JIST Journal of Information Science and Technology
Volume 14, NO 2 | JUL – DEC 2024 | 1-11
ISSN: 2651-1053 (Online)
ISSN: 1906-9553 (Print)

Thai Language Sentiment Analysis with a Hybrid Method on WangchanBERTa-CNN-BiLSTM

Kasidhit Suraratchai and Suronapee Phoomvuthisarn*

College of Digital Innovation Technology, Rangsit University

Received: April 30, 2024; Revised: November 5, 2024; Accepted: December 11, 2024; Published: December 28, 2024

ABSTRACT – Understanding emotions conveyed in text, especially in non-global languages such as Thai, sentiment analysis is particularly important in Thailand. However, this endeavor faces challenges due to variations in text length, which significantly impact sentiment analysis outcomes. Previous research has employed neural network and machine learning models in the process, yet each model specializes in different aspects, making comprehensive sentiment analysis coverage unattainable. Recent research has delved into hybrid models like CNN-BiLSTM and BiLSTM-CNN. Although they demonstrate efficacy, their performance still varies across different datasets. For instance, CNN-BiLSTM excels with short sentences by considering surrounding word context, while BiLSTM-CNN is more effective with long sentences due to its bidirectional learning capability. While showing promise, these models perform effectively, but varied text lengths in datasets often lead to sentiment misinterpretation. To address these challenges, inspired by recent advances, we propose an innovative solution: the Parallel Hybrid model. This approach integrates WangchanBERTa into both CNN-BiLSTM and BiLSTM-CNN architectures, harnessing ensemble techniques to improve overall performance and adaptability. Our experiments, conducted on datasets like Wisesight, a highly imbalanced dataset with mostly longer texts, and Thai Children's Tales, a less imbalanced dataset with mostly shorter texts, confirm the effectiveness of the Parallel Hybrid model, which outperforms other model configurations with Macro F1 scores of 0.6270 and 0.7859, respectively. This research marks a significant advancement in sentiment analysis for the Thai language.

KEYWORDS: Sentiment Classification, Deep Learning, Natural Language Processing, Convolutional Neural Network, Bi-LSTM, Thai Language

1. Introduction

Sentiment analysis, a crucial component of Natural Language Processing (NLP), deciphers user sentiments in both text and speech, focusing on discerning positive or negative emotions conveyed. It evaluates sentiment polarity (positive, negative, or neutral), offering businesses valuable insights into customer perceptions. Diverse methodologies, Zhang et al. (2018) (1) leveraged multiple dictionaries for sentiment analysis on Chinese microblog messages, highlighting the importance of diverse linguistic resources. Zhang et al.'s (2020) (2) research on the Bidirectional Long Short Term Memory Network (BiLSTM), Cahyanti et al. (2020) (3) employing the Support Vector Machine (SVM) and Mishra et al.'s (2023) (4) exploration of the Long Short Term Memory Network (LSTM) and Convolutional Neural Network (CNN), signify ongoing progress in sentiment analysis methodologies. These diverse approaches lay a foundation for exploring sentiment analysis in Thai, considering its unique linguistic traits.

However, according to research [10], sentiment analysis poses challenges due to its uniqueness, such as the text length of each dataset. This necessitates adapting each model to the relevant dataset. Despite these obstacles, advances in NLP can be tailored to different types of data. However, ongoing efforts are underway to enhance the effectiveness of sentiment analysis in Thailand. By integrating a neural network architecture customized for the Thai language, coupled with insights from the broader NLP literature, there is potential for significant improvement in accuracy in Thai text processing.

**Corresponding Author: kasidhitdan@gmail.com*

Pasupa et al. (2022) [10] explored a hybrid CNN and BiLSTM approach for sentiment analysis on Thai language datasets. Their study systematically compared model architectures and feature extraction techniques, showcasing the effectiveness of this hybrid model. However, they utilized Thai2fit (ULMFiT) for word embedding, lacking the bidirectional understanding of BERT. Recognizing BERT's limitations in Thai due to its English-centric pre-training, this spurred the development of WangchanBERTa, a specialized BERT model tailored for Thai.

Inspired by Na Chen et al. (2023) [12] research on a parallel neural network model that combines BERT, CNN, and BiLSTM to perform sentiment analysis, new ideas are generated in integrating WangchanBERTa. Here, CNN-BiLSTM and BiLSTM-CNN are mixed in parallel. The goal is to leverage the strengths of CNN-BiLSTM, which can analyze well in short sentences, with BiLSTM-CNN, which can analyze well in long sentences.

In this research, we aim to integrate WangchanBERTa, a specialized transformer model for Thai. This methodology seeks to enhance sentiment analysis accuracy and effectiveness in Thai by integrating with the CNN-BiLSTM and BiLSTM-CNN models. We conduct experiments on two datasets: the Wisesight dataset, comprising diverse Thai language social media texts, and the 40 Thai Children's Tales Dataset, offering a unique linguistic perspective. By leveraging state-of-the-art NLP models and diverse datasets, this research aims to develop advanced sentiment analysis techniques tailored specifically for the Thai language context. Using the macro F1 score as the measuring method, preliminary experimental results indicate that employing WangchanBERTa improves sentiment analysis outcomes. Moreover, the integration of a parallel hybrid approach, combining CNN-BiLSTM and BiLSTM-CNN, further enhances sentiment analysis results.

2. Literatures Review

The literature review explores various hybrid approaches for sentiment analysis, highlighting their effectiveness in enhancing model performance across diverse domains. Amrani et al. (2018) (5) proposed a hybrid framework combining Random Forest and Support Vector Machine techniques to classify Amazon product reviews. Despite achieving high accuracy and showcasing the synergistic benefits of both algorithms, the study highlighted computational costs and limitations in retaining comprehensive sentence contexts crucial for accurate sentiment analysis.

