JSAT
Journal of Science and Agricultural Technology
Research Article
https://doi.org/10.14456/jsat.2025.4
e-ISSN 2730-1532, ISSN 2730-1524

The use of artificial intelligence methods to predict income and expense trends from semi-unstructured data for accounting planning for pineapple farmers in Phitsanulok province

Wiraiwan Sanchana[1], Wanida Junsri[2], Pornpimol Tawee[2], Wijitra Koonkum[2], Tassanee Muenwicha[2], Pensri Phu-uthai[2], and Pramote Sittijuk[3*]

[1]Major in Agricultural Technology, Faculty of Science and Technology, Rajamangala University of Technology Lanna Phitsanulok, Phitsanulok 65000, Thailand
[2]Major in Accounting, Faculty of Business and Accounting, Phitsanulok University, Phitsanulok, 65000, Thailand
[3]Major in Information Technology, Faculty of Business and Accounting, Phitsanulok University, Phitsanulok, 65000, Thailand

***Corresponding author**: Pramotes@plu.ac.th*

Received: January 13, 2025. Revised: June 17, 2025. Accepted: June 23, 2025.

ABSTRACT

This study examined the use of artificial intelligence (AI), specifically Natural Language Processing (NLP), to predict income and expense trends for pineapple farmers in Ban Yang Subdistrict, Nakhon Thai District, Phitsanulok Province. The research focused on processing semi-unstructured accounting data, mainly from PDF files, provided by 30 farmers with prior accounting experience. A custom NLP algorithm was used to classify financial records into payment and income categories. Three time series forecasting models—Prophet, LSTM, and ARIMA—were applied to comparatively predict future trends. LSTM excels at capturing complex long-term patterns, Prophet effectively models seasonal and event-driven fluctuations, and ARIMA is well suited for identifying linear trends and short-term changes. The results showed that ARIMA outperformed both LSTM and Prophet in terms of accuracy and explanatory power. ARIMA achieved the lowest Mean Absolute Error (MAE) of 34.1084 and the highest R-squared (R^2) of 0.9901, indicating superior prediction performance. LSTM had a MAE of 71.0920 and an R^2 of 0.9511, showing good accuracy but with higher MAE and lower R^2 compared to ARIMA. Prophet had the highest MAE of 603.8044 and the lowest R^2 of -3.2273, reflecting poor performance. Based on these results, ARIMA is identified as the most suitable model for this dataset. ARIMA's performance improved with longer forecast periods. For a 10-day forecast, it showed relatively low accuracy. As the forecast period extended to 20 and 30 days, accuracy and explanatory power increased, with the best results observed for the 30-day forecast. These findings suggest that ARIMA performs better for longer-term predictions. Future work could optimize the model and incorporate additional features for improved accuracy.

Keywords: artificial intelligence, income and expense accounting trend, semi-unstructured data; pineapple farmer

INTRODUCTION

Accounting for income and expenses is crucial for managing finances in farming, especially for pineapple farmers in Phitsanulok Province, where weather conditions and fluctuating prices impact their earnings. By tracking income and expenses, such as costs for fertilizers, seeds, and labor, farmers can assess their operations and plan their finances more effectively. However, some farmers may lack the necessary knowledge or tools to manage financial data efficiently. External factors, such as market fluctuations or weather conditions, can also affect the accuracy of income and expense forecasts. Therefore, adopting accounting practices in agriculture requires a simplified approach that emphasizes the importance of farm records for evaluating performance and supporting decision-making (Singh, 2024). Despite the challenges, accounting remains a valuable tool for managing finances efficiently and mitigating risks from uncontrollable factors.

