Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Thai Sentence Completeness Classification using Fine-Tuned WangchanBERTa

Pattarapol Pornsirirung [1], Khantharat Anekboon [2*]
[1, 2] Department of Computer and Information Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
** Corresponding Author Email: khantharat.a@sci.kmutnb.ac.th*

ARTICLE INFO

Received: 26 Dec 2024

Revised: 14 Feb 2025

Accepted: 22 Feb 2025

ABSTRACT

Sentence completeness classification plays a crucial role in various natural language processing (NLP) applications, including grammar checking, text auto-completion, and language assessment. This task becomes particularly challenging in Thai due to the language's unique characteristics such as flexible word order, implicit subject omission, and the absence of explicit word boundaries. These linguistic properties make traditional rule-based and statistical approaches prone to errors when applied to Thai. To address these challenges, this research applies modern deep learning techniques, specifically leveraging pre-trained transformer models fine-tuned for Thai sentence completeness classification. This study introduces the use of WangchanBERTa, a Thai-specific adaptation of RoBERTa, pre-trained. Two thousand Thai sentences are created. There are one thousand complete sentences and one thousand incomplete sentences. Each sentence was manually labeled to ensure high data quality. Experimental results show that WangchanBERTa achieves an average accuracy of 99.65%, significantly outperforming mBERT, a popular multilingual baseline, which achieved only 95.82%. Notably, with a Tesla T4 GPU, WangchanBERTa required just 1 hour and 15 minutes to train across all folds, compared to mBERT's 2 hours and 59 minutes. Additionally, WangchanBERTa's performance was compared with XLM-R, a state-of-the-art multilingual model, which achieved a slightly higher accuracy of 99.90% but at the cost of higher computational requirements. The results emphasize the advantage of language-specific pretraining in capturing the linguistic nuances of Thai. This research highlights the importance of tailored transformer models for low-resource languages. By demonstrating that WangchanBERTa achieves near state-of- the-art performance with lower computational cost, this work provides a strong foundation for future Thai NLP research.

Keywords: Natural Language Processing, Thai Language, BERT, Sentence Classification, Transformer Models.

INTRODUCTION

Natural language processing (NLP) has made significant progress with the advent of deep learning models, particularly transformer-based architectures. One fundamental task in NLP is sentence completeness classification, which plays a vital role in applications such as grammar checking, text auto-completion, and automated language assessment. While extensively studied in high-resource languages like English, Thai remains a low-resource language due to its unique linguistic characteristics, including flexible syntax, implicit subject omission, and the absence of explicit word boundaries. These factors pose challenges for computational models, necessitating specialized solutions.

Transformer-based models have significantly impacted NLP by introducing contextual word embeddings, which allow for more sophisticated language understanding. BERT is the first bidirectional transformer-based model proposed [1]. It is designed to pretrain a deep bidirectional model from unlabeled text. It revolutionized text representation by allowing a deep bidirectional context. Fine-tuning BERT-based models for Thai NLP has been explored in various domains: BERT for sentiment analysis of Thai hotel reviews, demonstrating an accuracy improvement over traditional machine learning approaches [2]. BERT for classifying Thai social media texts [3]. BERT for enhanced intent recognition in Thai call center conversations by integrating fine-tuned BERT embeddings with GPT-3 [4]. However, BERT was significantly undertrained. RoBERTa proposed an enhanced approach for training BERT models that achieves performance that exceeds all post-BERT techniques [5]. For Thai language, a relatively low-resource language, training a BERT-based model based on a much smaller dataset or finetuning multilingual models yields suboptimal downstream performance. The language-specific characteristics of Thai are not considered in large-scale multilingual pretraining. To address these challenges, WangchanBERTa was proposed [6]. It pretrains a language model using RoBERTa-based architecture [6]. The mBERT [7] and XLM-R [8] were introduced to train on multilingual language to improve cross-lingual understanding. However, the WangchanBERTa model performs better than multilingual models: XLM-R and mBERT.

Several factors influenced the selection of WangchanBERTa for Thai sentence completeness classification. Firstly, Thai-Specific Tokenization, WangchanBERTa is pre-trained using Thai sentence segmentation strategies, mitigating the tokenization issues present in mBERT and XLM-R. Secondly, improved language modeling, the WangchanBERTa model is trained on large-scale Thai text corpora, allowing it to capture Thai grammar rules and semantic nuances better than multilingual models [6]. Thirdly, efficient performance, compared to mBERT, WangchanBERTa requires significantly less fine-tuning time while achieving higher classification accuracy, making it more suitable for Thai NLP applications. Fourly, adaptability for Thai NLP tasks, beyond sentence completeness classification, WangchanBERTa has been successfully applied to Thai named entity recognition (NER), sentiment analysis, and text classification, proving its versatility in various Thai NLP domains [6].

