Data Science and Engineering (DSE) Record, Volume 6, issue 1.

COMPARATIVE STUDY OF LLM MODELS FOR SENTIMENT CLASSIFICATION IN THAI FINANCIAL NEWS HEADLINES

Nuttawut Thuayhanruksa [1] and Pree Thiengburanathun [2]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
nuttawut_thuayhan@cmu.ac.th

Abstract. This paper explores the application of various natural language processing (NLP) models for sentiment analysis on financial news articles sourced from Thai financial news websites, focusing on Thai-language data. The study evaluates machine learning and deep learning models, including Logistic Regression, Bidirectional Long Short-Term Memory (Bi-LSTM), Convolutional Neural Networks (CNN), WangChanBERTa, OpenAI's GPT-3.5 and OpenThaiGPT. The models' performance is assessed using accuracy, precision, recall, and F1-score. The findings reveal that the Fine-tuned WangChanBERTa model achieved the highest accuracy of 0.84 on the testing set, demonstrating its superior ability in classifying sentiment in Thai financial news. BI-LSTM and CNN models also performed well, with testing accuracies of 0.781 and 0.791 In contrast, OpenAI's GPT-3.5 and OpenThaiGPT, which lacked fine-tuning and optimized prompts due to computational constraints, exhibited practical limitations in resource-constrained settings.

Keywords: Sentiment Analysis, Thai Natural Language Processing, Thailand Stock Market, Classification

1 Introduction

The stock market's dynamic nature is driven by various factors, including economic indicators, geopolitical events, and market sentiment. Understanding these influences is vital for investors and traders to make informed decisions in a volatile market. In the digital age, news plays a key role in shaping investor sentiment and impacting market trends. Corporate announcements, economic reports, and geopolitical developments often cause significant market movements, making news sentiment analysis crucial for understanding trading activities. Beyond traditional news, social media platforms such as Twitter, Facebook, and Pantip have become important sources of financial information. Social media allows for real-time sentiment analysis, helping investors gauge public perception and identify emerging trends. Consequently, integrating social media sentiment analysis into trading strategies is gaining popularity, offering investors a competitive advantage.

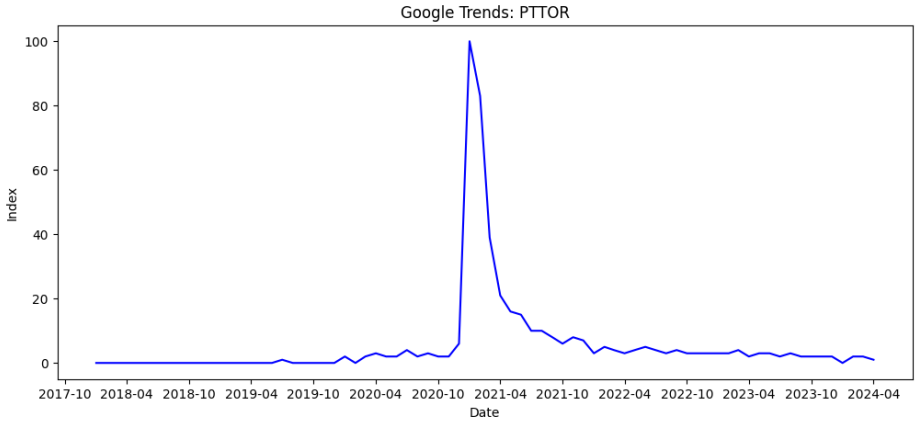


Fig. 1. Interest Over Times of PTTOR since 2018-2023

From Fig 1, the influence of news on market sentiment is the 2021 Initial Public Offering (IPO) of PTT Oil and Retail Business Public Company Limited (PTTOR), where positive coverage drove a surge in investment activity. This IPO led to a significant rise in stock prices and a record increase in new investor accounts in Thailand. Financial news portals like Investing, Kaohoon, and Hoonsmart provide valuable market insights, especially in the Thai language, which can be analyzed using Natural Language Processing (NLP) techniques to further enhance market sentiment analysis.

Table 1. PTTOR Stock price since 2021-2023 (THB)

Date	Price	Open	High	Low	Vol.
2021-02-11	18	-	-	-	-
2021-02-15	34	30.75	36.5	30.5	1.15B
2021-02-16	32.75	35	35	32	479.38M
2021-02-17	29.5	31.5	32.5	29.5	338.39M
2021-02-18	30.25	30.25	31.5	29.75	280.95M
2021-02-19	31.5	31	31.5	30	221.28M
2021-02-22	31.75	32.25	32.5	31.25	175.38M
2021-02-23	31	32	32.25	30.75	184.26M
2021-02-24	30.75	30.75	31.25	30	156.59M
2021-02-25	29.5	31	31.5	29.5	436.98M
2021-03-31	32.25	32.5	32.75	32	52.50M
2021-04-30	30.5	30.75	30.75	30	27.83M
2021-05-31	30	30.25	30.25	30	17.95M
2021-06-30	30.5	30.25	30.75	30.25	24.18M
2021-07-30	27.75	28	28.5	27.5	29.54M
2021-08-31	30.25	30	30.5	29.75	47.57M

Date	Price	Open	High	Low	Vol.
2021-09-30	27.5	27.75	28	27.5	28.45M
2021-10-29	27.5	27.75	27.75	27.5	12.76M
2021-11-30	25	25.5	26	24.9	53.88M
2021-12-30	27	27	27.25	26.75	28.11M
2022-01-31	24.6	24.4	24.8	24.4	22.60M
2022-02-28	25.75	26	26.5	25.75	26.57M
2022-03-31	25	25.25	25.25	25	6.67M
2022-04-29	25	24.6	25.25	24.5	50.61M
2022-05-31	27.75	27.25	27.75	27	69.82M
2022-06-30	25.5	26.25	26.5	25.5	31.29M
2022-07-27	25.25	25.5	25.75	25.25	13.21M
2022-08-31	27.25	27	27.25	26.75	28.96M
2022-09-30	25.75	25.75	26	25.75	8.18M
2022-10-31	24.1	24.2	24.4	24.1	24.45M
2022-11-30	24.4	24.4	24.4	24.2	14.37M
2022-12-30	23.8	23.9	23.9	23.7	14.29M
2023-01-31	22.4	22.6	22.7	22.3	24.79M
2023-02-28	21.9	22	22.2	21.8	22.22M
2023-03-31	21.3	21	21.3	21	12.36M
2023-04-28	22.2	22.3	22.3	22	15.76M
2023-05-31	20	20.1	20.2	20	24.22M

From Table 1, presents stock price movements and trading volumes between February 2021 and May 2023. The highest stock price of 36.5 THB occurred on February 15, 2021, with an exceptionally high trading volume of 1.15 billion shares. After that peak, trading volumes and prices fluctuated, with notable activity on February 25, 2021, when 436.98 million shares were traded. Over time, both the price and volume generally declined, with the volume dropping significantly in later periods, such as 6.67 million shares in March 2022, while the stock price settled at 20 THB by May 2023.

To improve the understanding of market dynamics and investor sentiment, supporting decision-making processes in the financial sector. Various sentiment analysis models were compared, focusing on their application to Thai financial news topics and assessing their effectiveness through key evaluation metrics. Additionally, a comprehensive analysis of different NLP and deep learning architectures was conducted, highlighting their strengths and limitations in capturing nuanced sentiment expressions within the financial context. The findings offer valuable insights into the models' suitability for analyzing market sentiment and trends.

2 Literature Review

2.1 Research related to Sentiment Analysis

Mohan Saloni (2019) [1] investigated stock price prediction through financial news sentiment analysis using a dataset of S&P500 stock prices and over 265,000 news articles. Various models, including ARIMA, Facebook Prophet, and RNN, were tested, with RNN showing superior performance. The study highlighted challenges in predicting volatile stock prices and suggested future research into domain-specific models and broader news integration.

Dilesh Tanna (2020) [2] explored sentiment analysis on social media to understand user emotions and platform dynamics. Using datasets from user engagements and feedback, the study applied lexicon-based libraries and machine learning models, achieving sentiment scores from -1 to +1. These insights aid in content management and user experience improvement, addressing potential issues such as user depression.

Wikanda Phaphan (2020) [3] examined stock price direction forecasting on the Stock Exchange of Thailand by analyzing Thai news with Python's pythainlp and TF-IDF. The study's classification model showed varied accuracy: 100% for Intouch Holdings, 66.67% for Thai Oil and Bumrungrad Hospital, and 66.67% for CP ALL and Kasikornbank, illustrating NLP's potential in stock market predictions.

Kostadin Mishev (2020) [4] evaluated various sentiment analysis models for finance, including lexicon-based methods, ML classifiers, and transformers, using datasets like the Financial Phrase Bank and SemEval 2017-Task 5. The study found BART transformer to be the most effective, achieving the highest MCC score of 0.895, demonstrating the superiority of transformer-based models in capturing financial sentiment nuances.

Kittisak Prachyachuwong (2021) [5] developed a deep learning model integrating numerical and textual data for forecasting stock market trends on Thailand's SET50 index. The model, utilizing industry-specific news vectors, achieved 61.28% accuracy and 59.58% F1-score, outperforming baselines and demonstrating an annualized return of 8.47% in simulated trading.

Priyank Sonkiya (2021) [6] examined stock price prediction using BERT and GAN techniques with news sentiment analysis for Apple Inc. The S-GAN model, which incorporates sentiment analysis as a latent vector, outperformed traditional models in RMSE, showcasing its effectiveness in various prediction intervals.

Pongsatorn Harnmetta (2022) [7] explored sentiment analysis of Thai stock reviews using BERT multilingual and WangchanBERTa. WangchanBERTa achieved the highest accuracy of 92.52%, outperforming BERT multilingual and TF-IDF, illustrating its superior capability in analyzing Thai financial sentiment.