Similarly, Erşahin et al. (2019) (6) introduced a pioneering hybrid methodology for Turkish sentiment analysis, combining lexicon-based and machine learning strategies. Their approach, integrating a sentiment dictionary with supervised classifiers like naive Bayes and support vector machines, outperformed standalone methods. Notably, the hybrid model achieved impressive accuracy across datasets from movies, hotels, and Twitter, surpassing both lexicon-based and machine learning-based approaches. However, the study highlighted challenges in hybrid models, particularly in effectively handling cross-domain data and capturing nuanced sentence contexts crucial for sentiment analysis.

Naseem et al. (2019) (7) introduced a groundbreaking hybrid approach, merging word representations with Bi-directional Long Short Term Memory (BiLSTM), outperforming traditional methods in accuracy. By integrating diverse word embeddings like Character, Context (Elmo), Glove, Lexicon, and Part of Speech embeddings with BiLSTM, the study achieved impressive accuracies across various datasets, including those from US airlines. Despite its success, the study underscored the evolving landscape of language modeling, notably with the rise of BERT (Bidirectional Encoder Representations from Transformers), renowned for its bidirectional nature and robust contextual understanding.

Cai et al. (2020) (8) introduced a hybrid strategy combining BERT and BiLSTM, aiming to leverage BERT's proficiency in adjacent word statistics and BiLSTM contextual learning capabilities. This fusion resulted in notable improvements in accuracy and recall rates, surpassing the individual performance of BERT and BiLSTM. However, the study emphasized the limitations of standalone models, particularly in capturing local patterns within input data, highlighting the importance of complementary approaches like CNN.

Pasupa et al. (2022) (9) employed a hybrid architecture integrating CNN and BiLSTM for sentiment analysis on Thai datasets, highlighting the advantage of synergistically combining multiple models for enhanced performance. Their study evaluated various feature extraction techniques before entering the deep learning model, showcasing the superiority of the amalgamated CNN-BiLSTM model over individual counterparts. However, the study recognized the limitations of existing word embedding techniques, particularly in effectively representing Thai language data, which necessitated the development of specialized models such as WangchanBERTa.

Gupta et al. (2022) (10) proposed a hybrid approach that integrates BERT, BiLSTM-BiGRU, and a 1-D CNN model for binary sentiment classification analysis of movie reviews, achieving an accuracy of 93.89%. Additionally, Na Chen et al. (2023) (11) introduced a parallel hybrid neural network model combining BERT, CNN, and BiLSTM for analyzing hotel review datasets (ChnSentiCorp) in the context of Chinese text sentiment analysis. This hybrid model demonstrated an accuracy of 92.35%.

In Chen et al.'s (2023) (11) study, they utilize parallel hybrid models, building upon Pasupa et al.'s (2022) (9) research. BiLSTM-CNN is employed for longer documents and translations, while CNN-BiLSTM is used for shorter ones, resulting in a parallel hybrid model. This configuration, acting as a set of models, increases overall performance and generality. Ensembling proves advantageous as different architectures may capture varying aspects of the data, ensuring robustness by providing alternative perspectives when one network fits a specific model.

3. Method

In this research, we introduce a novel hybrid model aimed at enhancing sentiment analysis performance. The proposed model, termed the Parallel Hybrid method, integrates WangchanBERTa with both CNN-BiLSTM and BiLSTM-CNN architectures. WangchanBERTa is a Transformer Model that converts text data into vectors based on the context of words. CNN-BiLSTM is good at capturing the context surrounding words rather than the sentence as a whole, and BiLSTM-CNN is good at capturing the context of words as a whole rather than surrounding words by doing Parallel. The Hybrid Model combines the advantages of each model to improve sentiment analysis results. To conduct the experiment, two datasets were utilized: the Wisesight dataset and the 40 Thai Children’s Tales dataset, both containing pre-labeled sentiments in Thai text. The model will be compared with two other models: WangchanBERTa with CNN-BiLSTM and WangchanBERTa with BiLSTM-CNN, as depicted in Figure 1.

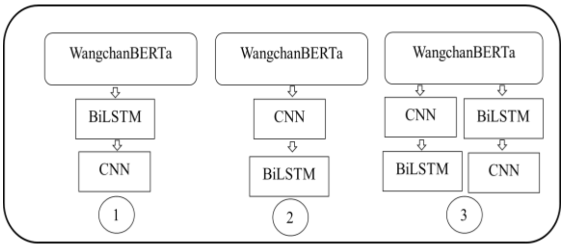

Figure 1. Architecture

The selection of the Wisesight Dataset and the 40 Thai Children’s Tales Dataset addresses the scarcity of labeled Thai language datasets for sentiment analysis, a significant challenge. These datasets minimize errors from manual labeling or unlabeled data and offer diverse social media texts and children's stories, providing unique linguistic perspectives. By leveraging these datasets, comparisons with past research on Thai sentiment analysis can be made, contributing to advancements in the field and facilitating a comprehensive assessment of model performance, enriching Thai sentiment analysis research. There are also differences between the two datasets in overall message length that can be compared across models.

3.1 Data Exploration

The Wisesight dataset was collected from public pages on Facebook, Twitter, YouTube, Pantip.com, and other web forums between 2016 and early 2019. The dataset from PyThaiNLP comprises 26,737 sentences categorized into four classes as follows: 6,823 negative sentences, 4,778 positive sentences, 14,561 neutral sentences, and 575 questions. It is notable that the distribution among these four classes is imbalanced

(PyThaiNLP, 2019). The Wisesight Sentiment Dataset is available at: https://github.com/PyThaiNLP/wisesight-sentiment.

