers the best overall performance in Thai sentiment classification, its computational demands may limit practical use in real-time or resource-constrained environments. In scenarios where fast inference and minimal resource consumption are priorities, a traditional ML model such as XGBoost may offer an optimal balance between performance and efficiency.

Table 5 Performance metrics of machine learning and transformer-based models on Thai sentiment classification.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Training Time (s)
LR	86.42	86.51	86.42	86.41	0.08
MNB	83.17	83.18	83.17	83.17	**0.01**
SVM	87.50	87.67	87.50	87.49	130.00
XGBoost	87.00	87.09	87.00	86.99	15.30
WangchanBERTa	**92.08**	**92.18**	**92.08**	**92.08**	1516.20

Table 6 Detailed classification results, including confusion matrix components and AUC-ROC scores.

Model	Correct class 1	Correct class 0	Incorrect class 0	Incorrect class 1	AUC-ROC
LR	534	503	97	66	0.9325
MNB	504	494	106	96	0.9110
SVM	545	505	95	55	0.9390
XGBoost	537	507	93	93	0.9370
WangchanBERTa	**567**	**538**	**62**	**33**	**0.9691**

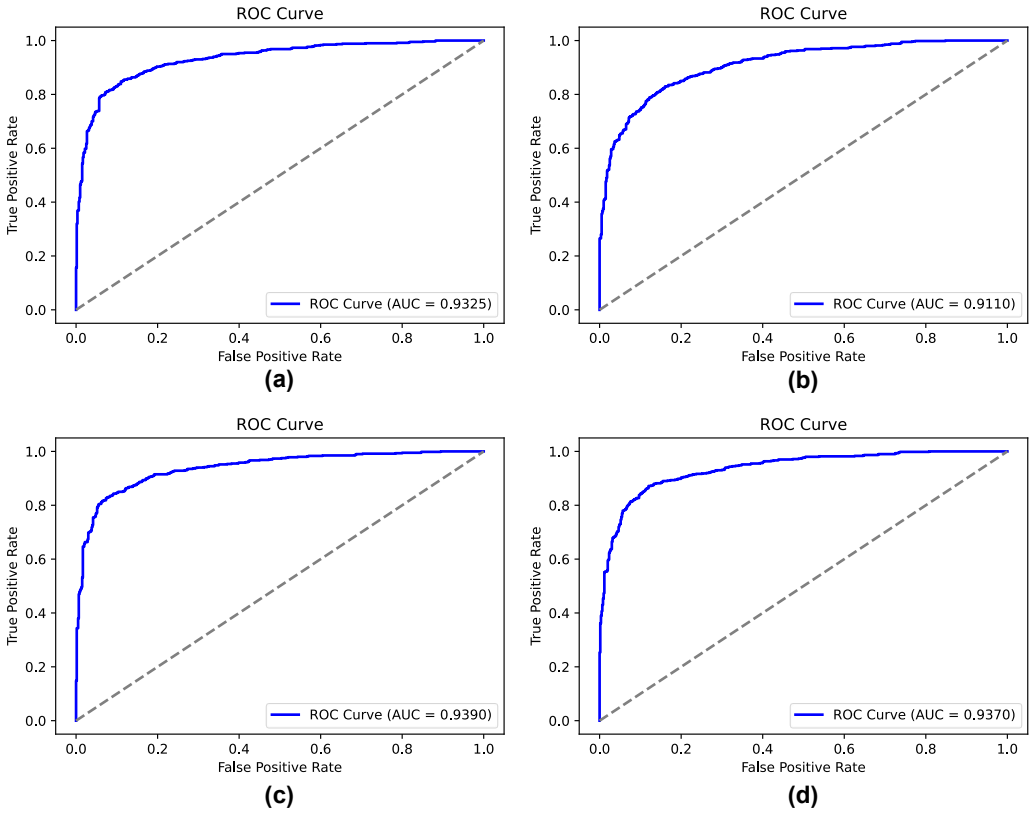

Figure 6 Receiver operating characteristic (ROC) curves for traditional machine learning models: (a) LR, (b) MNB, (c) SVM, and (d) XGBoost.

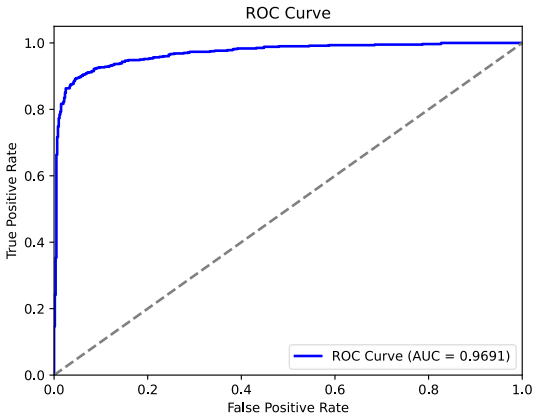

Figure 7 ROC curve of the WangchanBERTa model.