Ponrudee Netisopakul (2022) [8] investigated the impact of daily stock news on stock price direction in Thailand, employing multiple ML text classification methods.

Augmenting news with sentiment and meaningful grouping improved accuracy from 78.6% to 90.6%, highlighting the value of sentiment-enhanced predictions.

Ali Raheman (2022) [9] assessed sentiment analysis models for predicting cryptocurrency market trends using a dataset of 100,000 social media posts. The fine-tuned Aigents model showed the highest correlation (0.57) with sentiment values, emphasizing its relevance for Bitcoin market predictions.

Bledar Fazlija (2023) [10] applied BERT models to predict S&P 500 index direction based on sentiment scores from news articles. FinBERT demonstrated high accuracy, precision, recall, and F1-score, reinforcing the importance of sentiment scores in stock price prediction.

Sindhu (2023) [11] developed a sentiment analysis system for Twitter data focusing on Indian users' responses to events like lockdowns. Using BERT, the system accurately captured sentiment with tokenization and masking techniques, offering valuable insights for decision-making through a user-friendly dashboard.

Suntarin Sangsavate (2023) [12] compared supervised and semi-supervised learning for Thai financial news sentiment analysis. SVM with BERT and LSTM with BERT achieved the highest accuracy of 83.38% and 84.07%, respectively, showing the effectiveness of these models in sentiment classification.

Nattawat Khamphakdee (2023) [13] investigated sentiment analysis of Thai hotel reviews using various deep learning models and word embeddings. WangchanBERTa achieved the highest accuracy (0.9225), outperforming other models, including CNN and skip-gram embeddings.

Chalisa Jitboonyapinit (2023) [14] assessed sentiment analysis in Thai social media using CNN, LSTM, and GRU models. The hybrid CNN-LSTM model achieved the best accuracy of 85.0%, highlighting its superior performance over individual models.

2.2 Research related to Large Language Model Sentiment Analysis

Georgios Fatouros (2023) [15] assessed ChatGPT and FinBERT on forex news headlines with sentiment labels, focusing on metrics like confusion matrix analysis, Mean Absolute Error (MAE), and the correlation of sentiment predictions with market returns. Results showed ChatGPT outperformed FinBERT, achieving about 35% higher accuracy in sentiment classification and a 36% greater correlation with market returns. These findings highlight ChatGPT's superior performance and potential for predicting market trends based on sentiment analysis.

Kiana Kheiri (2023) [16] evaluated GPT-based models on the SemEval Twitter Dataset 2017, comparing them with Convolutional Neural Networks (CNNs), Gated Recurrent Neural Networks (GRNNs), and RoBERTa. Performance was measured using Accuracy, Recall, and F1-score. The study found that GPT-3.5 Turbo achieved a high accuracy of 0.9732, surpassing previous models. Despite this, the research noted

challenges such as data privacy concerns and the potential influence of social biases inherent in the training data on model predictions.

Markus Leippold (2023) [17] examines how sentiment analysis methods, especially keyword-based ones, are vulnerable to adversarial attacks using GPT-3. The study finds GPT-3 effective at changing sentiment while preserving meaning, but more advanced models like FinBERT are more resilient to manipulation. This underscores the need for robust NLP models in the face of adversarial threats.

Boyu Zhang (2023) [18] introduces a retrieval-augmented LLM framework for financial sentiment analysis, utilizing datasets from Twitter Financial News and FiQA. This novel framework outperforms baseline models and other LLMs, including BloombergGPT, ChatGPT, Llama-7B, ChatGLM2-6B, and FinBERT, achieving the highest accuracy and F1 score. The integration of a retrieval-augmented module significantly boosts the framework's effectiveness, highlighting its potential to advance sentiment analysis in finance.

Rodríguez-Ibáñez (2023) [19] reviews sentiment analysis in social networks, highlighting its applications in finance, health, marketing, and politics, with a focus on platforms like Twitter. The review discusses the use of traditional methods and neural networks, while noting limited use of advanced techniques like Transformers. It identifies challenges in applying computationally intensive tools such as GPT-3 and highlights uneven research distribution and emerging research opportunities for improving sentiment analysis methods.

Patthamanan Isaranontakul (2023) [20] evaluated the effectiveness of synthetic text datasets generated by GPT-3 for sentiment analysis using deep learning models BI-LSTM and Bi-GRU. The study utilized various datasets, including manually labeled data, GPT-3 labeled data, synthetic text, and a combination of these. Results showed that models trained with synthetic text achieved notable accuracy improvements, with BI-LSTM reaching 0.84 and Bi-GRU achieving 0.85 accuracy. This underscores the potential of GPT-3-generated synthetic datasets to enhance sentiment analysis accuracy.

From Table 2, the The literature reviewed highlights significant advancements in financial sentiment analysis and stock market trend forecasting., a variety of machine learning and deep learning models, including transformer learning and large language models,. These studies emphasize the need for robust, context-aware techniques to enhance the accuracy and reliability of sentiment analysis in the financial domain. This evolving landscape of methodologies offers promising avenues for further research and innovation in financial sentiment analysis..

Table 2. Literature Review Summary Table

Year	Author(s)	Application/Focus	Input Data	Feature Extraction Models	Machine/Deep Learning Models	Performance Evaluation	Result
2019	Mohan, S.	Stock price prediction using financial news sentiment analysis	S&P500 stock prices, 265,000 financial news articles	Mannual Annotation	ARIMA, Facebook Prophet, RNN	MAPE	The RNN-LSTM model with prices and text polarity as input was identified as the best performing model (MAPE = 2.03 - 2.17) across all experiments, particularly for companies with a higher volume of textual data.
2020	Tanna, D.	Sentiment analysis on social media platforms	User engagements, feedback, and posts	SentiWordNet, Senticnet, LIWC	Naïve Bayes, Maximum Entropy, Support Vector Machine	Sentiment Scores (-1 to +1)	Enhanced analysis efficacy for platform dynamics, content curation, and feedback management.
2020	Phaphan, W.	Forecasting stock price direction in Thai Stock Exchange	Thai news content, stock prices of selected companies	Pythainlp, TF-IDF	Naïve Bayes	Accuracy	Varying prediction accuracies, e.g., Intouch Holdings (100%), Thai Oil (66.67%).
2020	Mishev, K.	Sentiment analysis models comparison in finance	Financial Phrase Bank, SemEval	Mannual Annotation	Lexicon-based, Machine Learning	Matthews's correlation	BART transformer achieved highest MCC score (0.895), highlighting advanced NLP techniques.

Year	Author(s)	Application/Focus	Input Data	Feature Extraction Models	Machine/Deep Learning Models	Performance Evaluation	Result
			2017-Task 5 datasets		classifiers, Deep neural networks	coefficient Score (MCC)	
2021	Prachyachuwong, K	Stock market trend forecasting using deep learning	Historical and technical indicators, news headlines	BERT	Novel deep learning model	Accuracy, F1-score	Achieved 61.28% accuracy and 59.58% F1-score, outperforming baselines with notable annualized return.
2021	Sonkiya, P	Stock price prediction with BERT and GAN	Apple Inc. news headlines from Seeking Alpha	NLTK, finBERT	ARIMA, LSTM, GRU, Vanilla GAN, S-GAN	RMSE	S-GAN outperformed traditional approaches with lowest RMSE (0.5606)
2022	Harnmetta, P	Sentiment analysis of Thai stock reviews	Financial content from Bank of Ayudhya (2017-2021)	Tokenization techniques	BERT multilingual, WangchanBERTa, Logistic Regression	Accuracy	WangchanBERTa achieved superior accuracy 92.52% over BERT multilingual 89.12% and TF-IDF 85.03%.
2022	Netisopakul, P	Impact of daily stock news on stock price direction	Stock news collected in 2018	Word frequency by positive and negative sentiment	Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine, Neural Network	Accuracy	Incorporating sentiments and grouping increased accuracy from 78.6% to 90.6%.

Year	Author(s)	Application/Focus	Input Data	Feature Extraction Models	Machine/Deep Learning Models	Performance Evaluation	Result
2022	Raheman, A	Sentiment analysis for cryptocurrency market prediction	100,000 tweets and Reddit posts	BERT-based models	Lexicon-based, rule-based	Correlation	Fine-tuned Aigents model showed highest correlation with ground truth sentiment values (0.57).
2023	Fazlija, B	Financial markets sentiment analysis using BERT	S&P 500 index news articles	FinBERT,	Random Forest Classifier (RFC)	Accuracy, Precision, Recall	FinBERT demonstrated high classification accuracy 83.60%, Precision 83.90% and Recall 83.60%
2023	Sindhu, M	Sentiment analysis system for Twitter data	Tweets related to significant events in India	Aspect-based method, BERT	Recurrent Neural Network (RNN)	Accuracy	BERT-based model provided accurate 88.52% and syntactic information, aiding decision-making processes.
2023	Sangsavate, C S	Supervised vs. semi-supervised learning for Thai financial news	Thai financial news	PyThaiNLP	SVM, Random Forest, CNN, LSTM	Accuracy	SVM with BERT achieved 83.38% accuracy; LSTM with BERT performed best in deep learning 84.07%.
2023	Khamphakdee, N	Sentiment analysis of Thai hotel reviews	Thai hotel reviews from Agoda.com and Booking.com	Word2Vec (CBOW, Skip-gram)	CNN, LSTM, Bi-LSTM, GRU, Bi-GRU, WangchanBERTa	Accuracy	WangchanBERTa achieved highest accuracy 92.25% among models, highlighting its superiority.