The following table (Table 1) presents a comprehensive statistical analysis of textual characteristics, including average number of characters, average number of words, maximum and minimum word lengths, maximum and minimum character lengths, and quartile 99 values for both word and character lengths, offering valuable insights into the composition of the Wisesight dataset.

Table 1 Key statistics about the Wisesight Dataset, after pre-processing

Average number of characters	85
Longest characters	2111
shortest characters	1
Number of characters, quantile 0.99	689
Average number of words	25
Longest words	587
shortest words	1
Number of words, quantile 0.99	190

The 40 Thai Children’s Tales dataset includes 1,115 sentences from 40 Thai tales, which were classified by three expert annotators into positive, neutral, or negative sentiment categories. The annotators reached consensus on all 1,115 sentences, resulting in 309 sentences expressing positive sentiment, 508 exhibiting neutral sentiment, and 298 conveying negative sentiment. It is notable that the distribution among these three classes is imbalanced. The dataset is available for download at: https://github.com/dsmlr/40-Thai-Children-Stories.

The following table (Table 2) presents a comprehensive statistical analysis of textual characteristics, including average number of characters, average number of words, maximum and minimum word lengths, maximum and minimum character lengths, and quartile 99 values for both word and character lengths, offering valuable insights into the composition of the 40 Thai Children’s Tales dataset.

Table 2 Key statistics about the 40 Thai Children's Tales, after pre-processing

Average number of characters	**69**
Longest characters	347
shortest characters	10
Number of characters, quantile 0.99	194
Average number of words	18
Longest words	73
shortest words	1
Number of words, quantile 0.99	48

3.2 Data preprocess

Data preprocessing involved removing punctuation, URLs, HTML tags, tabs, whitespace, numerical characters, hashtags, user mentions, and special characters such as "#" and "@", as well as converting emojis to their corresponding text representations. Duplicate rows and empty entries were also eliminated.

Following preprocessing, the Wisesight Dataset comprises 26,676 labeled messages, with imbalanced class distributions: 6,811 negative, 14,513 neutral, 4,777 positive, and 575 questions.

Similarly, the 40 Thai Children's Tales Dataset includes 1,115 labeled messages with imbalanced class distributions: 309 expressing positive sentiment, 508 exhibiting neutral sentiment, and 298 conveying negative sentiment.

For the purpose of creating robust sentiment analysis models, the datasets were meticulously partitioned into distinct subsets. The Wisesight dataset, comprising 26,676 messages, was split into three subsets using the train_test_split method: 60% designated for training, 20% for validation, and another 20% for testing, resulting in 16,005, 5,336, and 5,335 messages, respectively. Similarly, the 40 Thai Children's Tales Dataset contains 1,115 messages, with 669, 223, and 223 messages allocated to the training, validation, and testing sets, respectively.

The train_test_split method was chosen because it is simple and efficient, making it ideal for research projects with time and resource constraints. Unlike cross-validation, which involves multiple splits and repeated training cycles, train_test_split partitions the data in a single step, significantly reducing computation time. This efficiency is particularly beneficial when working with complex models that require extensive training, as it minimizes resource usage while ensuring the datasets are effectively partitioned.

3.3 Model Creation

BiLSTM-CNN Model

The model depicted in Figure 2 undergoes WangchanBERTa encoding upon receiving a sentence input. It then passes through a BiLSTM layer to capture sequential information bidirectionally. The output from BiLSTM, incorporating long-range dependencies, is fed into a CNN to extract local text features.

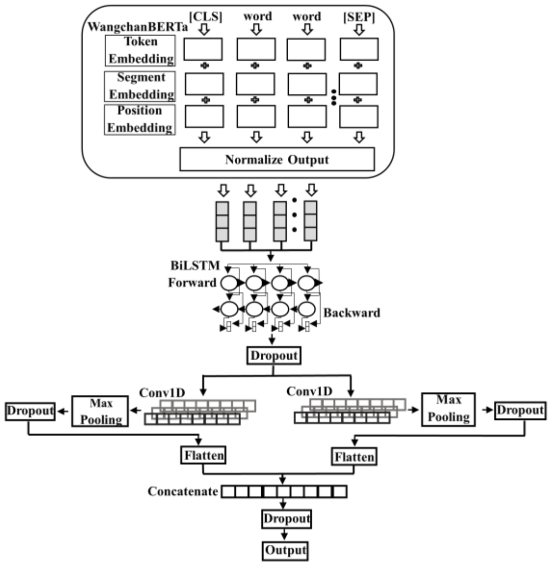

Figure 2. BiLSTM-CNN Model

CNN-BiLSTM

The model depicted in Figure 3 starts with WangchanBERTa encoding for a sentence input. It then proceeds to a CNN to extract local text features. Subsequently, the output from the CNN is passed to a BiLSTM layer to capture sequential information bidirectionally.

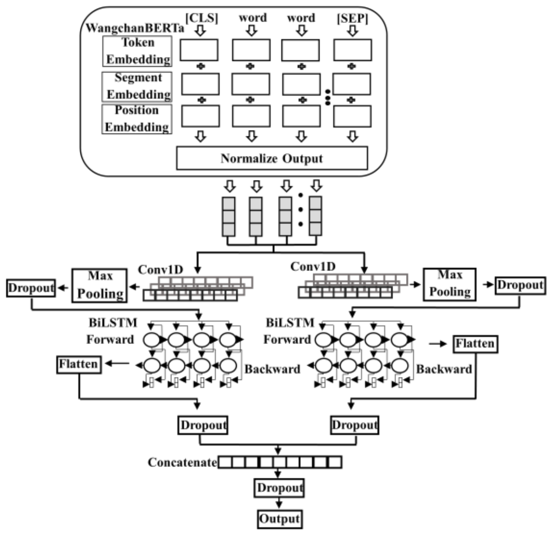

Figure 3. CNN-BiLSTM model

Parallel Hybrid Model

The model in Figure 4 begins with WangchanBERTa encoding for sentences. These encoded sentences are then passed to either BiLSTM-CNN and CNN-BiLSTM architectures, leveraging the strengths of both models to improve prediction outcomes.