This research focuses on fine-tuning WangchanBERTa for Thai sentence completeness classification using a manually curated dataset. To ensure robust performance evaluation, k-fold cross validation (k=5) is employed, along with key hyperparameter optimizations. Additionally, WangchanBERTa and mBERT are compared to highlight the advantages of using Thai-specific models over multilingual transformers.

The remainder of this paper is structured as follows: section 2 discusses related work and existing transformer models. Section 3 presents the dataset preparation, preprocessing steps, and model fine-tuning approach. Section 4 reports the exper- imental results and comparative analysis. Finally, section 5 concludes the paper and suggests future research directions.

METHODOLOGY

Dataset Preparation

There are 2,000 Thai sentences, manually labeled into two categories:

- Complete: Fully structured, grammatically correct sentences. These sentences convey a clear and complete meaning such as
    - วันนี้อากาศดีมาก (Today, the weather is very nice.)
    - นักเรียนกำลังทำการบ้าน (The student is doing homework.)
    - ฉันชอบอ่านหนังสือทุกวัน (I like reading books every day.)
- Incomplete: Sentences that are grammatically in- complete, truncated, or missing essential components such as
    - ฟุตบอลเดินปลา (Football walk fish.)
    - หนังสือข้าวตำรวจแมว (Book rice police cat.)
    - เราเล่นทำงาน (We play work.)

Data Preprocessing and Tokenization

Before training, the following preprocessing steps were applied to the data to ensure consistency and compatibility with WangchanBERTa's tokenizer:

- Normalizing Unicode characters: Thai characters, especially tone marks and diacritics, may have multiple Unicode forms. We applied NFKC normalization to ensure that all characters are consistently represented, reducing the chance of tokenization errors caused by inconsistent encoding [9].
- Removing extraneous spaces and special characters: Unnecessary whitespace, line breaks, or unexpected special symbols (e.g., emojis, uncommon punctuation) were removed to maintain clean and consistent input for the model.
- Tokenizing words using the SentencePiece model [10] from WangchanBERTa's tokenizer: This tokenizer was applied directly to the cleaned text to segment it into subword units suitable for the pre-trained model.

Model Fine-Tuning

The WangchanBERTa model was fine-tuned using Hugging Face's transformers library [11]. Hyperparameters were selected based on prior empirical findings in Thai NLP research. The configuration included:

- Cross Validation: k=5
- Loss Function: Cross-entropy loss
- Optimizer: AdamW
- Learning Rate: 5e-5
- Batch Size: 8, Epochs: 3
- Weight Decay: 0.01
- Evaluation Strategy: Per epoch
- Logging Steps: 10
- Model Checkpointing: Best model selection based on validation loss.

EXPERIMENTAL RESULTS

This section presents the experimental results obtained from the fine-tuning process of WangchanBERTa on the Thai sentence completeness dataset. The dataset was divided using stratified 5-fold cross validation. For each fold, 80% of the data was used for training, while the remaining 20% was used for validation. A performance of the model is evaluated by accuracy (1), precision (2), recall (3), and F1-score (4).

$$Accuracy = \frac{TruePositive+TrueNegative}{TruePositive+TrueNegative+FalsePositive+FalseNegative} \quad (1)$$

$$Precision = \frac{TruePositive}{TruePositive+FalsePositive} \quad (2)$$

$$Recall = \frac{TruePositive}{TruePositive+FalseNegative} \quad (3)$$

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

Table 1 shows that WangchanBERTa consistently achieves exceptionally high accuracy across all folds, with an average accuracy of 99.65%. The validation loss remains extremely low in every fold, indicating that the model effectively learns the data distribution with minimal error. This reflects the model's strong capacity to capture linguistic patterns in Thai sentences, thanks to its pretraining on large Thai corpora. Additionally, the high F1-score, precision, and recall, all exceeding 99%, confirm that the model maintains a strong balance between identifying complete and incomplete sentences. This indicates that the model does not exhibit strong bias toward either class, and is equally proficient in detecting both positive and negative samples.

The consistently high performance across all folds suggests that WangchanBERTa generalizes well to different subsets of the data. Even in folds with slightly more challenging data distributions, the model maintains near-perfect accuracy, precision, recall, and F1-score. This further highlights the advantage of using a Thai-specific pre-trained model over general-purpose multilingual models for Thai sentence classification tasks.