4.2 Effect of Data Imbalance

To evaluate the robustness of WangchanBERTa, its performance was reassessed using the original dataset, without any balancing techniques such as RUS being applied. The model was fine-tuned using the same configuration described in Table 4, ensuring consistency across both balanced and imbalanced conditions.

As shown in Table 7, although accuracy on the imbalanced dataset slightly improved from 92.08% to 93.55%, other metrics noticeably declined. This performance degradation indicates that while the model correctly identified more overall samples, it became less effective at detecting minority class instances, which is a typical challenge in imbalanced scenarios. The decline in metrics such as recall and

F1-score demonstrates that the model's ability to capture negative sentiment instances was compromised when trained on the original dataset, highlighting the importance of addressing class imbalance to ensure robust model performance.

Table 7 Comparison of WangchanBERTa performance on imbalanced versus RUS-balanced datasets.

Metric (%)	Original dataset	Balanced dataset (with RUS)
Accuracy	**93.55**	92.08
Precision	90.24	**92.18**
Recall	87.00	**92.08**
F1-score	88.50	**92.08**

4.3 Model Interpretability

Model interpretability is essential for understanding how an ML model generates its predictions, particularly in domains where transparency and accountability are critical. In this study, interpretability was a central focus to ensure that the influence of each input feature on the model's output could be clearly understood and effectively communicated.

Despite its strong performance, WangchanBERTa is considered a black box due to its deep architecture and complex nonlinear feature interactions. To address this, XAI techniques were employed to enhance transparency. Among the available XAI methods, LIME (Local Interpretable Model-agnostic Explanations) and SHAP are the most prominent. While both are valuable for interpreting ML models, SHAP is generally preferred from a theoretical standpoint because it provides mathematically grounded guarantees of consistency and accuracy.

In this study, SHAP was applied to the best-performing model using three representative test cases. SHAP analysis revealed how specific input features, particularly individual words, contributed to the model's sentiment predictions. By uncovering the decision-making rationale behind WangchanBERTa's outputs, SHAP analysis significantly enhanced the model's interpretability and facilitated more transparent and trustworthy sentiment analysis.

Table 8 presents the top five feature contributions to the model's output for Case I, based on SHAP analysis. These results provide insight into how individual words influence the prediction, with the actual label being positive.

Table 8 Top five feature contributions to model output for Case I.

Feature name	SHAP value
ยอดเยี่ยม (excellent)	+1.165
มหัศจรรย์ (wonderful)	+1.014
เป็นเลิศ (outstanding)	+0.593
ทุกครั้ง (every time)	+0.508
แห่งนี้ (this)	+0.429

The sentence for this case is:

*"**มหัศจรรย์**เราชอบโรงแรม**แห่งนี้**มากและจะพักที่นี่**ทุกครั้ง**ที่มากรุงเทพมหานคร พนักงานทุกคน**เป็นเลิศ**และเป็นมิตร อาหารเช้าก็**ยอดเยี่ยม**เช่นกัน"*

*("**Wonderful**, we really like **this** hotel and will stay here **every time** we visit Bangkok. All the staff are **outstanding** and friendly. The breakfast is also **excellent**.")*

The base value of the model's prediction is 0.634, representing the expected output without any input features. The sum of the positive feature contributions is 4.351, reflecting the combined effect of words that increased the predicted sentiment score. This results in a final predicted sentiment score significantly higher than the baseline, indicating strong positive sentiment. Notably, the actual label for this case is also positive, confirming the model's accurate prediction. SHAP values clearly illustrate how each word contributed to the prediction.

Table 9 presents the top five feature contributions for Case II, where both the sentiment prediction and the actual label are negative.

Table 9 Top five feature contributions to model output for Case II.