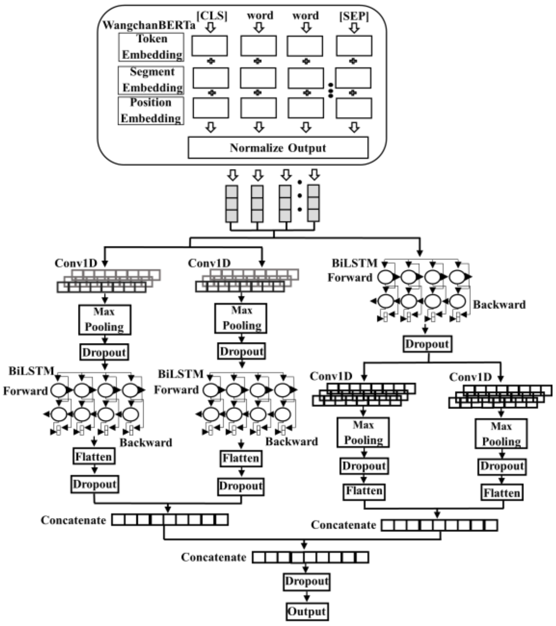

Figure 4. Parallel Hybrid Model

All three models consist of the following parts:

a. Word Embedding Layer: Utilizes pre-trained WangchanBERTa to transform tokenized text into embeddings, comprising Token, Segment, and Positional Embeddings for context-aware representations.
b. Feature Extraction Layer: Utilizes CNN for local semantic features and BiLSTM for global contextual feature extraction, constructing both CNN-BiLSTM and BiLSTM-CNN architectures.
c. Max Pooling Layer: Reduces spatial dimensions of convolutional layer outputs while preserving important information.
d. Dropout Layer: Prevents overfitting by randomly dropping units during training to improve generalization.
e. Concatenate Layers: Combines outputs from CNN-BiLSTM or BiLSTM-CNN for further processing.
f. Output Layer: Fuses feature vectors, applies dropout regularization, and classifies output using sigmoid activation function with a fully connected neural network.

BiLSTM ×CNN (Pasupa Model)

The BLSTM×CNN model is an ensemble approach that combines the predictions from both the Bidirectional Long Short-Term Memory (BLSTM) and Convolutional Neural Network (CNN) models using a soft voting scheme. In this configuration, the sentiment probabilities generated by each model are averaged to arrive at a final prediction. This method harnesses the strengths of both architectures, allowing for improved accuracy in sentiment classification. By blending the outputs, the model can effectively capture diverse features from the data, making it more robust against misclassifications.

BiLSTM +CNN (Pasupa Model)

The BLSTM+CNN model simultaneously processes input sequences through both the BLSTM and CNN layers to leverage their complementary strengths. In this structure, input sentences are first transformed into feature vectors and then passed through both layers concurrently. The BLSTM layer captures long-range dependencies in the text, while the CNN layer extracts local features. The outputs from these layers are concatenated, creating a rich feature set that embodies both contextual information and local patterns. This combined representation is then subjected to a dropout layer before reaching the output layer, enhancing the model's capability to classify sentiment accurately.

3.4 Parameter Setting

The parameters for the models BiLSTM-CNN, CNN-BiLSTM, and Parallel Hybrid, as determined by research utilizing the Wisesight dataset, are as follows.

Table 3 BiLSTM-CNN, CNN-BiLSTM, AND Parallel Hybrid Wisesight Dataset Parameter setting.

BiLSTM-CNN, CNN-BiLSTM, and Parallel Hybrid Wisesight dataset	
Number of hidden units in Bidirectional LSTM layer	128
Bidirectional LSTM layer activation function	ReLU
Number hidden units in Convolution kernels	32
Convolution layer kernel size	2,3
Convolution layer activation function	ReLU
MaxPooling1D size	2
Dropout	0.5
Loss	categorical_crossentropy
Optimizer	Adam
Learning rate	0.001
Epoch	30

The hyperparameters for the models BiLSTM-CNN, CNN-BiLSTM, and Parallel Hybrid, as determined by research utilizing the 40 Thai Children’s Tales Dataset, are as follows.

Table 4 BiLSTM-CNN the 40 Thai Children's Tales Dataset Parameter setting.

BiLSTM-CNN 40 Thai Children's Tales Dataset	
Number of hidden units in Bidirectional LSTM layer	128
Bidirectional LSTM layer activation function	ReLU
Number hidden units in Convolution kernels	128
Convolution layer kernel size	3,5
Convolution layer activation function	ReLU
MaxPooling1D size	2
Dropout	0.6
Loss	categorical_crossentropy
Optimizer	Adam
Learning rate	0.00001
Epoch	100

Table 5 CNN-BiLSTM, AND Parallel Hybrid the 40 Thai Children's Tales Dataset Parameter setting.