Household accounting for farmers has been continuously promoted, as good financial management is an important factor affecting the economic stability of farming households. The recording of income and expenses in the form of unstructured data and semi-unstructured data refer to data that is not organized or structured clearly, making it challenging to manage and utilize the information (Tredinnick and Laybats, 2024). This is in contrast to structured data, which is stored in standardized formats, such as tables or databases with clearly defined fields, making management and analysis easier (Bhatt et al.,2024). When data is stored in an unstructured form, such as handwritten notes or documents without proper organization, finding the necessary information can become difficult

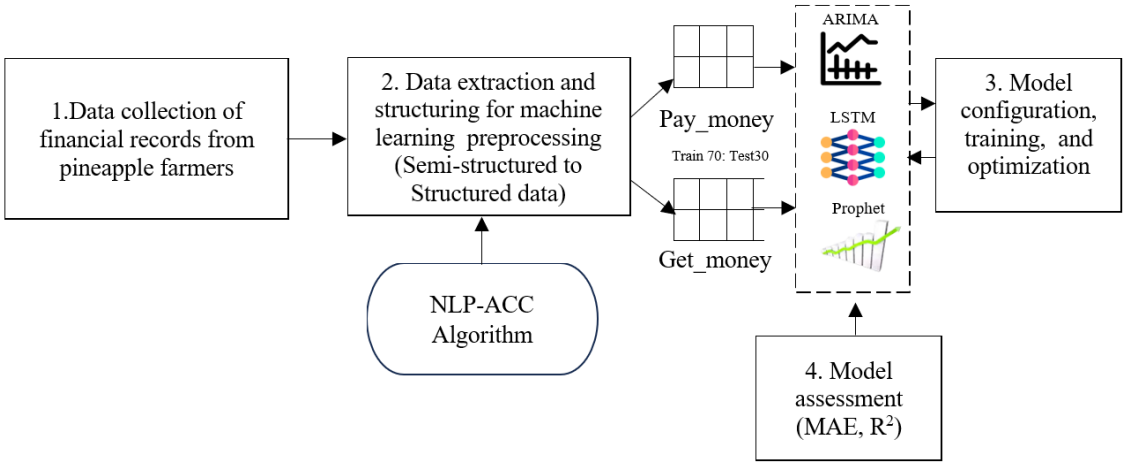

Figure 1. Research methodology steps diagram

and time-consuming. For example, if someone wants to know past expenses or look up income from product sales in different seasons, semi-unstructured data can make it impossible to retrieve the information quickly, requiring significant time to compile and analyze the results. If the data can be organized using artificial intelligence-related analytical techniques (Cao et al.,2024), it can be categorized into clear sections, such as income from product sales, expenses for materials and inputs, or other costs, making data management and analysis more efficient. Moreover, having organized data helps farmers track their financial status more effectively and accurately evaluate the performance of their agricultural operations.

At present, research works have developed various techniques for extracting semi-unstructured data and transforming it into structured data, which is highly significant in an era where data is abundant and complex. Artificial Intelligence (AI) technology has become a key tool in enhancing the efficiency and accuracy of managing this type of data. One of the critical techniques is Natural Language Processing (NLP), which enables computers to analyze text in various formats, such as documents, emails, or social media posts. NLP enables the extraction of key information, such as names, dates, and keywords, from textual data, organizing it into clear structures. It also helps analyze sentiment trends on social media, providing an automated approach to knowledge discovery and information extraction (VanGessel, 2024). Another prominent technique is Computer Vision, which is used to process image or video data. AI in this area can detect objects in images, read text using Optical Character Recognition (OCR), and convert image data into a format suitable for further analysis, such as digitizing information from image-based documents. Additionally, machine learning models are used in processes after transforming semi-unstructured data into structured data for clustering, classification, and prediction of data trends. These models are essential for extracting valuable insights from large-scale or semi-unstructured data to forecast sales, stock prices, or changes in a business's financial status. Predictive models analyze historical data to identify patterns, which they then use to predict future events (Dhawas et al.,2024).

This paper discusses the use of Natural Language Processing (NLP) techniques to organize and understand semi-unstructured financial data by matching expected accounting keywords. It focuses on identifying and categorizing key terms in financial documents, such as revenue and income statements. Additionally, predictive models, such as linear regression and time series algorithms, are used to forecast future financial outcomes based on historical data. This approach aims to improve traditional manual accounting methods, particularly for pineapple farmers in Phitsanulok Province, by automating the categorization and analysis of financial data, ultimately saving time and improving the tracking of income, expenses, and profits.