Year	Author(s)	Application/Focus	Input Data	Feature Extraction Models	Machine/Deep Learning Models	Performance Evaluation	Result
2023	Jitboonyapinit, C	Sentiment analysis in Thai social media	Wongnai product and service dataset	Used Dataset sentiment	CNN, LSTM, GRU, CNN-LSTM	Accuracy	CNN-LSTM hybrid model achieved highest accuracy 85.0%, outperforming individual models.
2023	Georgios, F	Evaluation of ChatGPT and FinBERT for forex news sentiment	Forex news headlines	ChatGPT, FinBERT	ChatGPT, FinBERT	Accuracy, MAE, Correlation	ChatGPT showed approximately 35% improvement in sentiment classification accuracy and 36% higher correlation with market returns compared to FinBERT.
2023	Kiana, K	Efficacy of GPT-based models in sentiment analysis	SemEval Twitter Dataset 2017	Used Dataset sentiment	GPT models, CNN, GRNN, RoBERTa	Accuracy, Recall, F1-score	GPT 3.5 Turbo achieved highest accuracy 97.32%, outperforming previous methods, but highlighted data privacy and social bias challenges.
2023	Leippold, M	Vulnerability of sentiment analysis methods to adversarial attacks	Financial Phrase Bank dataset	Used Dataset sentiment	GPT-3, FinBERT, Keyword-based sentiment analysis	Accuracy, Recall	Keyword-based sentiment analysis from 100% to 1.2% for negative sentences in the Financial Phrase Bank, while FinBERT's accuracy dropped from 98% to 91%. This indicates a 99% success rate for GPT-3 against the keyword-based method

Year	Author(s)	Application/Focus	Input Data	Feature Extraction Models	Machine/Deep Learning Models	Performance Evaluation	Result
							and a 7% reduction in accuracy for FinBERT
2023	Zhang, B	Retrieval-augmented LLM for financial sentiment analysis	Twitter Financial News, FiQA datasets	Used Dataset sentiment	Retrieval-augmented LLM, BloombergGPT, ChatGPT	Accuracy, F1-score	Finetuned LLM achieved highest accuracy 81.8 and F1 score 84.2, indicating enhanced predictive capabilities.
2023	Rodríguez-Ibánez, M	Review of sentiment analysis in social networks	Twitter data	N/A	Dictionaries, neural networks, Transformers	N/A	Emphasized diverse applications, highlighting challenges in applying advanced methods and ongoing research opportunities.
2023	Isaranontakul, P	Efficacy of synthetic text datasets from GPT-3 for sentiment analysis	COVID-19 Tweets dataset	Manual Annotation, GPT-3	BI-LSTM, Bi-GRU	Accuracy	Synthetic text from GPT-3 improved model accuracy: BI-LSTM (0.84) and Bi-GRU (0.85).

3 Data and Methodology

3.1 Data

The dataset was collected over a period of 880 days, from January 1, 2021, to May 31, 2023, using reputable sources such as Investing, Kaohoon, and Hoonsmart, which are recognized platforms for Thai financial news. Data extraction was facilitated by the Web Scraper Google Chrome Extension, configured to capture news headlines, dates, and descriptions with pagination support. Data cleansing involved removing unnecessary symbols and standardizing text using the PyThaiNLP library, along with expanding abbreviations to ensure consistency and accuracy in the sentiment analysis. During preprocessing, tokenization was performed using PyThaiNLP, and short sentences were filtered out to focus on substantial content. During the manual annotation and data labeling phase, a systematic approach was adopted to ensure dataset accuracy and relevance. Stratified random sampling was used to select news headlines based on the source and specific dates, representing a comprehensive dataset. These headlines, originally in Thai, were translated into English using Google Translate via the deep-translator package to facilitate annotation. Sentiment classification was conducted using the FinBERT model, designed specifically for financial text analysis. FinBERT provided initial sentiment classifications into positive, neutral, and negative categories, which were manually reviewed and adjusted to ensure accuracy. After this review process, the results were mapped back to the original Thai headlines, with a final dataset of 5,940 labeled headlines: 2,243 positive, 2,197 neutral, and 1,300 negative.

Before training the classification models, different dataset splitting strategies were employed to ensure balanced and representative samples for model evaluation. For models like Logistic Regression, BI-LSTM, and CNN, a Stratified K-Fold cross-validation approach was used, employing 5 splits while shuffling the data and setting a fixed random state of 112. The performance metrics (accuracy, precision, recall, F1-score) from each fold are then aggregated to provide a robust estimate of the model's performance.

For transformer models and large language models, a different strategy was applied. The dataset was split into training (80%) and test sets (20%), again using a random state of 112 to ensure reproducibility. This train-test split method evaluates the model's generalization ability, offering a realistic measure of performance on unseen data. Both approaches help avoid overfitting and ensure that the models perform well across different subsets of the data.

3.2 Methodology

1) **Logistic Regression:** Logistic Regression is a widely used machine learning algorithm for classification tasks, particularly binary and multi-class classification. In this study, it was applied to financial sentiment analysis by converting the text data into numerical features, such as TF-IDF vectors, and modeling the probability of

different sentiment classes (positive, neutral, negative). Logistic Regression is efficient, interpretable, and provides a baseline for comparison against more complex models used in this research.

2) **Bidirectional Long Short-Term Memory (BI-LSTM):** BI-LSTM is an advanced neural network architecture designed for sequential data. It extends the standard LSTM by processing input sequences in both forward and backward directions, capturing dependencies from both past and future words. In this study, BI-LSTM was applied to the sentiment analysis task, using word embeddings to represent financial news headlines. Its bidirectional structure helped capture contextual information more effectively, improving the model's ability to classify sentiment accurately.

3) **Convolutional Neural Network (CNN):** A 1D CNN was utilized in this study for sentiment analysis of financial news headlines. Although typically used for image data, CNNs are effective for text classification by using convolutional filters to detect important patterns and n-grams in the text. In this case, the model applied multiple convolutional layers, followed by pooling and fully connected layers, to predict sentiment labels (positive, neutral, negative). The ability of CNNs to capture local dependencies made it a powerful tool for understanding sentiment in short financial text.

4) **WangChanBERTa:**WangChanBERTa is a transformer model pre-trained on Thai language data, particularly useful for text-based tasks in the Thai language. In this study, WangChanBERTa was fine-tuned on a dataset of financial news headlines for sentiment classification. By leveraging its transformer architecture, which includes self-attention mechanisms, WangChanBERTa could understand the context of each word within a headline, providing highly accurate sentiment predictions tailored to the Thai financial context.

5) **OpenAI:** OpenAI's GPT-based model was employed in this study to analyze fi-nancial news sentiment. Known for its powerful language generation and under-standing capabilities, the model was fine-tuned to classify sentiment in financial text. Using its transformer-based architecture, the model could capture nuanced meanings in the headlines, leading to accurate sentiment classification. The mod-el's strength lies in its ability to process complex text and produce reliable pre-dictions based on deep contextual understanding.

6) **OpenThaiGPT:** OpenThaiGPT is a Thai-language variant of the GPT architecture designed for tasks involving Thai text. In this research, it was fine-tuned for sentiment analysis of financial news headlines in the Thai language. With its ability to understand the unique structure and context of Thai text, OpenThaiGPT effec-tively classified sentiments as positive, neutral, or negative. Its application in this study demonstrated the model's proficiency in handling sentiment analysis tasks in the

Thai financial domain, benefiting from deep learning's advanced natural language processing capabilities.a strong capacity to handle Thai financial sentiment analysis tasks.

3.3 Evaluation Matrix

1) **Accuracy:** Accuracy is the ratio of correctly predicted instances (both positive and negative) to the total number of instances. It is a common evaluation metric but can be misleading when dealing with imbalanced datasets, as it does not distinguish between the types of errors (false positives and false negatives).

2) **Precision:** Precision measures how many of the instances predicted as positive are actually positive. It is crucial when the cost of false positives is high, such as in financial sentiment analysis where incorrectly labeling negative news as positive can lead to misleading conclusions.

3) **Recall:** Recall (or sensitivity) measures how many actual positive instances were correctly predicted. It is important when the cost of missing positive instances (false negatives) is high, such as in identifying significant financial market shifts.

4) **F1-Score:** The F1-score is the harmonic mean of Precision and Recall, offering a balanced metric when you want to consider both false positives and false negatives. It is useful for imbalanced datasets where Accuracy might be misleading

The Knowledge Discovery in Databases (KDD) process is an iterative and structured methodology designed to extract valuable insights from large datasets. This process begins with developing a thorough understanding of the application domain, which defines the project's goals and the context in which the knowledge discovery will be applied. Following this, data selection, preprocessing, and cleansing are carried out to ensure data quality, while transformation techniques such as dimension reduction are applied to tailor the data for analysis. The core phase involves selecting the appropriate data mining task, implementing algorithms, and evaluating the results. Finally, the discovered knowledge is applied in practical settings, ensuring that insights from the analysis are actionable and relevant. This methodical approach guarantees a comprehensive and effective analysis, which forms the theoretical foundation for the practical implementation in this study, as demonstrated in Fig. 2