Feature name	SHAP value
สกปรก (dirty)	–1.236
ฉันจะไม่ (I will not)	–1.148
ห้อง (room)	–0.589
อื่น (another)	–0.569
เก่า (old)	–0.443

The sentence for this case is:

*"**ฉันจะไม่**อยู่อีกครั้ง ไม่มีอะไรพิเศษมากนัก แค่**ห้อง**เล็ก ๆ ที่มองเห็นผนังอาคาร**อื่น**ได้ **เก่า**ไปหน่อย สระว่ายน้ำ**สกปรก**"*

*("**I will not** stay again. Nothing special, just a small **room** overlooking the wall of **another** building. It's a bit **old**, and the swimming pool is **dirty**.")*

The model's base value is 0.781, and the total feature contribution is −3.188, reflecting a strong negative shift in sentiment due to the influence of negatively charged words. As a result, the final sentiment prediction is significantly lower than the baseline, indicating a clearly negative sentiment. The model's prediction aligns with the actual label, confirming its accuracy.

Table 10 highlights the SHAP analysis for Case III, which includes both positive and negative features, while the actual label remains positive.

Table 10 Top five feature contributions to model output for Case III

Feature name	SHAP value
สวย (beautiful)	+1.012
ดีที่สุด (the best)	+0.939
สถานที่ (place)	+0.636
มีราคาแพง (expensive)	-0.583
ตลอด (throughout)	+0.497

The sentence for this case is:
*"ที่นี่เป็น**สถานที่ที่ดีที่สุด**ที่เราพัก**ตลอด**การเดินทาง เตียงนุ่ม บริเวณสระว่ายน้ำ**สวย** แต่บาร์บนดาดฟ้า**มีราคาแพง**"*
*("The **place** is **the best** place we stayed **throughout** the trip. The bed is soft, the pool area is **beautiful**, but the rooftop bar's menu is **expensive**.")*

The base value is again 0.781, and the sum of the feature contributions is +3.192. Despite the negative impact of *"มีราคาแพง"* (expensive), the overall sentiment is strongly positive due to other influential positive terms. The SHAP values provide a detailed, interpretable breakdown of how each feature impacted the final prediction.

The SHAP analysis reveals how specific Thai words influence WangchanBERTa's sentiment predictions. This enhanced interpretability makes the model's decision-making process transparent and trustworthy. By showing which features contribute positively or negatively to predictions, the TRUE framework bridges the gap between complex modeling and practical applications and offers actionable insights for stakeholders such as hotel managers.

TRUE combines state-of-the-art performance with explainability techniques specifically designed for Thai sentiment analysis. This approach emphasizes both high classification accuracy and practical interpretability to support a better understanding and deployment.

WangchanBERTa outperformed traditional ML models by 4.58% to 8.91% in accuracy, demonstrating the effectiveness of this monolingual transformer-based model for Thai sentiment analysis. This finding aligns with the work of Harnmetta & Samanchuen (2022), which showed that monolingual models outperform both multilingual and traditional ML models on single-language datasets. They achieved accuracy of over 92%, while other models were below 90%.

As shown in Table 7, training on the imbalanced dataset slightly increased accuracy, but other metrics dropped significantly. This indicates the model identified more samples overall but was less effective at detecting minority class instances. This highlights the critical need to address class imbalance for robust performance.

The SHAP analysis within the TRUE framework offers valuable insights into the features driving predictions, enabling targeted fine-tuning to further improve performance and robustness. Fine-tuning WangchanBERTa with domain-specific data or more nuanced sentiment labels may enhance its ability to handle complex or infrequent sentiment expressions, thereby broadening its applicability.

While these results are promising, the generalizability of the TRUE framework's performance beyond the hospitality domain remains unclear. Variations in language use, sentiment expression, and domain-specific vocabulary may affect model performance in other contexts, requiring additional adaptation or fine-tuning. Moreover, potential biases introduced during translation processes or data labeling, such as subjective interpretation of sentiment or annotation inconsistencies, could influence model outcomes. Addressing these biases will be crucial for applying the framework reliably in broader settings.

In summary, the TRUE framework presents a well-balanced approach that integrates a high-performing transformer model with explainable techniques, ensuring both accuracy and transparency in Thai sentiment classification.

5. Conclusion

This study contributes to the advancement of NLP for low-resource languages by demonstrating the effectiveness of domain-specific fine-tuning combined with XAI in practical applications. Fine-tuning WangchanBERTa for Thai sentiment classification in hotel reviews led to significant performance improvements over traditional ML models across all

evaluation metrics. This confirms the model's suitability for handling nuanced sentiment tasks in the Thai language.

To enhance interpretability and user trust, SHAP was applied to the best-performing model, offering clear explanations of how individual features influenced predictions. This integration of an XAI technique with a complex transformer-based model strengthens transparency and trustworthiness, critical aspects for real-world deployment. While fine-tuning LLM requires substantial computational resources, the trade-off is justified by improvements in both predictive accuracy and interpretability.