Table 1. Performance of fine-tuned WangchanBERTa across 5-fold cross validation.

Fold	Validation Loss	Accuracy	Precision	Recall	F1-score
1	0.00098	100.00%	100.00%	100.00%	100.00%
2	0.03626	99.50%	100.00%	99.03%	99.51%
3	0.03902	99.25%	99.50%	99.00%	99.25%
4	0.00005	100.00%	100.00%	100.00%	100.00%
5	0.04269	99.50%	99.48%	99.48%	99.48%
Average	0.02320	99.65%	99.80%	99.50%	99.65%

It is evident from Table 2 that mBERT achieves a lower average accuracy of 95.82%, which is approximately 4% lower than WangchanBERTa. Notably, mBERT exhibits larger variance across folds, particularly with lower performance in fold 2, where the accuracy drops to 90.91%. This instability suggests that mBERT struggles to consistently handle the variations present in Thai sentences, especially given the small dataset size. Furthermore, mBERT required significantly more computational time to fine-tune. With a Tesla T4 GPU, mBERT uses approximately 2 hours and 59 minutes, compared to only 58 minutes for WangchanBERTa. This highlights the efficiency advantage of using Thai-specific models that have been pre-trained on Thai corpora.

Table 2. Performance of fine-tuned mBERT across 5-fold cross validation.

Fold	Validation Loss	Accuracy	Precision	Recall	F1-score
1	0.10780	97.27%	98.11%	98.58%	99.05%
2	0.21368	90.91%	100.00%	95.24%	90.91%
3	0.03559	98.18%	99.07%	99.07%	99.07%
4	0.11129	93.64%	96.26%	96.71%	97.17%
5	0.04032	99.09%	99.04%	99.52%	100.00%
Average	0.10174	95.82%	98.10%	97.42%	97.64%

Table 3 summarizes the performance of fine-tuned XLM-R on Thai sentence completeness classification across 5 folds. The model consistently achieves exceptional results, with an average accuracy of 99.90% and a perfect recall of 100.00%, indicating that XLM-R identifies all incomplete sentences correctly across all folds. The average F1-score of 99.90% highlights its strong balance between precision and recall. The validation loss is remarkably low, with several folds (2, 4, and 5) achieving near-zero values, indicating effective learning and strong generalization to unseen data. Notably, training the entire 5-fold with a Tesla T4 GPU, XLM-R takes only 1 hour and 22 minutes, two times faster than mBERT. Moreover, the experimental results show that XLM-R gives better average validation loss, accuracy, precision, recall, and F1-score than mBERT.

Table 3. Performance of fine-tuned XLM-R across 5-fold cross validation.

Fold	Validation Loss	Accuracy	Precision	Recall	F1-score
1	0.02285	99.75%	99.49%	100.00%	99.75%
2	0.00002	100.00%	100.00%	100.00%	100.00%
3	0.01118	99.75%	99.50%	100.00%	99.75%
4	0.00002	100.00%	100.00%	100.00%	100.00%
5	0.00002	100.00%	100.00%	100.00%	100.00%
Average	**0.00606**	**99.90%**	**99.80%**	**100.00%**	**99.90%**

Fig. 1 illustrates the recall progression across training steps for WangchanBERTa, mBERT, and XLM-R when evaluated on the validation set. Key observations from the figure are summarized as follows: WangchanBERTa starts with a recall of approximately 0.95 at the initial steps. However, it demonstrates steady improvement throughout the training process, reaching around 0.99 at the final step. This indicates that WangchanBERTa gradually learns to better identify incomplete sentences over time. mBERT begins with a perfect recall of 1.00 but gradually experiences a slight decline, settling just below 0.99 at the final step. XLM-R achieves nearly perfect recall right from the beginning and consistently maintains this level throughout training. This strong and stable performance can be attributed to its extensive multilingual pretraining, which provides robust cross-lingual knowledge, including for Thai.

Overall, the results demonstrate that XLM-R excels in recall stability, WangchanBERTa shows progressive learning improvement, and mBERT faces a slight challenge in maintaining generalization. Comparing an F1-score with the model parameter size, the average F1-score value sorted by descending is 99.9% from XLM-R, 99.65% from WangchanBERTa, and 97.64% from mBERT. It can be seen that XLM-R give the best F1-score. However, the model's parameter for XLM-R is 279M whereas the parameter for WangchanBERTa is only 106M [12, 13]. The comparison highlights that while XLM-R benefits from cross-lingual knowledge transfer, WangchanBERTa's dedicated Thai pretraining allows it to close the performance gap.