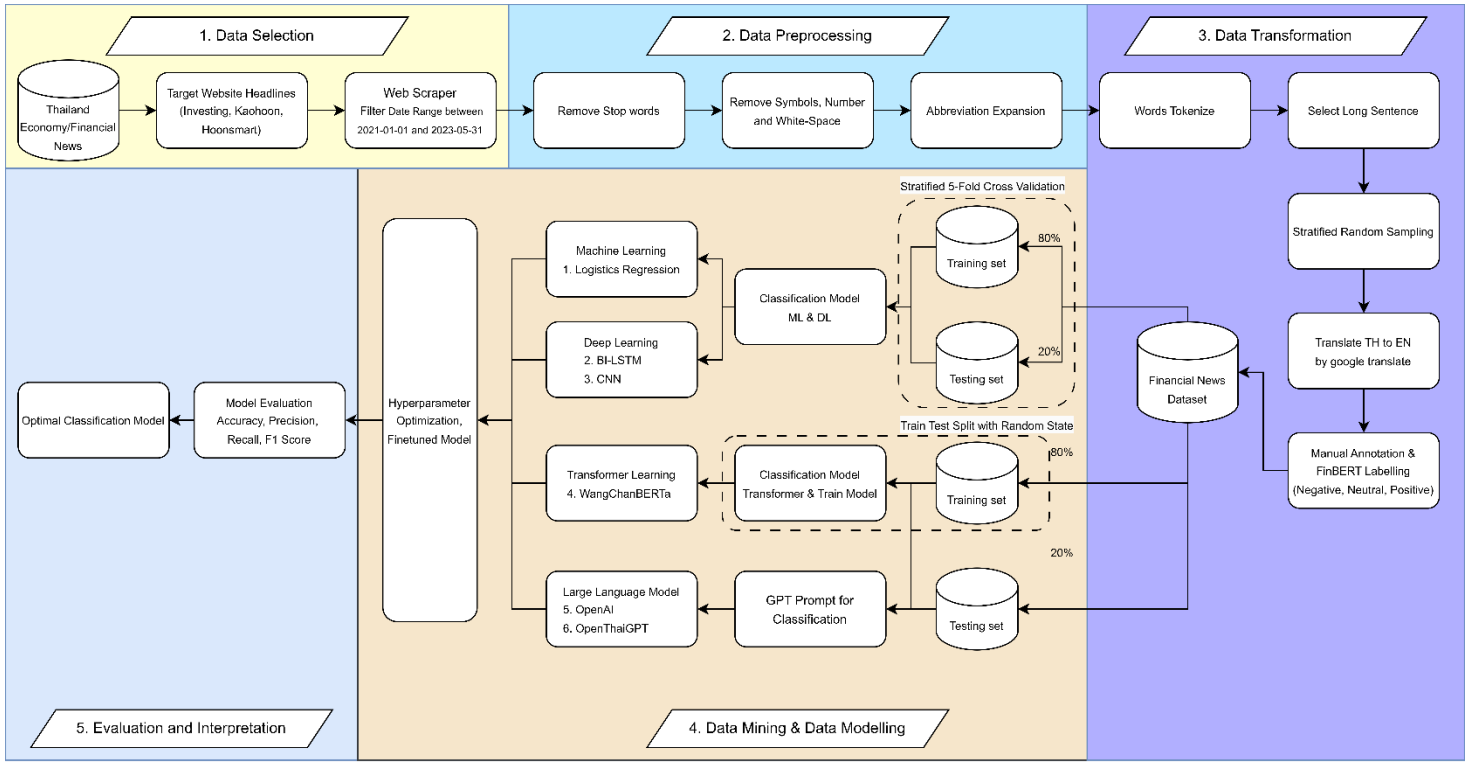


Fig. 2. An overall illustration of sentiment analysis of Thai financial news framework

4 Exploratory Data Analysis and Sentiment Analysis

4.1 Exploratory Data Analysis

In natural language processing (NLP), exploratory data analysis (EDA) is fundamental for comprehending the structure, patterns, and relationships within textual datasets. EDA serves as a preliminary step before implementing advanced models, enabling researchers to gain insights that inform the subsequent stages of preprocessing, feature engineering, and model development. This chapter emphasizes the application of EDA methods tailored for textual data, showcasing techniques and visualizations used to analyze word distributions, sentence structures, and significant linguistic patterns. By uncovering these insights, EDA provides a foundation for more informed and effective NLP modeling.

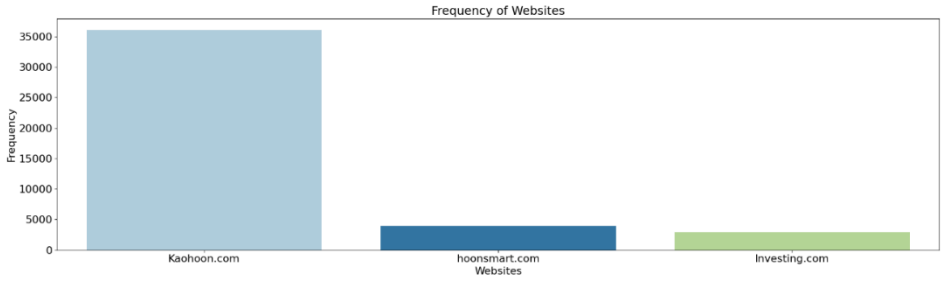


Fig. 3: Frequency of website in collected Thai financial news headlines between 1st Jan 2021 to 30th Jun 2023

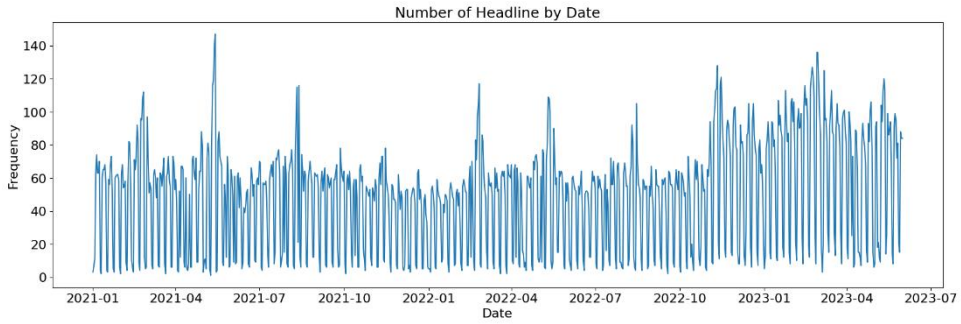


Fig. 4: Number of headlines in collected Thai financial news headlines between 1st Jan 2021 to 30th Jun 2023

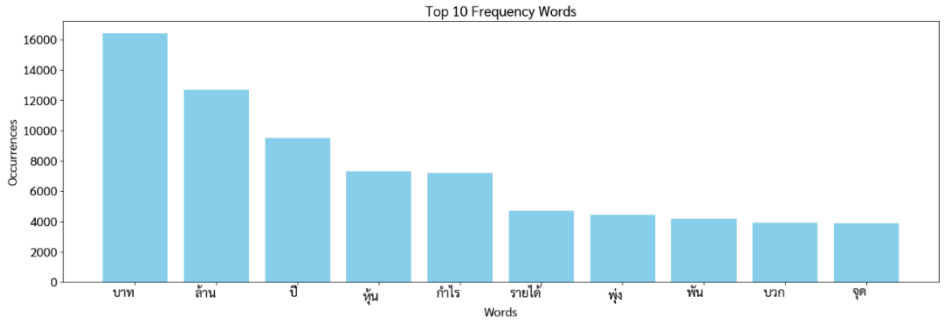


Fig. 5: Top frequency words in collected Thai financial news headlines between 1st Jan 2021 to 30th Jun 2023

The top frequency website for news headlines is Kaohoon.com, as illustrated in Fig. 3. The date with the highest number of headlines is May 14, 2024, In May 2021, Thai financial markets were influenced by the Bank of Thailand's measures to support economic recovery amid the COVID-19 pandemic, maintaining policy rates and providing support to affected sectors. The stock market saw fluctuations due to global economic uncertainties and domestic economic data, with significant attention on tourism, healthcare, and technology sectors. In February 2023, Thai Airways' major restructuring plan to improve efficiency and reduce debt was a key highlight, alongside robust performance in banking, energy, and real estate sectors, driven by positive corporate earnings and investor confidence boosted by government economic policies and infrastructure development plans in Fig. 4 Furthermore, the most frequently occurring words in the dataset are Baht, Million, and Year, highlighting common themes of currency, financial quantities, and timeframes in the news, as shown in Fig. 5

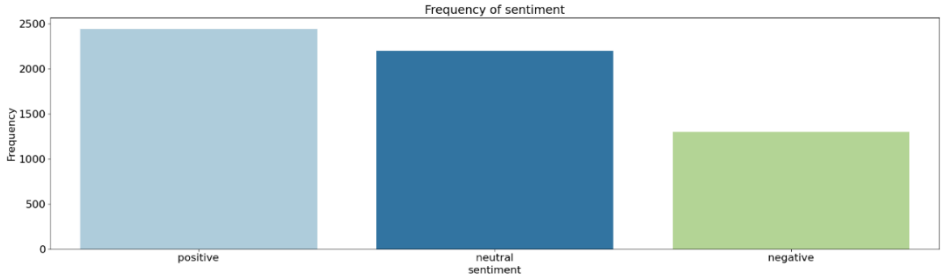


Fig. 6: The number of Sentiment in sampled Thai financial news headline dataset

Fig. 6 illustrates the distribution of sentiment labels in the dataset, with positive sentiment being the most frequent at around 2500 occurrences, followed by neutral sentiment with approximately 2200 instances, and negative sentiment being the least common at around 1500 occurrences. This indicates an imbalanced distribution, where positive and neutral sentiments dominate the dataset compared to negative sentiments. Understanding this distribution is important for further analysis, as it provides insight into the dataset's overall sentiment tendencies and potential challenges in sentiment classification.

4.2 Sentiment Analysis

The sentiment analysis conducted on financial news headlines employed a range of machine learning and deep learning models to assess their effectiveness in capturing nuanced sentiment within the finance domain. The models included traditional machine learning techniques, such as logistic regression, deep learning architectures like Bidirectional Long Short-Term Memory (BI-LSTM) and Convolutional Neural Networks (CNN). In addition, advanced transformer-based models, such as WangchanBERTa and fine-tuned WangchanBERTa, and large language models, including OpenAI and OpenThai GPT, were utilized. The evaluation of these methodologies was centered on key performance metrics, namely accuracy, precision, recall, and F1-score.