The practical implications of this research extend beyond hotel reviews to other service sectors such as transportation, retail, and mobile applications, where understanding user sentiment is vital. Future research will explore systematic fine-tuning of LLMs for specialized domains such as healthcare, consumer behavior, and government services, each presenting unique linguistic and contextual challenges.

Moreover, advancing XAI methods tailored to the linguistic and cultural characteristics of Thai and other low-resource languages will be key to developing NLP systems that are not only high-performing but also transparent, trustworthy, and sustainable. Such systems have the potential to support informed decision-making and enhance user experience across diverse real-world applications.

6. Abbreviations

Abbreviation	Full Term
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BERT	Bidirectional Encoder Representations from Transformers
CPU	Central Processing Unit
Delta TF-IDF	Delta Term Frequency–Inverse Document Frequency
FN	False Negative
FP	False Positive
GPU	Graphics Processing Unit
GPT	Generative Pre-trained Transformer
LIME	Local Interpretable Model-Agnostic Explanation
LLM	Large Language Model
LR	Logistic Regression
ML	Machine Learning
MLM	Masked Language Modeling
MNB	Multinomial Naïve Bayes
NLP	Natural Language Processing
ROC	Receiver Operating Characteristic
RoBERTa	Robustly Optimized BERT Pretraining Approach
RUS	Random Under-Sampling
SHAP	Shapley Additive Explanation
SVM	Support Vector Machines
TF-IDF	Term Frequency–Inverse Document Frequency
TN	True Negative
TP	True Positive
TRUE	Thai Review Understanding with Explainable Sentiment
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

7. CRediT Statement

Sabsathit Takaew: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing – Original Draft, Visualization, Project Administration.

Walisa Romsaiyud: Conceptualization, Methodology, Validation, Formal Analysis, Writing – Review & Editing, Supervision.

8. References

Almuayqil, S. N., Humayun, M., Jhanjhi, N. Z., Almufareh, M. F., & Khan, N. A. (2022). Enhancing sentiment analysis via random majority under-sampling with reduced time complexity for classifying tweet reviews. *Electronics*, *11*(21), Article 3624. https://doi.org/10.3390/electronics11213624

Arreerard, R., Mander, S., & Piao, S. (2022). Survey on Thai NLP language resources and tools [Conference presentation]. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France. https://aclanthology.org/2022.lrec-1.697/

Bimagambetova, Z., Rakhymzhanov, D., Jaxylykova, A., & Pak, A. (2023). Evaluating large language models for sentence augmentation in low-resource languages: A case study on Kazakh [Conference presentation]. *2023 19th International Asian School-Seminar on Optimization Problems of Complex Systems (OPCS)*, Moscow, Russia. https://doi.org/10.1109/OPCS59592.2023.10275753

Büyük, O. (2024). A comprehensive evaluation of large language models for turkish abstractive dialogue summarization. *IEEE Access*, *12*,

124391–124401. https://doi.org/10.1109/ACCESS.2024.3454342

Chaisen, T., Charoenkwan, P., Kim, C. G., & Thiengburanathum, P. (2024). A zero-shot interpretable framework for sentiment polarity extraction. *IEEE Access*, *12*, 10586–10607. https://doi.org/10.1109/ACCESS.2023.3322103

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale [Conference presentation]. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. https://doi.org/10.18653/v1/2020.acl-main.747

Devlin, J. (2019). *Bert/multilingual.md*. Retrieved from https://github.com/google-research/bert/blob/master/multilingual.md#list-of-languages

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding [Conference presentation]. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, United States. https://doi.org/10.18653/v1/N19-1423

Gatchalee, P., Waijanya, S., & Promrit, N. (2023). Thai text classification experiment using CNN and transformer models for timely-timeless content marketing. *ICIC Express Letters*, *17*(01), 91–101. https://doi.org/10.24507/icicel.17.01.91

Golchubian, A., Marques, O., & Nojoumian, M. (2021). Photo quality classification using deep learning. *Multimedia Tools and Applications, 80*(14), 22193–22208. https://doi.org/10.1007/s11042-021-10766-7

Goyal, N., Du, J., Ott, M., Anantharaman, G., & Conneau, A. (2021). Larger-scale transformers for multilingual masked language modeling [Conference presentation]. *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, Bangkok, Thailand. https://doi.org/10.18653/v1/2021.repl4nlp-1.4

Han, J., Yang, T., Zhang, C., & Jiang, H. (2024). Fine-tuning large language models for the electric power industry with specialized Chinese electric power corpora [Conference presentation]. *2024 5th International Conference on Computers and Artificial Intelligence Technology (CAIT)*, Zhejiang, China. https://doi.org/10.1109/CAIT64506.2024.10963259