CNN-BiLSTM, and Parallel Hybrid 40 Thai Children's Tales Dataset	
Number of hidden units in Bidirectional LSTM layer	64
Bidirectional LSTM layer activation function	ReLU
Number hidden units in Convolution kernels	128
Convolution layer kernel size	3,5
Convolution layer activation function	ReLU
MaxPooling1D size	2
Dropout	0.6
Loss	categorical_crossentropy
Optimizer	Adam
Learning rate	0.00001
Epoch	100

3.5 Model Evaluation

Evaluating the models with test data and labels allows for the assessment of their performance using the F1-Score metric. The comparison includes Pasupa et al.'s study (2022) [14], which encompasses BLSTM, CNN, BLSTM-CNN, CNN-BLSTM, BLSTM+CNN, and BLSTM×CNN models. Each model undergoes feature extraction before entering the deep learning model, incorporating word embedding, POS tag, sentic, word embedding + POS tag, word embedding + sentic, POS tag + sentic, and word embedding + POS tag + sentic. Subsequently, the F1 scores of the models WangchanBERTa-BiLSTM-CNN, WangchanBERTa-CNN-BiLSTM, and the Parallel Hybrid model are compared. The evaluation process is repeated with 10 different random splits in each model, and the F1 macro averages are compared collectively.

4. Experimental Results

4.1 Model Performance

Wisesight Dataset BiLSTM-CNN model

Table 6. Wisesight Dataset BiLSTM-CNN Model Performance

	precision	recall	f1-score	support
neg	0.7452	0.7681	0.7565	13630
neu	0.7548	0.8008	0.7771	29030
pos	0.6084	0.4850	0.5398	9550
q	0.4723	0.3704	0.4152	1150
accuracy			0.7266	53360
macro avg	0.6452	0.6061	0.6221	53360
weighted avg	0.7201	0.7266	0.7216	53360

Table 6 presents the evaluation results for both overall sentiment analysis and sentiment analysis for each class on the test dataset. The results indicate a precision of 0.7201, recall of 0.7402, F1-score of 0.7216, macro precision of 0.6452, macro recall of 0.6061, and macro F1-score of 0.6221.

Table 7. Wisesight Dataset BiLSTM-CNN Model Confusion Matrix

Confusion Matrix

Actual Values / Predicted Values

	Neg	Neu	Pos	Q	Recall
Neg	10469 19.62 %	2721 5.10 %	381 0.71%	59 0.11 %	76.81 %
Neu	2817 5.28%	23247 43.57 %	2587 4.85%	379 0.71 %	80.08 %
Pos	677 1.27%	4203 7.88 %	4632 8.68 %	38 0.07 %	48.50%
Q	85 0.16%	626 1.17 %	13 0.02 %	426 0.80 %	37.04 %
Precision	74.52%	75.48 %	60.84 %	47.23%	Acc. : 72.66 %

As shown in Table 7, the BiLSTM-CNN model on the Wisesight Dataset predominantly misclassified samples as neutral. It accurately identified 48.50% of positive samples but misclassified 44.01% as neutral. For question samples, the model achieved an accuracy of 37.04%, misclassifying 54.43% as neutral. The model performed better with negative samples, correctly classifying 76.81% while misclassifying 19.96% as neutral. Additionally, it accurately classified 80.08% of neutral samples, but misclassified 9.70% as negative.

Wisesight Dataset CNN-BiLSTM model

Table 8. Wisesight Dataset CNN-BiLSTM Model Performance

	precision	recall	f1-score	support
neg	0.7674	0.7621	0.7648	13630
neu	0.7435	0.8366	0.7873	29030
pos	0.6501	0.4373	0.5228	9550
q	0.5007	0.3191	0.3898	1150
accuracy			0.7350	53360
macro avg	0.6654	0.5888	0.6162	53360
weighted avg	0.7276	0.7350	0.7257	53360

Table 8 presents the evaluation results for both overall sentiment analysis and sentiment analysis for each class on the test dataset. The results indicate a precision of 0.7276, recall of 0.7350, and F1-score of 0.7257. The macro precision, recall, and F1-score are 0.6654, 0.5888, and 0.6162, respectively.

Table 9. Wisesight Dataset CNN-BiLSTM Model Confusion Matrix

Confusion Matrix

Actual Values \ Predicted Values	Neg	Neu	Pos	Q	Recall
Neg	10388 19.47 %	2930 5.49 %	261 0.50%	45 0.08 %	76.21 %
Neu	2484 4.66%	24287 45.52 %	1974 3.70%	285 0.53 %	83.66 %
Pos	591 1.11%	4747 8.90 %	4176 8.68 %	36 0.07 %	43.73 %
Q	74 0.14%	702 1.32 %	7 0.01 %	367 0.69 %	31.91 %
Precision	76.74%	74.35 %	65.01 %	50.07%	Acc. : 73.50 %

Predicted Values

As shown in Table 9, the CNN-BiLSTM model on the Wisesight Dataset primarily misclassified samples as neutral. It accurately identified 43.73% of positive samples but misclassified 49.71% as neutral. For question samples, the model achieved an accuracy of 31.91%, misclassifying 61.04% as neutral. It correctly classified 76.21% of negative samples, but misclassified 21.50% as neutral. The model also performed well with neutral samples, accurately classifying 83.66% while misclassifying 10.09% as negative.

Wisesight Dataset Parallel Hybrid model

Table 10. Wisesight Dataset Parallel Hybrid Model Performance

	precision	recall	f1-score	support
Neg	0.7662	0.7685	0.7674	13630
Neu	0.7453	0.8349	0.7876	29030
Pos	0.6553	0.4356	0.5233	9550
Q	0.5146	0.3687	0.4296	1150
accuracy			0.7364	53360
macro avg	0.6704	0.6019	0.6270	53360
weighted avg	0.7296	0.7364	0.7274	53360

Table 10 presents the evaluation results for both overall sentiment analysis and sentiment analysis for each class on the test dataset. The results indicate a precision of 0.7296, recall of 0.7364, and F1-score of 0.7274. The macro precision, recall, and F1-score are 0.6704, 0.6019, and 0.6270, respectively.