4.2.1 Logictic Regression

Logistic Regression model was used to classify sentiments in financial news article titles, utilizing a 5-fold Stratified K-Fold Cross-Validation for robust evaluation. The dataset comprised cleaned and tokenized titles as features, with the sentiment column serving as the target variable. Feature extraction was limited to a maximum of 5,000 features to manage dimensionality. The model was trained to identify patterns between features and sentiment labels and evaluated using precision, recall, F1-score, and a confusion matrix. Hyperparameter tuning was performed using GridSearchCV to identify optimal model parameters Table 3.

Table 3. Logistic regression hyperparameter tuning GridSearchCV

Hyperparameter	Values Tried	Optimal Value
C	0.01, 0.1, 1, 10	1
Solver	lbfgs, newton-cg, sag, saga	saga
Penalty	l2 (for lbfgs, newton-cg, sag)	-
	elasticnet, l1, l2 (for saga)	elasticnet
l1 ratio	0.1, 0.3, 0.5, 0.7, 0.9 (only for elasticnet)	0.7
max_iter	100	100

For each trial, data was split into training and testing sets using 5-fold Stratified K-Fold Cross-Validation to ensure class balance in each fold. The model was trained on the training set and evaluated on the testing set, with key metrics including accuracy, precision, recall, F1-score, and log loss computed for both sets. These metrics were averaged across the folds to provide an overall performance estimate for each trial. The entire process was repeated across five trials, and the results were aggregated to compute average metrics and their standard deviations, offering a comprehensive view of the model's central tendency and variability. The logistic regression model achieved

robust training accuracies between 86.48% and 86.50%, with a mean accuracy of 86.49%. However, validation accuracies ranged from 76.58% to 77.61%, with a mean of 76.61%, highlighting the model's generalization limitations. Table 4. presents these training and testing metrics, including precision, recall, and F1-score, revealing insights into the model's performance in classifying sentiment categories.

Table 4. Logistic regression model trials comparison table

Dataset	Trials	Accuracy	Precision	Recall	F1-score
Training Set	Trial-1	0.8650	0.8651	0.8650	0.8647
	Trial-2	0.8648	0.8649	0.8648	0.8645
	Trial-3	0.8648	0.8650	0.8648	0.8646
	Trial-4	0.8650	0.8651	0.8650	0.8647
	Trial-5	0.8648	0.8650	0.8648	0.8645
	MEAN ± SD	**0.8649 ± 0.0001**	**0.8651 ± 0.0001**	**0.8649 ± 0.0001**	**0.8647 ± 0.0001**
Testing Set	Trial-1	0.7661	0.7670	0.7661	0.7657
	Trial-2	0.7658	0.7667	0.7658	0.7653
	Trial-3	0.7663	0.7672	0.7663	0.7658
	Trial-4	0.7663	0.7671	0.7663	0.7658
	Trial-5	0.7659	0.7668	0.7659	0.7655
	MEAN ± SD	**0.7661 ± 0.0002**	**0.7670 ± 0.0002**	**0.7661 ± 0.0002**	**0.7657 ± 0.0002**

4.2.2 Bidirectional Long Short-Term Memory (BI-LSTM)

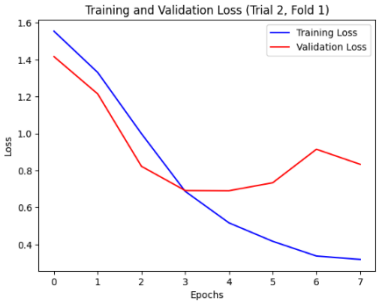
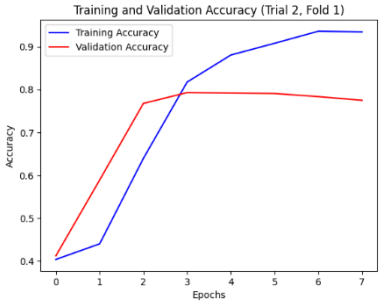
Bidirectional Long Short-Term Memory (BI-LSTM) model for sentiment classification on financial news articles. To ensure rigorous evaluation, a 5-fold Stratified K-Fold Cross-Validation technique was employed with a fixed random state (random_state=112) to maintain consistency and reproducibility in data splits. The dataset consisted of preprocessed and tokenized article titles from the 'title_cleaned' column, with the sentiment labels from the 'sentiment' column. The text data was tokenized and padded to create uniform sequence lengths suitable for LSTM input. The model architecture included an embedding layer followed by two Bidirectional LSTM layers with dropout regularization, dense layers for feature extraction, and an output layer for classification. The BI-LSTM was trained with an embedding dimension of 100, 64 LSTM units, a 0.7 dropout rate, and L2 regularization ($\lambda = 0.001$), using the Adam optimizer and sparse categorical cross-entropy loss. The model was trained for 10 epochs with a batch size of 128. Hyperparameter tuning was performed to further enhance the model's performance, as detailed in Table 5.

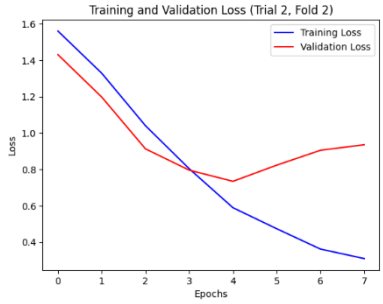
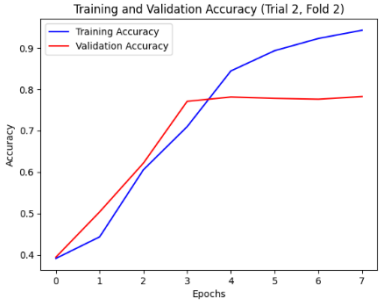
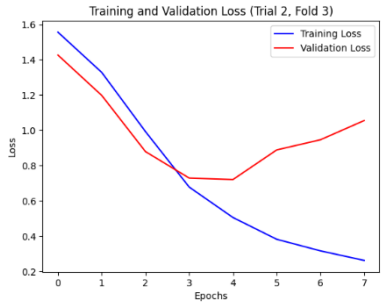
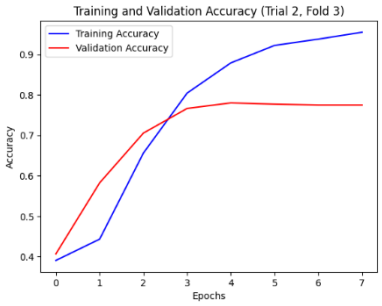
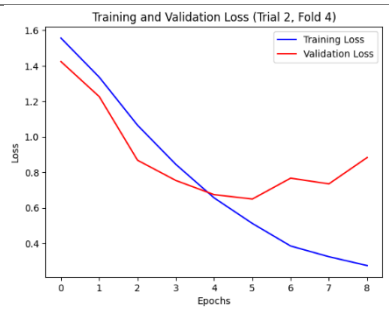
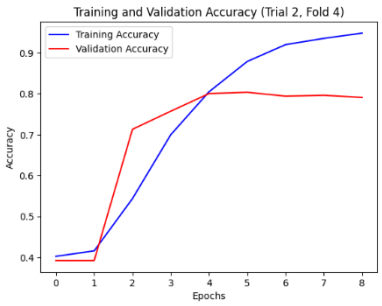
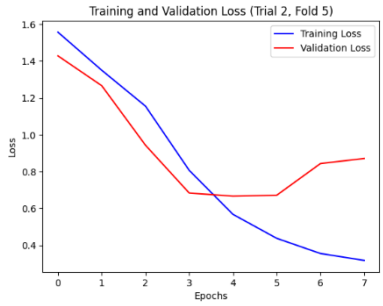
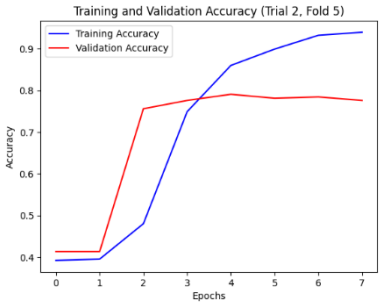
Table 5. BI-LSTM hyperparameter tuning GridSearchCV

Hyperparameter	Values Tried	Optimal Value
embedding_dim	50, 100, 200	100
lstm_units	64, 128, 256	64
dropout_rate	0.3, 0.5, 0.7	0.7
l2_reg	0.01, 0.001, 0.0001	0.001
batch_size	32, 64, 128	128
epochs	5, 10, 15	10
earlystopping(patience)	1, 2, 3	3

The BI-LSTM model's training yielded its optimal accuracy in Trial 2, with the highest performance observed in the first fold. The training process demonstrated the model's ability to learn and adapt to the data effectively, though it also revealed challenges related to overfitting. To enhance generalization in future iterations, strategies such as refining regularization techniques, adjusting dropout rates, or applying early stopping criteria could be beneficial. These adjustments would likely improve the model's ability to generalize beyond the training data, as outlined in Table 6.

Table 6. BI-LSTM Model training and validation result

Train/Validation Loss	Train/Validation Accuracy
	

Train/Validation Loss	Train/Validation Accuracy
	
	
	
	

The BI-LSTM model exhibited strong learning capabilities across trials, with peak training accuracy reaching 87.87%. However, validation accuracy fluctuated between 75% and 80.34%, indicating potential overfitting as the training progressed. While the model demonstrated effectiveness in learning from the training data, further enhancements such as early stopping and dropout regularization could help improve generalization and stabilize validation performance. Performance evaluation was conducted using accuracy, precision, recall, and F1-score metrics, calculated through the scikit-learn classification_report() function. Training accuracies ranged from 87.79% to 89.73%, with the highest accuracy recorded at 89.73%. Validation accuracies varied between 77.11% and 79.28%, and testing accuracies were between 78.11% and 78.28%. These results reflect the model's ability to generalize while also revealing areas for improvement. The precision, recall, F1-score metrics, and the Model Trials Comparison Table are summarized in Table 7., providing a comprehensive evaluation of the model's performance and guiding future enhancements in sentiment analysis for financial applications.