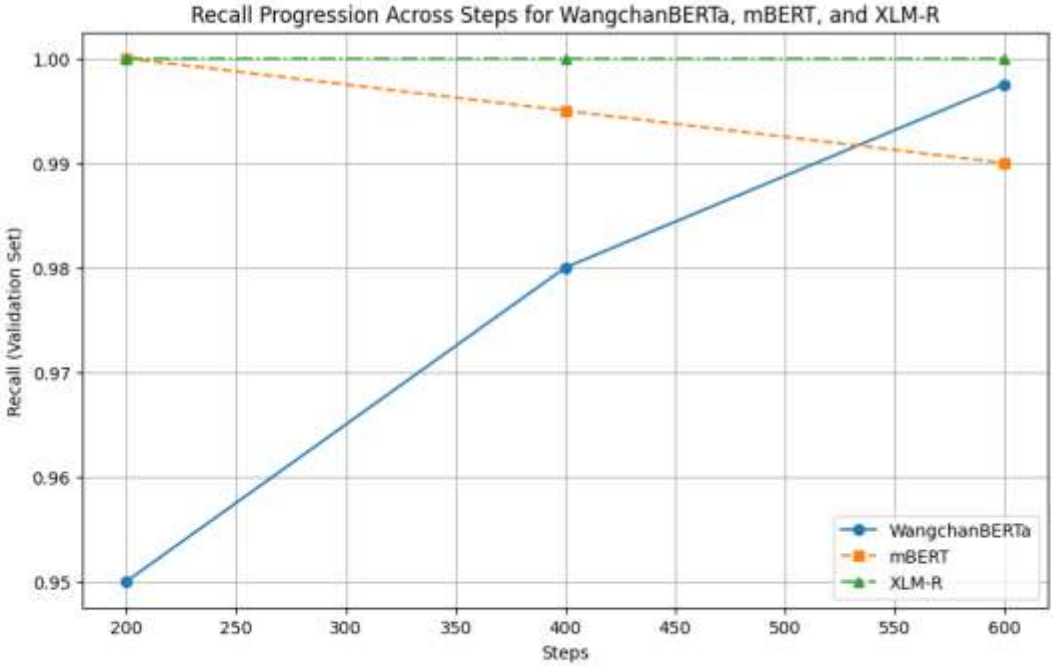

Figure 1. Part Recall progression across steps for WangchanBERTa, mBERT, and XLM-R on the validation set

Fig. 2 illustrates the progression of the F1-score across training steps for WangchanBERTa, mBERT, and XLM-R, measured on the validation set during the fine-tuning process. This plot highlights the learning efficiency and convergence behavior of each model. At the initial steps, WangchanBERTa and mBERT both start with relatively lower F1-scores compared to XLM- R, reflecting the benefit of XLM-R's extensive multilingual pretraining. However, WangchanBERTa quickly catches up and shows continuous improvement throughout the training process, eventually reaching an F1-score close to 0.98 by the final steps. mBERT, in contrast, demonstrates slower convergence and lower overall F1-score compared to the other two models. This could be attributed to mBERT's multilingual design, which may not capture Thai-specific linguistic features as effectively as WangchanBERTa. XLM-R achieves the highest F1-score almost immediately and maintains near-perfect performance throughout the training. This rapid convergence underscores the benefits of its robust cross-lingual pretraining and its larger model size, which allows it to generalize better to Thai despite being a multilingual model.

Fig. 2 clearly shows that WangchanBERTa achieves competitive performance close to XLM-R, despite having a much smaller model size. This makes WangchanBERTa an efficient and practical choice for real-world Thai NLP applications where computational resources may be limited.