MATERIALS AND METHODS

From Figure 1, This research methodology began with the collection of financial data from pineapple farmers, including income, expenses, and related accounting information. The data are transformed from semi-structured formats into well-structured data using the NLP-ACC algorithm, making them suitable for machine learning preprocessing. Next, the LSTM model was configured, trained, and optimized, as it is effective in capturing complex and long-term data patterns. The ARIMA model focuses on analyzing linear trends and short-term fluctuations, while the Prophet model captures seasonal patterns and special events. The data were split into 70% for training and 30% for testing to accurately evaluate the models' performance. Finally, the results from each model were assessed using metrics such as MAE and R^2 to measure forecasting accuracy and suitability.

Data collection for accounting management methods of pineapple farmers

This research aimed to investigate the accounting practices and use of accounting formulas among 30 pineapple farmers in Ban Yang Subdistrict, Nakhon Thai District, Phitsanulok Province. The participants were selected using purposive sampling, focusing on farmers with prior experience in accounting. The data collection focused on how these farmers document their income and expenses, calculate production costs, and apply accounting formulas to assess their profits and losses. The data collection was conducted from October to November 2024.

The results of the data collection can be summarized by stating that most farmers believe accounting plays a vital role for pineapple farmers, enabling them to effectively manage and control their finances and production activities. The process begins with the preparation of income and expense accounts, which involves recording data on revenue from selling fresh produce and processed products, as well as expenditures such as fertilizer costs, pesticide costs, and labor expenses. Detailed record-keeping allows farmers to accurately track their expenses and net profits.

Additionally, categorizing production costs into fixed and variable costs helps farmers better understand the structure of their expenses. This classification also facilitates cost-per-unit calculations, which can be used to analyze profit and loss for each production cycle. Recording assets such as land and agricultural tools, along with liabilities like loans or other financial obligations, is another critical step in resource management.

Financial data analysis, such as calculating net profit and liquidity, enables farmers to plan their long-term finances more effectively. Moreover, leveraging technology, such as accounting applications or Excel programs, adds convenience and accuracy to data recording and numerical calculations using the common accounting formulas outlined in Equations 1–3.

$$\text{Total Income} = \text{Sum of all incomes} \quad (1)$$

$$\text{Total Expenditure} = \text{Sum of all expenditures} \quad (2)$$

$$\text{Net Income or Loss} = \text{Total Income} - \text{Total Expenditure} \quad (3)$$

From the study of accounting practices among farmers, it was found that fewer farmers recorded accounting data for revenue and payment summarization in the initial stages. They used an accounting form created by the Cooperative Audit Department (2009) as a structured data format, which was then input into Excel software. The summarized accounting data of the 30 farmers is presented as an average structured accounting record in Table 1. Additionally, some farmers used semi-unstructured data formats, such as text files on smartphone applications and Microsoft Word documents. While these methods are more convenient for the farmers, they lack a standardized structure. This reliance on informal formats reflects the farmers' limited accounting proficiency and the lack of user-friendly tools tailored to their needs. As a result, converting such unstructured data (Figure 2) into a format suitable for machine learning forecasting poses a significant challenge.

An example of semi-unstructured data records from farmers, who recorded data in text files on smartphone applications and document files in Microsoft Word, poses a challenge for structuring the data in the machine learning prediction process, as illustrated in Figure 2. From Figure 2, it is shown as a prototype of the farmers' accounting records in the form of semi-unstructured data in text files. This presents a challenge in building an understanding and tracking the data, such as lists of dates, income and expenditure types, and amounts. For this data collection process, 150 semi-unstructured accounting data samples can be collected from 30 farmers for testing the performance of data extraction using NLP techniques and creating a dataset for developing predictive models using machine learning algorithms. However, when this semi-unstructured data is transformed into structured data, it can be enhanced to develop accounting trends for farmers, which can be used in artificial intelligence models to predict future agricultural accounting risk states.