Table 7. BI-LSTM Model Trials Comparison Table

Dataset	Trials	Accuracy	Precision	Recall	F1-score
Training Set	Trial-1	0.8931	0.8935	0.8931	0.8930
	Trial-2	0.8973	0.8986	0.8973	0.8973
	Trial-3	0.8779	0.8799	0.8779	0.8781
	Trial-4	0.8787	0.8801	0.8787	0.8785
	Trial-5	0.8882	0.8900	0.8882	0.8883
	MEAN ± SD	**0.8870 ± 0.0077**	**0.8884 ± 0.0074**	**0.8870 ± 0.0077**	**0.8870 ± 0.0077**
Testing Set	Trial-1	0.7763	0.7768	0.7763	0.7761
	Trial-2	0.7828	0.7849	0.7828	0.7829
	Trial-3	0.7825	0.7849	0.7825	0.7823
	Trial-4	0.7822	0.7843	0.7822	0.7816
	Trial-5	0.7811	0.7840	0.7811	0.7809
	MEAN ± SD	**0.7810 ± 0.0024**	**0.7830 ± 0.0031**	**0.7810 ± 0.0024**	**0.7808 ± 0.0024**

4.2.3 Convolutional Neural Networks (CNN)

The Convolutional Neural Network (CNN) model was implemented to classify sentiments from financial news articles. A 5-fold Stratified K-Fold Cross-Validation technique was employed, using a fixed random state (random_state=112) to ensure consistency in data splitting, enhancing the robustness and reproducibility of the results. The dataset consisted of cleaned and tokenized article titles from the 'title_cleaned' column, with the 'sentiment' column as the target variable. To ensure uniform input lengths, the text data was tokenized and padded, with the maximum sequence length determined dynamically. The CNN model architecture featured an embedding layer for word embeddings, followed by a one-dimensional convolutional layer and a global

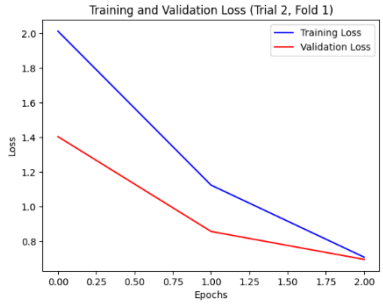
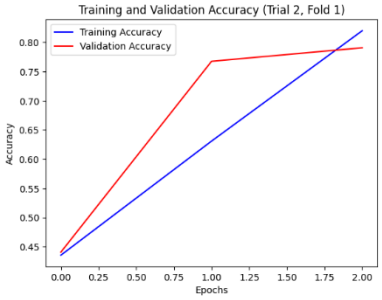
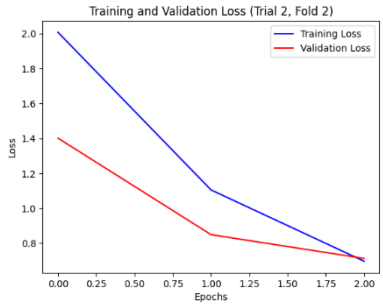
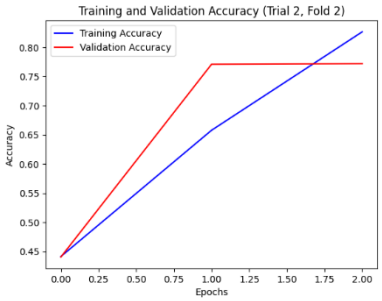
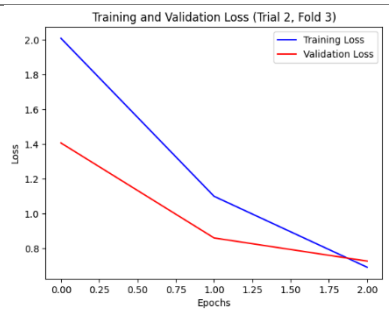
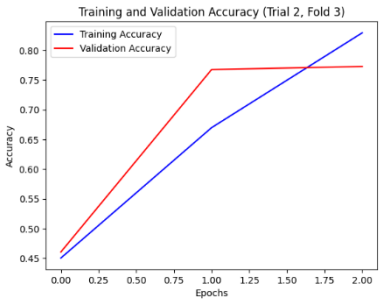
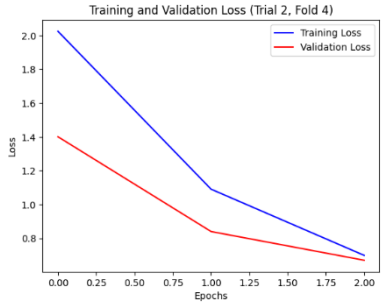
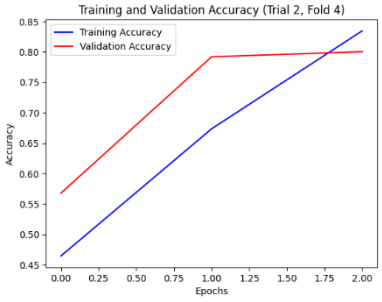
max-pooling layer to capture key features. Dense layers were incorporated for feature extraction and sentiment classification. The model was trained using the Adam optimizer and sparse categorical cross-entropy loss. Key hyperparameters included an embedding dimension of 100, 128 filters, a kernel size of 5, a dropout rate of 0.5, and L2 regularization of 0.01. Training was conducted for 3 epochs with a batch size of 32. Hyperparameter tuning results are detailed in Table 8

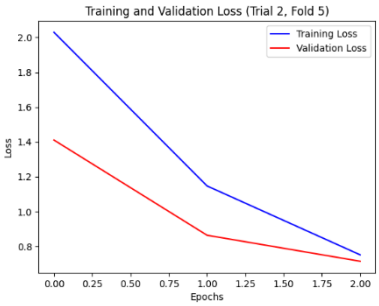
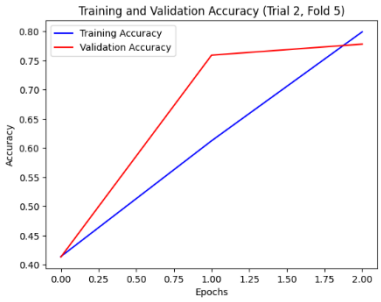
Table 8. CNN hyperparameter tuning GridSearchCV

Hyperparameter	Values Tried	Optimal Value
filters	128	128
kernel_size	5	5
dense_units	64	64
dropout_rate	0.5	0.5
l2_reg	0.01, 0.001, 0.0001	0.01
batch_size	32, 64, 128	32
epochs	3, 5, 10	3
earlystopping (patience)	1, 3, 5	1

During the CNN model training, the optimal accuracy was observed in Trial 2, particularly in the first fold of the cross-validation process. The training demonstrated the model's effective learning and adaptability, with clear improvements over multiple trials. However, the results also indicated the potential for overfitting, which could limit the model's generalization to unseen data. To address this, future iterations could benefit from additional regularization techniques, such as adjusting dropout rates or employing early stopping to prevent overtraining. These strategies would likely improve the model’s generalization capabilities, as shown in Table 9

Table 9. CNN Model training and validation result

Train/Validation Loss	Train/Validation Accuracy
	
	
	
	

Train/Validation Loss	Train/Validation Accuracy
	

The CNN model exhibited strong learning capabilities across trials, with training accuracies consistently improving, reaching an average of 87.08%. However, validation accuracy fluctuated significantly between 44.06% and 80.02%, suggesting potential overfitting as training progressed. While the model effectively captured patterns in the training data, incorporating strategies such as early stopping or adjusting dropout rates may help improve generalization and stabilize validation performance. Training accuracies ranged from 86.92% to 87.55%, with the best trial achieving 87.55%. Validation and testing accuracies showed lower variability, ranging from 78.70% to 79.34%, reflecting robust performance on unseen data with good generalizability. For a comprehensive comparison of trials, Table 10 presents detailed metrics, offering insights into the model's performance across experiments.

Table 10. CNN Model Trials Comparison Table

Dataset	Trials	Accuracy	Precision	Recall	F1-score
Training Set	Trial-1	0.8753	0.8764	0.8753	0.8746
	Trial-2	0.8708	0.8722	0.8708	0.8700
	Trial-3	0.8731	0.8738	0.8731	0.8727
	Trial-4	0.8755	0.8764	0.8755	0.8748
	Trial-5	0.8692	0.8706	0.8692	0.8682
	MEAN ± SD	**0.8728 ± 0.0024**	**0.8739 ± 0.0023**	**0.8728 ± 0.0024**	**0.8721 ± 0.0026**
Testing Set	Trial-1	0.7919	0.7943	0.7919	0.7909
	Trial-2	0.7934	0.7973	0.7934	0.7922
	Trial-3	0.7923	0.7950	0.7923	0.7917
	Trial-4	0.7870	0.7889	0.7870	0.7858
	Trial-5	0.7899	0.7933	0.7899	0.7887
	MEAN ± SD	**0.7909 ± 0.0022**	**0.7938 ± 0.0027**	**0.7909 ± 0.0022**	**0.7899 ± 0.0024**

4.2.4 WangchanBERTa

The WangChanBERTa model, based on the BERT architecture, is designed for sentiment analysis, employing attention mechanisms and SentencePiece tokenization for effective text classification. Using the pre-trained "airesearch/wangchanberta-base-att-spm-uncased" model, researchers can load weights and tokenize input text with ease. Initially applied to the Wisesight Sentiment dataset, the model struggled to distinguish between sentiment categories, particularly the Negative class. To address this, a Train-Test split (random_state=112) was used for consistent evaluation.