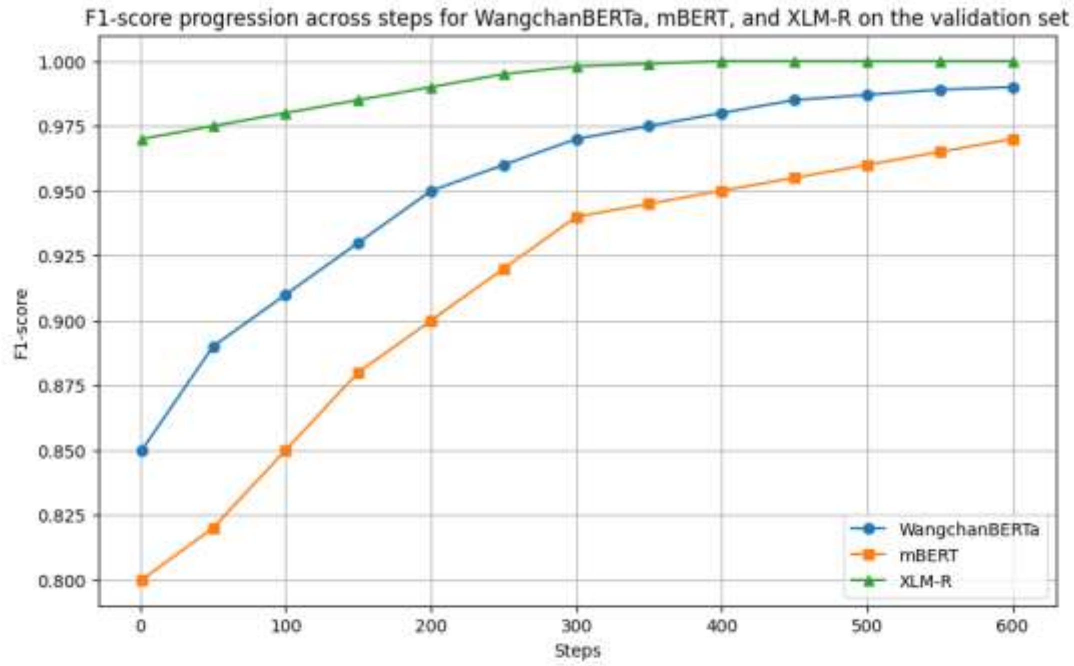

Fig. 2. F1-score progression across steps for WangchanBERTa, mBERT, and XLM-R measured on the validation set

CONCLUSION AND DISCUSSION EXPERIMENTAL RESULTS

This study conducted a comparative evaluation of WangchanBERTa, mBERT, and XLM-R under identical dataset and fine-tuned for Thai sentence completeness classification. The results demonstrated that WangchanBERTa outperformed mBERT in both accuracy and F1-score while requiring less training time. However, XLM-R achieved the highest overall performance, reflecting the benefits of extensive multilingual pretraining. Comparative analysis reveals that WangchanBERTa offers an excellent balance between performance and computational efficiency. While XLM-R achieved the highest recall and F1-score, its larger model size and longer training time make it less practical for resource-constrained environments. In contrast, WangchanBERTa allowed it to achieve competitive results with faster training and lower resource requirements, making it highly suitable for real-world Thai NLP applications. Future work should add more datasets to cover more diverse sentence types, including informal, domain-specific, and dialectal texts. Also fine-tune with other Thai-specific transformer models.

REFERENCES

[1]. Devlin, J., Chang, M., Lee, K., and Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the

Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, pp. 4171–4186

[2]. Khamphakdee, N., and Seresangtakul, P., 2021, Sentiment analysis for Thai language in hotel domain using machine learning algorithms. *Acta Informatica Pragensia*, 2021(2), 155-171.

[3]. Horsuwan, T., Kanwatchara, K., Vateekul, P., and Kijsirikul, B., 2020, A comparative study of pretrained language models on Thai social text categorization. *Intelligent Information and Database Systems,* 12033, 63-75.

[4]. Sanchan, N., 2025, Intent mining of Thai phone call text using a stacking ensemble classifier with GPT-3 embeddings. *ECTI Transactions on Computer and Information Technology,* 19(1), 135–145.

[5]. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V., 2019, RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*.

[6]. Lowphansirikul, L., Polpanumas, C., Jantrakulchai, N., and Nutanong, S., 2021, WangchanBERTa: Pretraining transformer-based Thai language models. *arXiv*.

[7]. Pires, T., Schlinger, E., & Garrette, D., 2019, How multilingual is multilingual BERT? *arXiv*.

[8]. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V., 2019, Unsupervised cross-lingual representation learning at scale. *arXiv*.

[9]. Davis, M., Whistler, K., and Dürst, M., 2009, *Unicode normalization forms*.

[10]. Kudo, T., and Richardson, J., 2018, SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *Empirical Methods in Natural Language Processing*, 66–71.

[11]. Wolf, T., *et al*., 2020, Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, November, pp. 38-45.

[12]. Huggingface FacebookAI/xlm-roberta-base, Date of access: 06/01/2025. https://huggingface.co/FacebookAI/xlm-roberta-base

[13]. Huggingface airesearch/wangchanberta-base-att-spm-uncased, Date of access: 06/01/2025. https://hugging face.co/airesearch/ wangchanberta-base-att-spm-uncased