Table 11. Wisesight Dataset Parallel Hybrid Model Confusion Matrix

Confusion Matrix

Actual Values \ Predicted Values	Neg	Neu	Pos	Q	Recall
Neg	10475 19.63 %	2881 5.40 %	236 0.44%	38 0.07 %	76.85 %
Neu	2512 4.71%	24236 45.42 %	1949 3.65%	333 0.62 %	83.49 %
Pos	615 1.15%	4746 8.89 %	4160 7.80 %	29 0.05 %	43.56 %
Q	69 0.13%	654 1.23 %	3 0.01 %	424 0.79 %	36.87 %
Precision	76.62%	74.53 %	65.53 %	51.46%	Acc. : 73.64 %

Predicted Values

As shown in Table 11, the Parallel Hybrid model on the Wisesight Dataset primarily misclassified samples as neutral. It accurately classified 43.56% of positive samples but misclassified 49.70% as neutral. For question samples, the model achieved an accuracy of 36.87%, misclassifying 56.87% as neutral. It correctly classified 76.85% of negative samples but misclassified 18.43% as neutral. The model also performed well with neutral samples, accurately classifying 83.48% while misclassifying 8.65% as negative.

The 40 Thai Children's Tales Dataset BiLSTM-CNN model

Table 12. the 40 Thai Children's Tales Dataset BiLSTM-CNN Model Performance

	precision	recall	f1-score	support
neg	0.7123	0.6992	0.7057	595
neu	0.7832	0.8079	0.7953	1015
pos	0.7195	0.6952	0.7071	620
accuracy			0.7475	2230
macro avg	0.7384	0.7341	0.7361	2230
weighted avg	0.7466	0.7475	0.7469	2230

Table 12 presents the evaluation results for both overall sentiment analysis and sentiment analysis for each class on the test dataset. The results indicate a precision of 0.7466, recall of 0.7475, and F1-score of 0.7469. The macro precision, recall, and F1-score are 0.7384, 0.7341, and 0.7361, respectively.

Table 13. the 40 Thai Children's Tales Dataset BiLSTM-CNN Model Confusion Matrix

Confusion Matrix

Actual Values

	Neg	Neu	Pos	Recall
Neg	416 18.65 %	99 4.44 %	80 3.59%	69.92 %
Neu	107 4.80%	820 36.77 %	88 3.95%	80.79 %
Pos	61 2.74%	128 5.74 %	431 19.33 %	69.52 %
Precision	71.23%	78.32 %	71.95 %	Acc. : 74.75 %

Predicted Values

As shown in Table 13, the BiLSTM-CNN model on the 40 Thai Children's Tales Dataset primarily misclassified samples as neutral. It accurately classified 69.92% of positive samples but misclassified 20.65% as neutral. For negative samples, the model achieved an accuracy of 69.52%, misclassifying 16.64% as neutral. Regarding neutral samples, it performed well with an accuracy of 80.79% while misclassifying 10.54% as negative.

The 40 Thai Children's Tales Dataset CNN-BiLSTM model

Table 14. the 40 Thai Children's Tales Dataset CNN-BiLSTM Model Performance

	precision	recall	f1-score	support
neg	0.7581	0.7479	0.7530	595
neu	0.7965	0.8522	0.8234	1015
pos	0.7702	0.6919	0.7290	620
accuracy			0.7798	2230
macro avg	0.7749	0.7640	0.7685	2230
weighted avg	0.7789	0.7798	0.7784	2230

Table 14 presents the evaluation results for both overall sentiment analysis and sentiment analysis for each class on the test dataset. The results indicate a precision of 0.7789, recall of 0.7798, and F1-score of 0.7467. The macro precision, recall, and F1-score are 0.7749, 0.7640, and 0.7685, respectively.

Table 15. the 40 Thai Children's Tales Dataset CNN-BiLSTM Confusion Matrix

Confusion Matrix

Actual Values

	Neg	Neu	Pos	Recall
Neg	445 19.96 %	95 4.26 %	55 2.47%	74.79 %
Neu	77 3.45%	865 38.79 %	73 3.27%	85.22 %
Pos	65 2.91%	126 5.65 %	429 19.24 %	69.19 %
Precision	75.81%	79.65 %	77.02 %	Acc. : 77.98 %

Predicted Values

As shown in Table 15, the CNN-BiLSTM model on the 40 Thai Children's Tales Dataset primarily misclassified samples as neutral. It accurately classified 69.19% of positive samples but misclassified 20.32% as neutral. For negative samples, the model achieved an accuracy of 74.79%, misclassifying 15.97% as neutral. Regarding neutral samples, it performed well with an accuracy of 85.22% while misclassifying 7.59% as negative.

The 40 Thai Children's Tales Dataset Parallel Hybrid model

Table 16. the 40 Thai Children's Tales Dataset Parallel Hybrid Model Performance

	precision	recall	f1-score	support
neg	0.7965	0.7563	0.7759	595
neu	0.8065	0.8502	0.8278	1015
pos	0.7697	0.7387	0.7539	620
accuracy			0.7942	2230
macro avg	0.7909	0.7818	0.7859	2230
weighted avg	0.7936	0.7942	0.7934	2230

Table 16 presents the evaluation results for both overall sentiment analysis and sentiment analysis for each class on the test dataset. The results indicate a precision of 0.7936, recall of 0.7942, and F1-score of 0.7934. The macro precision, recall, and F1-score are 0.7909, 0.7818, and 0.7859, respectively.