Using natural language processing (NLP) techniques to understand semi-unstructured accounting data

Review of NLP techniques for understanding semi-unstructured data from related research works

In the field of Natural Language Processing (NLP), several new techniques have been developed to improve the efficiency of analyzing and handling text data. One of the key techniques is word embedding proposed by Mikolov et al. (2013) , which transforms words or phrases into mathematical vectors that capture the meaning of the words in context. Models like Word2Vec or GloVe are used, helping words with similar meanings to be located close together in vector space. For example, the words "expense" and "payment" will have similar vector values. This technique allows NLP models to better understand the relationships between words in a deeper and more efficient way.

Table 1. Average structured accounting data recording of 30 farmers

Date	Average income (Baht)	Average expenditure (Baht)
Feb 2, 2024	2,750 (Carry-forward cash balance)	-
Mar 3, 2024	-	250 (Household expenses)
Mar 5, 2024	2,000 (Farm wages)	400 (Household expenses)
Mar 8, 2024	1,500 (Mushroom nursery wages)	500 (Kitchen supplies)
Mar 10, 2024	20,000 (Loan)	7,000 (Agricultural equipment)
Mar 12, 2024	-	3,500 (Fertilizers)
Mar 15, 2024	30,000 (Loan)	20,000 (Bank deposit)
Mar 15, 2024	-	2,400 (Fertilizers and pesticides)
Mar 20, 2024	-	8,000 (Soil preparation)
Mar 22, 2024	5,000 (Pineapple sales)	700 (Transport costs)
Mar 25, 2024	-	100 (Medical expenses)
Mar 28, 2024	-	

"On March 3, 2024, the farmer paid 200 Baht for food, gave 300 Baht to their child, and paid 100 Baht for electricity and water bills. On March 8, 2010, the farmer received 1,500 Baht for wages from building a mushroom nursery, paid 300 Baht for food, and spent 500 Baht on kitchen supplies.

On March 15, 2024, the farmer borrowed 30,000 Baht and deposited 20,000 Baht into the bank, spent 2,400 Baht on fertilizers and pesticides, paid 100 Baht for car fuel, and spent another 300 Baht on food.

On March 20, 2024, the farmer paid 8,000 Baht for plowing soil preparation and 200 Baht for food. On March 25, 2010, the farmer went to see a doctor, spending 100 Baht for medical expenses and another 200 Baht for food. On March 28, 2010, the farmer gave 300 Baht to their child."

Figure 2. Semi-unstructured accounting data recording of farmers**.**

Another widely used technique is sentiment analysis proposed by Pang and Lee (2008), which aims to analyze the opinions or emotions expressed in text. This technique is particularly useful for evaluating feedback on products or services. For instance, the text "This product is great" would be classified as a positive sentiment, while "The product is damaged" would be classified as negative. Sentiment analysis enables a better understanding of the emotions and opinions of people online effectively.

Text summarization is another important technique proposed by Nallapati et al. (2017). that helps create concise summaries from large volumes of information, such as summarizing news articles or financial reports. By using models like Sequence-to-Sequence or Transformer, it can produce abstractive summaries (using new words), making the content easier to understand and quicker to process without having to read all the data.

Topic modeling is a technique used to identify hidden topics in large datasets of text. This method proposed by Blei and Lafferty (2007), does not require prior input and is used, for example, to classify financial transactions according to the type of product or service. Models such as Latent Dirichlet Allocation (LDA) or Correlated Topic Model (CTM) can help find relevant topics from vast amounts of data efficiently.

Lastly, Named Entity Recognition (NER) proposed by Finkel et al. (2005), is a technique used to identify important information within text, such as names, places, dates, or monetary amounts. For example, in the sentence "Paid 500 baht for kitchen equipment on March 8, 2010," the model can identify "500 baht" and "March 8, 2010" as key entities. This technique helps extract specific, meaningful data from text more accurately.

These five techniques are crucial for developing and enhancing the performance of NLP, helping to analyze and understand text data with greater precision and efficiency.

Development of NLP techniques for understanding semi-unstructured accounting data

The operation of this code involves the collaboration of multiple libraries, which enables efficient extraction of data from PDF files, translation, word similarity checking, and saving data into CSV files. The process is classified into 4 steps as follows:

Opening and reading the PDF file

The pypdf library is used to open the PDF file named 'doc3.pdf' with the command PdfReader('doc3.pdf') to read the data from the file. The code checks the total number of pages in the PDF using len(reader.pages), which tells the total number of pages in the file. Then, the code selects the desired page using reader.pages[i], which extracts the text from the chosen page and stores it in the text variable. This text is then split into words using text.split(), which breaks the text into individual words and stores them in the list selected_text_split. The content from the PDF page is stored in the dictionary datax, with the page number as the key.