Harnmetta, P., & Samanchuen, T. (2022). Sentiment analysis of Thai stock reviews using transformer models [Conference presentation]. *2022 19th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Bangkok, Thailand. https://doi.org/10.1109/JCSSE54890.2022.9836278

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification [Conference presentation]. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia. https://doi.org/10.18653/v1/P18-1031

Indira Kumar, A. K., Sthanusubramoniani, G., Gupta, D., Nair, A. R., Alotaibi, Y. A., & Zakariah, M. (2025). Multi-task detection of harmful content in code-mixed meme captions using large language models with zero-shot, few-shot, and fine-tuning approaches. *Egyptian Informatics Journal*, *30*, Article 100683. https://doi.org/10.1016/j.eij.2025.100683

Kastrati, M., Imran, A. S., Hashmi, E., Kastrati, Z., Daudpota, S. M., & Biba, M. (2025). Unlocking language barriers: Assessing pre-trained large language models across multilingual tasks and unveiling the black box with explainable artificial intelligence. *Engineering Applications of Artificial Intelligence*, *149*, Article 110136. https://doi.org/10.1016/j.engappai.2025.110136

Khamphakdee, N., & Seresangtakul, P. (2021a). A framework for constructing thai sentiment corpus using the cosine similarity technique [Conference presentation]. *2021 13th International Conference on Knowledge and Smart Technology (KST)*, Chonburi, Thailand. https://doi.org/10.1109/KST51265.2021.9415802

Khamphakdee, N., & Seresangtakul, P. (2021b). Sentiment analysis for Thai language in hotel domain using machine learning algorithms. *Acta Informatica Pragensia*, *10*(2), 155-171. https://doi.org/10.18267/j.aip.155

Khamphakdee, N., & Seresangtakul, P. (2023). An efficient deep learning for Thai sentiment analysis. *Data*, *8*(5), Article 90. https://doi.org/10.3390/data8050090

Khoboko, P. W., Marivate, V., & Sefara, J. (2025). Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models. *Machine Learning with Applications*, *20*, Article 100649. https://doi.org/10.1016/j.mlwa.2025.100649

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach (Version 1). *arXiv*, https://doi.org/10.48550/ARXIV.1907.11692

Lowphansirikul, L., Polpanumas, C., Jantrakulchai, N., & Nutanong, S. (2021). WangchanBERTa: pretraining transformer-based Thai language models (Version 2). *arXiv*, https://doi.org/10.48550/ARXIV.2101.09635

Ozkaya, I. (2023). Application of large language models to software engineering tasks: Opportunities, risks, and implications. *IEEE Software*, *40*(3), 4–8. https://doi.org/10.1109/MS.2023.3248401

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations [Conference presentation]. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, LA, United States. https://doi.org/10.18653/v1/N18-1202

Phatthiyaphaibun, W., Chaovavanich, K., Polpanumas, C., Suriyawongkul, A., Lowphansirikul, L., & Pattarawat Chormai. (2020). *PyThaiNLP/pythainlp: PyThaiNLP 2.1.4 (Version v2.1.4)* [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.3659277

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018*). Improving Language Understanding by Generative Pre-Training*. OpenAI. Retrieved from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Saetia, C., Thonglong, A., Amornchaiteera, T., Chalothorn, T., Taerungruang, S., & Buabthong, P. (2024). Streamlining event extraction with a simplified annotation framework. *Frontiers in Artificial Intelligence*, *7*, Article 1361483. https://doi.org/10.3389/frai.2024.1361483

Sanchan, N. (2024). Comparative study on automated reference summary generation using BERT models and ROUGE score assessment. *Journal of Current Science and Technology*, *14*(2), Article 26. https://doi.org/10.59796/jcst.V14N2.2024.26

Simmachan, T., & Boonkrong, P. (2025). Effect of resampling techniques on machine learning models for classifying road accident severity in Thailand. *Journal of Current Science and Technology*, *15*(2), Article 99. https://doi.org/10.59796/jcst.V15N2.2025.99

ThAIKeras. (2018). *ThAIKeras/bert.* Retrieved from https://github.com/ThAIKeras/bert

Thiengburanathum, P., & Charoenkwan, P. (2023). SETAR: Stacking ensemble learning for Thai sentiment analysis using RoBERTa and hybrid feature representation. *IEEE Access*, *11*, 92822–92837. https://doi.org/10.1109/ACCESS.2023.3308951