Table 17. the 40 Thai Children's Tales Dataset Parallel Hybrid Model Confusion Matrix

Confusion Matrix

Actual Values \ Predicted Values	Neg	Neu	Pos	Recall
Neg	450 20.18 %	92 4.13 %	53 2.38%	75.63 %
Neu	68 3.05%	863 38.70 %	84 3.77%	85.02 %
Pos	47 2.11%	115 5.16 %	458 20.54 %	73.87 %
Precision	79.65%	80.65 %	76.97 %	Acc. : 79.42 %

As shown in Table 17, the Parallel Hybrid model on the 40 Thai Children's Tales Dataset primarily misclassified samples as neutral. It accurately classified 73.87% of positive samples but misclassified 18.55% as neutral. For negative samples, the model achieved an accuracy of 75.63%, misclassifying 15.46% as neutral. Regarding neutral samples, it performed well with an accuracy of 85.02% while misclassifying 8.28% as positive.

4.2 Comparing Models

The performance of the neural network models is presented in Table 18, with performance metrics measured as the macro F1-score on the test dataset. Evaluation was conducted by averaging the macro F1-score across ten random splits.

Table 18. Model Performance

Dataset	Model	Feature	Macro F1-score
Wisesight	BiLSTM	FW + FS	0.5483
	CNN	FW + FP + FS	0.5074
	BiLSTM-CNN	FW + FP + FS	0.5521
	CNN-BiLSTM	FW + FP	0.5609
	BiLSTM+CNN	FW + FP	0.5517
	BiLSTM×CNN	FW	0.5461
	BiLSTM-CNN	WangchanBERTa	0.6221
	CNN-BiLSTM	WangchanBERTa	0.6162
	Parallel Hybrid	WangchanBERTa	0.6270
The 40 Thai Children's Tales	BiLSTM	FW + FP + FS	0.6980
	CNN	FW + FS	0.7393
	BiLSTM-CNN	FW + FP + FS	0.7436
	CNN-BiLSTM	FW + FP + FS	0.6768
	BiLSTM+CNN	FW + FP + FS	0.7124
	BiLSTM×CNN	FW + FP + FS	0.7357
	BiLSTM-CNN	WangchanBERTa	0.7361
	CNN-BiLSTM	WangchanBERTa	0.7685
	Parallel Hybrid	WangchanBERTa	0.7859

Pasupa et al.'s study (2022) [9] groundbreaking study not only advances sentiment analysis with their innovative exploration of BLSTM, CNN, and their hybrids but also enriches the field by meticulously integrating various features such as word embedding, POS tags, and sentic features, offering invaluable insights into the nuances of deep learning architectures for natural language processing.

Result from [9], which encompasses BLSTM, CNN, BLSTM-CNN, CNN-BLSTM, BLSTM+CNN, and BLSTM×CNN models. Each model undergoes feature extraction before entering the deep learning model, incorporating word embedding (FW), POS tag (FP), sentic (FS), word embedding + POS tag, word embedding + sentic, POS tag + sentic, and word embedding + POS tag + sentic.

From Table 18, the standout model in the Wisesight dataset is the Parallel Hybrid model, boasting a commendable macro F1-score of 0.6270, surpassing both the CNN-BiLSTM and BiLSTM-CNN models in scores 0.6162 and 0.6221 each, respectively. Additionally, the Parallel Hybrid model outperformed Pasupa et al.'s (2022) leading model, the CNN-BiLSTM with word embedding + POS tags, which achieved a macro F1-score of 0.5609.

In Table 6, the best performing model on the 40 Thai Children's Tales dataset is again the Parallel Hybrid model, achieving an impressive macro F1 score of 0.7859, surpassing both the CNN-BiLSTM and BiLSTM-CNN models in scores. 0.7685 and 0.7361 each, respectively. Furthermore, it surpassed Pasupa et al.'s (2022) leading model, the CNN-BiLSTM with word embedding + POS + Sentic tags, which achieved a macro F1-score of 0.7436.

4.3 Example Sentence

The performance of the parallel hybrid model facilitates superior sentiment analysis compared to other models. For instance:

"เคยใช้หมด ไปบักหลายขวดคือกัน สรุปกูแพ้ เลยมาใช้ กานิเย้ 😑"
The BiLSTM-CNN model interprets feelings of guilt erroneously by shifting from negative to positive sentiment, in contrast, neither the CNN-BiLSTM model nor the Parallel model did not encounter this problem.

"หมาป่าหันกลับไปมองภูเขาอีกครั้ง ป่าเป็นที่ที่ไม่น่าอยู่ก็จริง มันกล่าว แต่ที่นั่นไม่มีโซ่ตรวน ไม่มีปลอกคอ ไม่มีใครห้ามให้เดินเล่นเพ่นพ่าน" The CNN-BiLSTM model interprets feelings of guilt erroneously by shifting from neutral to negative sentiment. However, both the BiLSTM-CNN and Parallel models did not encounter this issue.

From the given example sentence, it becomes apparent that the BiLSTM-CNN model considers the overall context, such as the phrase "เคยใช้หมด ไปบักหลายขวดคือกัน", but lacks attention to the surrounding context in the "สรุปกูแพ้" part, leading to erroneous analysis. Conversely, the CNN-BiLSTM model focuses on the surrounding context, like the segment "ป่าเป็นที่ที่ไม่น่าอยู่", but overlooks the broader context. This results in inaccurate analysis. The Parallel hybrid model could be developed to mitigate errors in this aspect.