Fine-tuning the WangChanBERTa model improved its ability to capture sentiment nuances, resulting in enhanced precision, recall, and F1-scores across all sentiment classes. The dataset used preprocessed text from financial news article titles, tokenized and padded for uniform input. The fine-tuned model significantly outperformed the pre-trained version, demonstrating its effectiveness in sentiment classification, especially in financial contexts. The fine-tuning process was conducted with a learning rate of 2e-5, a batch size of 5, and 3 epochs, with performance metrics detailed in Table 11

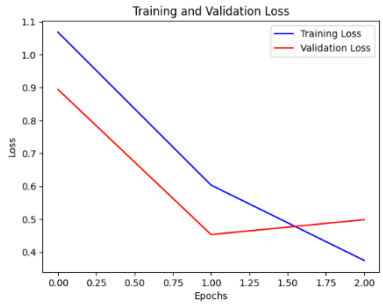
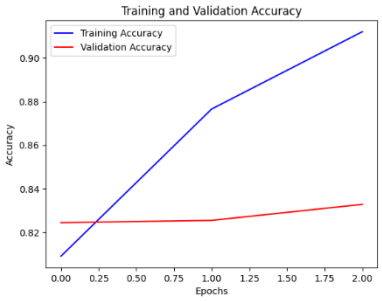
Table 11. Fine-tuned WangchanBERTa hyperparameter tuning GridSearchCV

Hyperparameter	Values Tried	Optimal Value
per_device_train_batch_size	5	5
num_train_epochs	3, 5, 7, 10	3
learning_rate	2e-5	2e-5
save_steps	200	200
save_total_limit	2	2
weight_decay	0.01	0.01
gradient_accumulation_steps	2	2
warmup_steps	500	500
logging_dir	"./logs"	"./logs"
logging_steps	200	200
evaluation_strategy	"steps"	"steps"
eval_steps	200	200
load_best_model_at_end	TRUE	TRUE

The training process highlights the effectiveness of the WangChanBERTa model in capturing sentiment nuances within the financial news dataset. Consistent improvements were noted in both accuracy and loss metrics over the course of the training,

demonstrating the model's capacity to learn and refine its sentiment classification abilities. These improvements, along with detailed performance metrics, are presented in Table 12

Table 12. Fine-tuned WangchanBERTa training and validation result

Train/Validation Loss	Train/Validation Accuracy
	

During the training of the WangChanBERTa model, notable improvements were observed across three epochs. In the first epoch, the model achieved a training accuracy of 80.90%, a validation accuracy of 82.44%, and losses of 1.0682 (training) and 0.8942 (validation). By the second epoch, the training accuracy increased to 87.66%, while the validation accuracy remained stable at 82.55%, with training and validation losses dropping to 0.6036 and 0.4533, respectively. In the final epoch, training accuracy rose further to 91.21%, with a validation accuracy of 83.28%, and corresponding losses of 0.3747 (training) and 0.4985 (validation). These metrics show continuous improvement in the model's performance over the epochs, as presented in Table 4.10.

The fine-tuning of the WangChanBERTa model significantly enhanced its performance. Initially, the model struggled, particularly with the Negative sentiment class, showing a precision of only 0.01 and a weighted average F1-score of 0.54, with an overall accuracy of 37%. However, after fine-tuning, the model's performance drastically improved, achieving precisions of 0.83 for Negative, 0.83 for Neutral, and 0.90 for Positive sentiment, with an overall weighted F1-score of 0.86 and an accuracy of 86% on the training set. In testing, the fine-tuned model maintained strong performance, attaining accuracies of 78%, 79%, and 90% for Negative, Neutral, and Positive sentiments, respectively, resulting in an overall testing accuracy of 84%. These results underscore the effectiveness of fine-tuning in enhancing the model's sentiment analysis capabilities. Detailed metrics can be found in Table 13

Table 13. Classification Report of WangchanBERTa Table

Model			Precision	Recall	F1-score	Support
WangChanBERTa	Training Set	Negative	0.01	0.37	0.01	19
		Neutral	1.00	0.37	0.54	4733
		Positive	0.00	0.00	0.00	0
		Accuracy			0.37	4752
		Macro AVG	0.33	0.25	0.19	4752
		Weighted AVG	0.99	0.37	0.54	4752
	Testing Set	Negative	0.02	0.83	0.04	6
		Neutral	1.00	0.36	0.53	1182
		Positive	0.00	0.00	0.00	0
		Accuracy			0.36	1188
		Macro AVG	0.34	0.40	0.19	1188
		Weighted AVG	1.00	0.36	0.53	1188
Fine-Tuned WangChanBERTa	Training Set	Negative	0.83	0.88	0.86	962
		Neutral	0.83	0.86	0.84	1716
		Positive	0.90	0.85	0.87	2074
		Accuracy			0.86	4752
		Macro AVG	0.85	0.86	0.86	4752
		Weighted AVG	0.86	0.86	0.86	4752
	Testing Set	Negative	0.78	0.87	0.83	247
		Neutral	0.79	0.84	0.81	399
		Positive	0.90	0.81	0.86	542
		Accuracy			0.84	1188
		Macro AVG	0.83	0.84	0.83	1188
		Weighted AVG	0.84	0.84	0.84	1188

4.2.5 OpenAI

This section details the application of OpenAI's GPT-3.5 model for sentiment analysis on financial news articles. A specific prompt command was crafted to direct the model's classification of sentiment into one of three categories: 'Negative,' 'Neutral,' or 'Positive.' The prompt used was:

"Classify the sentiment of the following financial news in 3 classes 'Negative', 'Neutral', 'Positive': '{text}'."

This prompt guided the model to focus on sentiment analysis based on the given financial news text. Utilizing this approach, GPT-3.5 provided predictions reflecting sentiment trends in the financial domain

To ensure consistent evaluation, an 80-20 train-test split was applied with a fixed random state (random_state=112), ensuring reproducible and reliable results. Key parameters for this analysis included the GPT-3.5 engine 'gpt-3.5-turbo-instruct,' a maximum token limit of 50 to regulate output length, and a temperature setting of 0.7, balancing creativity and determinism in model responses. These configurations, presented in Table 14, enabled effective sentiment classification within the financial context.

Table 14. OpenAI parameter

Parameter	Values
engine	gpt-3.5-turbo-instruct
prompt	Classify the sentiment of the following financial news in 3 classes 'Negative', 'Neutral', 'Positive'): '{text}'
max tokens	50
temperature	0.7

The results indicate that OpenAI's model performed moderately well in classifying Neutral sentiment but struggled to accurately identify Negative and Positive sentiments, especially in real-world testing scenarios. The classification report for financial news sentiment analysis showed that in the training set, the model achieved an overall accuracy of 63%, with the best performance in the Neutral category (precision of 0.53 and recall of 0.84). However, it had difficulty with Negative sentiment, showing lower precision (0.72) and recall (0.59).

In the testing set, the model's overall accuracy dropped slightly to 61%. While it maintained relatively good performance for Neutral sentiment, its precision and recall for Negative and Positive sentiments were suboptimal, indicating challenges in these categories. This suggests that while the model is effective at recognizing Neutral sentiment, it requires further refinement to enhance its classification capabilities for Negative and Positive sentiments, as detailed in Table 15

Table 15. Classification Report of OpenAI Table

Confusion Matrix		Precision	Recall	F1-score	Support
Training Set	Negative	0.72	0.59	0.65	1026
	Neutral	0.53	0.84	0.65	1771
	Positive	0.85	0.47	0.60	1955
	Accuracy			0.63	4752
	Macro AVG	0.70	0.63	0.63	4752
	Weighted AVG	0.70	0.63	0.63	4752

Confusion Matrix		Precision	Recall	F1-score	Support
Testing Set	Negative	0.75	0.57	0.64	274
	Neutral	0.50	0.85	0.63	426
	Positive	0.85	0.44	0.58	488
	Accuracy			0.62	1188
	Macro AVG	0.70	0.62	0.62	1188
	Weighted AVG	0.70	0.62	0.61	1188

4.2.6 OpenThai GPT

This section covers the application of OpenThaiGPT for sentiment analysis of Thai financial news articles. A Thai-language prompt was formulated to classify the sentiment of the given news text into 'Negative,' 'Neutral,' or 'Positive' categories, guiding the model's response generation effectively. The prompt used was:

“จำแนกอารมณ์ของข่าวการเงินต่อไปนี้ โดยตอบเป็น {'เชิงลบ', 'เป็นกลาง','เชิงบวก'} :{title}”

OpenThaiGPT's language processing capabilities were leveraged, with the LlamaCPP library used to fine-tune model parameters. This included setting the temperature to 0.1 and limiting responses to 256 tokens to optimize performance and efficiency. This methodology showcases the effective use of NLP for analyzing the sentiment of Thai financial news, providing valuable insights into the Thai financial market, as shown in Table 16.

Table 16. OpenThaiGPT parameter

Parameter	Values
engine	OpenThai GPT
prompt	จำแนกอารมณ์ของข่าวการเงินต่อไปนี้ โดยตอบเป็น {'เชิงลบ', 'เป็นกลาง','เชิงบวก'} :{title}”
max tokens	50
temperature	0.7

The results of the sentiment analysis using OpenThaiGPT on Thai financial news articles demonstrate the model's strength in identifying Negative sentiments, with good precision and recall in both the training and testing sets. However, the model struggles with accurately classifying Neutral sentiments, particularly in real-world testing scenarios. Despite achieving a precision of 0.78 for Neutral sentiment in the training set, its recall was notably low at 0.11, and the performance did not improve in the testing

set. The overall accuracies were 0.64 and 0.62 for the training and testing sets, respectively, indicating room for improvement as detailed in Table 17

These findings suggest that while OpenThaiGPT can effectively capture Negative sentiments, it requires further refinement to enhance its performance across all sentiment categories, particularly for Neutral sentiments. Future work could involve additional fine-tuning or incorporating a more diverse dataset to better equip the model for real-world applications.