5. Discussion and Conclusion

This study emphasizes the necessity of utilizing a Parallel Hybrid model, which integrates the strengths of the CNN-BiLSTM and BiLSTM-CNN architectures to enhance sentiment analysis performance. The BiLSTM-CNN architecture effectively captures sentence context, making it suitable for longer texts found in the Wisesight dataset. In contrast, the CNN-BiLSTM architecture excels at identifying local word features, which benefits shorter texts like those in the 40 Thai Children’s Tales dataset. By combining these two architectures, the Parallel Hybrid model aims to provide a comprehensive understanding of sentiment by effectively leveraging both word and sentence contexts.

The comparative performance metrics of the models evaluated on both datasets are presented in Table 19

Table 19. Comparative Macro F1 Scores of Sentiment Analysis Models

Model	Wisesight Dataset Macro F1 Score	40 Thai Children's Tales Dataset Macro F1 Score
BiLSTM-CNN	0.6221	0.7361
CNN-BiLSTM	0.6162	0.7685
Parallel Hybrid	0.6270	0.7859
Pasupa et al. (2022)	0.5609	0.7436

The results indicate that the Parallel Hybrid model consistently outperforms the individual models across both datasets, achieving Macro F1 scores of 0.6270 for the Wisesight dataset and 0.7859 for the 40 Thai Children's Tales dataset. This performance underscores the effectiveness and generalizability of the Parallel Hybrid model in addressing the complexities of Thai sentiment analysis. By integrating parallel hybrid techniques, it captures nuanced sentiment patterns more effectively, contributing to its superior performance.

Additionally, the choice of word embedding techniques is crucial for the models' effectiveness. WangchanBERTa embeddings, specifically designed to address the intricacies of the Thai language, significantly enhance sentiment analysis capabilities compared to alternatives such as Thai2Fit. The strategic combination of BiLSTM-CNN and CNN-BiLSTM architectures in the Parallel Hybrid model ensures robust performance across diverse datasets, indicating that while individual architectures may excel in specific contexts, the hybrid approach effectively maximizes their strengths.

In conclusion, the experimental findings highlight the versatility and effectiveness of the Parallel Hybrid model in Thai sentiment analysis. With higher Macro F1 scores and a proficient approach to Thai-specific challenges, this model opens promising avenues for future sentiment analysis research and significantly advances natural language processing methodologies tailored to the complexities of the Thai language domain.

References

[1] S. Zhang, Z. Wei, Y. Wang, and T. Liao, "Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary," Future Generation Computer Systems, vol. 81, pp. 395-403, 2018/04/01/ 2018, doi: https://doi.org/10.1016/j.future.2017.09.048.

[2] Y. Zhang and Z. Rao, "n-BiLSTM: BiLSTM with n-gram Features for Text Classification," in 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), 12-14 June 2020

2020, pp. 1056-1059, doi: 10.1109/ITOEC49072.2020.9141692. [Online]. Available: https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=9141692&ref=

[3] F. E. Cahyanti, Adiwijaya, and S. A. Faraby, "On The Feature Extraction For Sentiment Analysis of Movie Reviews Based on SVM," in 2020 8th International Conference on Information and Communication Technology (ICoICT), 24-26 June 2020 2020, pp. 1-5, doi: 10.1109/ICoICT49345.2020.9166397. [Online]. Available: https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=9166397&ref=

[4] M. Mishra and A. Patil, "Sentiment Prediction of IMDb Movie Reviews Using CNN-LSTM Approach," in 2023 International Conference on Control, Communication and Computing (ICCC), 19-21 May 2023 2023, pp. 1-6, doi: 10.1109/ICCC57789.2023.10165155. [Online]. Available: https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=10165155&ref=

[5] Y. Al Amrani, M. Lazaar, and K. E. El Kadiri, "Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis," Procedia Computer Science, vol. 127, pp. 511-520, 2018/01/01/ 2018, doi: https://doi.org/10.1016/j.procs.2018.01.150.

[6] B. ErŞAhİN, Ö. Aktaş, D. Kilinç, and M. ErŞAhİN, "A hybrid sentiment analysis method for Turkish," TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES, vol. 27, pp. 1780-1793, 05/15 2019, doi: 10.3906/elk-1808-189.

[7] U. Naseem, S. K. Khan, I. Razzak, and I. A. Hameed, "Hybrid Words Representation for Airlines Sentiment Analysis," in AI 2019: Advances in Artificial Intelligence, Cham, J. Liu and J. Bailey, Eds., 2019// 2019: Springer International Publishing, pp. 381-392.

[8] R. Cai et al., "Sentiment Analysis About Investors and Consumers in Energy Market Based on BERT-BiLSTM," IEEE Access, vol. 8, pp. 171408-171415, 2020, doi: 10.1109/ACCESS.2020.3024750.

[9] K. Pasupa and T. Seneewong Na Ayutthaya, "Hybrid Deep Learning Models for Thai Sentiment Analysis," Cognitive Computation, vol. 14, no. 1, pp. 167-193, 2022/01/01 2022, doi: 10.1007/s12559-020-09770-0.

[10] B. Gupta, P. Prakasam, and T. Velmurugan, "Integrated BERT embeddings, BiLSTM-BiGRU and 1-D CNN model for binary sentiment classification analysis of movie reviews," Multimedia Tools and Applications, vol. 81, no. 23, pp. 33067-33086, 2022/09/01 2022, doi: 10.1007/s11042-022-13155-w.

[11] N. Chen, Y. Sun, and Y. Yan, "Sentiment analysis and research based on two-channel parallel hybrid neural network model with attention mechanism," IET Control Theory & Applications, vol. 17, no. 17, pp. 2259-2267, 2023/11/01 2023, doi: https://doi.org/10.1049/cth2.12463