Table 17. Classification Report of OpenThaiGPT Table

Confusion Matrix		**Precision**	**Recall**	**F1-score**	**Support**
Training Set	Negative	0.66	0.92	0.77	1026
	Neutral	0.78	0.11	0.20	1771
	Positive	0.61	0.96	0.75	1955
	Accuracy			0.64	4752
	Macro AVG	0.98	0.66	0.57	4752
	Weighted AVG	0.69	0.64	0.55	4752
Testing Set	Negative	0.67	0.89	0.76	274
	Neutral	0.68	0.07	0.13	426
	Positive	0.60	0.95	0.73	488
	Accuracy			0.62	1188
	Macro AVG	0.65	0.64	0.54	1188
	Weighted AVG	0.64	0.62	0.52	1188

5 Conclusion and Future work

This chapter summarizes the key findings from sentiment analysis on financial news headlines, evaluating the performance of Logistic Regression, Bidirectional Long Short-Term Memory (BI-LSTM), Convolutional Neural Network (CNN), WangChan-BERTa, Fine-tuned WangChanBERTa, OpenAI's GPT-3.5, and OpenThaiGPT. The research aimed to determine the effectiveness of these methods in accurately classifying sentiment within the finance domain, highlighting both the challenges faced and the successes achieved during the analysis.

Table 18. Summary of sentiment analysis model performance metrics

	Model	Accuracy	Precision	Recall	F1-score
Training Set	Logistic Regression	0.8649 ± 0.0001	0.8651 ± 0.0001	0.8649 ± 0.0001	0.8647 ± 0.0001
	BI-LSTM	**0.8870 ± 0.0077**	**0.8884 ± 0.0074**	**0.8870 ± 0.0077**	**0.8870 ± 0.0077**
	CNN	**0.8728 ± 0.0024**	**0.8739 ± 0.0023**	**0.8728 ± 0.0024**	**0.8721 ± 0.0026**
	WangChanBERTa	0.3700± 0.0000	0.3300± 0.0000	0.2500± 0.0000	0.1900± 0.0000
	Fine-tuned WangChanBERTa	**0.8600± 0.0000**	**0.8500± 0.0000**	**0.8600± 0.0000**	**0.8600± 0.0000**
	Open AI (GPT3.5)	0.6300± 0.0000	0.7000± 0.0000	0.6300± 0.0000	0.6300± 0.0000
	OpenThaiGPT	0.6400± 0.0000	0.9800± 0.0000	0.6600± 0.0000	0.5700± 0.0000
Testing Set	Logistic Regression	0.7661 ± 0.0002	0.7670 ± 0.0002	0.7661 ± 0.0002	0.7657 ± 0.0002
	BI-LSTM	**0.7810 ± 0.0024**	**0.7830 ± 0.0031**	**0.7810 ± 0.0024**	**0.7808 ± 0.0024**
	CNN	**0.7909 ± 0.0022**	**0.7938 ± 0.0027**	**0.7909 ± 0.0022**	**0.7899 ± 0.0024**
	WangChanBERTa	0.3600± 0.0000	0.3400± 0.0000	0.4000± 0.0000	0.1900± 0.0000
	Fine-tuned WangChanBERTa	**0.8400± 0.0000**	**0.8300± 0.0000**	**0.8400± 0.0000**	**0.8300± 0.0000**
	Open AI (GPT3.5)	0.6200± 0.0000	0.7000± 0.0000	0.6200± 0.0000	0.6200± 0.0000
	OpenThaiGPT	0.6200± 0.0000	0.6500± 0.0000	0.6400± 0.0000	0.5400± 0.0000

Table 18 presents a comparative analysis of the performance metrics across various models used for sentiment analysis of financial news headlines, including Logistic Regression, BI-LSTM, CNN, WangChanBERTa, Fine-tuned WangChanBERTa, OpenAI (GPT-3.5), and OpenThaiGPT. Among these, the Fine-tuned WangChanBERTa model achieved the highest testing accuracy of 0.8400 ± 0.0000, demonstrating superior capability in accurately classifying sentiment in Thai financial news articles. The BI-LSTM model also performed robustly, with a testing accuracy of 0.7810 ± 0.0024, and the CNN model followed closely with a testing accuracy of 0.7909 ± 0.0022. The Logistic Regression model, while less complex, recorded a respectable testing accuracy of 0.7661 ± 0.0002, showcasing its reliability in sentiment classification tasks.

Fine-tuning LLMs on a domain-specific dataset, selecting optimal parameters, and using an effective prompt are essential for enhancing model performance. Training on labeled financial data while optimizing hyperparameters such as learning rate, batch

size, and epochs allows the model to better capture linguistic patterns in financial sentiment analysis. This study highlights the challenges of deploying large-scale models in resource-constrained environments, emphasizing the need for fine-tuning and prompt optimization. Future work should focus on refining these models through domain-specific training, parameter optimization, and improved prompt engineering to enhance sentiment classification accuracy in financial contextsThe results of this study suggest several directions for future research to enhance sentiment analysis models for Thai financial news. Proposed future work includes:

1. Training OpenAI's GPT-3.5 and OpenThaiGPT with Custom Datasets: Further research will involve training these models on the existing financial news dataset to evaluate their effectiveness in capturing nuanced sentiment expressions. This will help assess the models' ability to classify sentiment categories (Negative, Neutral, Positive) in Thai financial articles, offering deeper insights into market dynamics.
2. Optimizing Prompts for Large Language Models: Systematic exploration of different prompt formulations will be undertaken to guide models like GPT-3.5 and OpenThaiGPT more effectively. The goal is to develop optimal prompts that improve model performance in sentiment classification by experimenting with various phrasings, contexts, and formats.
3. Fine-tuning WangChanBERTa with Expanded Datasets: Future work will involve fine-tuning WangChanBERTa using a more comprehensive dataset, including new financial news articles. Adjustments to the model’s architecture and hyperparameters will be made to enhance its generalization capabilities, enabling it to perform accurately across a broader range of sentiments and contexts.

These initiatives aim to advance the field of sentiment analysis, providing valuable contributions for both academic research and practical applications within the financial domain.

References

1. S. Mohan, "Stock Price Prediction Using News Sentiment Analysis," 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications, pp. pp. 205-208, 2019.
2. D. Tanna, M. Dudhane, A. Sardar, K. Deshpande and N. Deshmukh, "Sentiment Analysis on Social Media for Emotion Classification," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 911-915, 2020.
3. W. Phaphan, "The predictions of a daily stock price direction from the Thai news content by using natural language processing," Journal of Applied Science, pp. 19(1):59-79, 2020.
4. K. Mishev, "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers," IEEE Access, vol. 8, pp. 131662-131682, 2020.
5. K. Prachyachuwong and P. Vateekul, "Stock Trend Prediction Using Deep Learning Approach on Technical Indicator and Industrial Specific Information," Information, vol. 250, p. 12, 2021.
6. P. Sonkiya, V. Bajpai and A. Bansal, "Stock price prediction using BERT and GAN," arXiv:2107.09055, 2021.

7. P. Harnmetta and T. Samanchuen, "Sentiment Analysis of Thai Stock Reviews Using Transformer Models," 19th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2022.
8. P. Netisopakul and W. Saewong, "Thai stock news classification based on price changes and sentiments," Int. J. Electronic Finance, vol. 11, 2022.
9. A. Raheman, "Social Media Sentiment Analysis for Cryptocurrency Market Prediction," Computation and Language (cs.CL); Machine Learning (cs.LG); Social and Information Networks (cs.SI), p. arXiv:2204.10185, 2022.
10. B. Fazlija and P. Harder, "Using Financial News Sentiment for Stock Price Direction Prediction," Mathematics, 2022.
11. M. Sindhu, T. Sabareeswari and A. Tamilselvi, "Twitter Sentiment Analysis and Prediction using NLP," 2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), pp. 1-6, 2023.
12. S. Sangsavate, S. Sinthupinyo and A. Chandrachai, "Experiments of Supervised Learning and Semi-Supervised Learning in Thai Financial News Sentiment: A Comparative Study," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 22, no. 7, pp. 1-36, 2023.
13. N. Khamphakdee and P. Seresangtakul, "An Efficient Deep Learning for Thai Sentiment Analysis," Data, vol. 90, p. 8(5), 2023.
14. C. Jitboonyapinit, P. Maneerat and N. Chirawichitchai, "Sentiment Analysis on Thai Social Media Using Convolutional Neural Networks and Long Short-Term Memory," int. sci. j eng. tech., vol. 7, pp. 74-80, 2023.
15. G. Fatouros, J. Soldatos, K. Kouroumali, G Makridis, and D. Kyriazis, "Transforming sentiment analysis in the financial domain with ChatGPT," Machine Learning with Applications, vol. 14, p. 100508, 2023.
16. K. Kheiri and H. Karimi, "SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning," arXiv, 2023.
17. M. Leippold, "Sentiment spin: Attacking financial sentiment with GPT-3," Finance Research Letters, vol. 103957, p. 55, 2023.
18. B. Zhang, H. Yang, T. Zhou, A. Babar and X. Liu, "Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models," Computation and Language, 2023.
19. M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos and P. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," Expert Systems with Applications, vol. 223, p. 119862, 2023.
20. P. Isaranontakul and W. Kreesuradej, "A Study of Using GPT-3 to Generate a Thai Sentiment Analysis of COVID-19 Tweets Dataset," 2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 106-111, 2023.