Language translation

In this step, the googletrans library is used, which is a tool for language translation. The translator.translate() function is used to translate words from Thai into English. This translation is important for converting Thai data in the PDF into English for easier processing. The translated word is stored in the translated_previous_value variable for use in the next step.

Checking word similarity

After translation, the code uses the difflib library and the get_close_matches() function to compare the translated word with words in a predefined dictionary. This dictionary contains words related to payment (e.g., "pay", "payment", "settle") and income (e.g., "receive", "income", "profit"). The get_close_matches() function checks how similar the translated word is to the words in the dictionary, considering a similarity threshold of 70% (using the cutoff=0.7 parameter). If the translated word is sufficiently similar to a word in the dictionary, the data is categorized into appropriate lists (e.g., pay_money or get_money). In this process, difflib.get_close_matches function from Python's Standard Library identifies close matches for a given word within a list using the SequenceMatcher algorithm. It compares similarity ratios to provide a ranked list of matches that meet or exceed a specified threshold, as shown in Equation 4.

$$Similarity\ Raio = \frac{2.M}{T_a + T_b} \quad (4)$$

From Equation 4, M is Number of matching elements, T_a is Total elements in the first sequence and T_b is Total elements in the second sequence.

Saving data into CSV files

After translating the words and categorizing the data, the code uses the csv library to save the data into CSV files. The csv.writer() function is used to write the data into the file. The code opens a new CSV file and uses writer.writerow() to write the header of the file and writer.writerows() to write the data in each row. The data saved includes both the translated text and the information related to payments and income, which is stored in separate CSV files based on category (e.g., corrected_data.csv, pay_money.csv, and get_money.csv). After the structured accounting data is stored in the .CSV files, the data (pay_money.csv and get_money.csv) will be processed using the accounting formulas provided in Equations 1–3 with time series algorithms for predicting the future accounting status. The algorithm is outlined in the procedure shown in Table 2.

Developing predictive accounting trend models using machine learning algorithms

After extracting the semi-unstructured accounting data recorded by farmers, it is categorized into two accounting types: pay_money and get_money, using the developed algorithm shown in Table 2. The categorized data in the CSV file is then organized into lists of payments and receipts, grouped by date, to create a suitable dataset for processing with time series algorithms, including Prophet, LSTM, and ARIMA. The structure of the categorized data is shown below.

Table 1: pay_money

Date	Description	Amount
2025-01-01	Purchased fertilizer	500
2025-01-02	Paid for labor	700
2025-01-04	Bought seeds	300

Figure 3. Pay money data table.

Table 2: get_money

Date	Description	Amount
2025-01-01	Sold rice	1,200
2025-01-03	Sold vegetables	800
2025-01-05	Received government aid	1,000

Figure 4. Get_money data table.

In processing the accounting data recorded by farmers in CSV files, following the data structures—including dates and related amounts shown in Figures 3 and 4, the data are imported using Python's Pandas library. This process prepares the data into a suitable format for management and analysis by converting the date columns into datetime types, enabling various models to accurately process the information.

To compare the performance of time series modeling and linear regression modeling systematically, we can break the process into three main stages:

Data preparation

The data preparation for time series modeling starts by organizing the data in chronological order, ensuring that the data is sorted by date and filling in any missing values. Then, the data is split into 70% for the training set and 30% for the testing set to preserve the chronological order and prevent data leakage, where future data would be used to predict the past.

Model configuration

The time series algorithms, including Prophet, LSTM, and ARIMA were defined with the following model configuration.

Facebook prophet

Prophet is based on an additive time series decomposition approach. Prophet is optimized for large datasets with strong seasonal patterns and is designed to handle seasonality and holidays. The model adjusts trends, seasonality, and events using either an additive or multiplicative approach. Its accuracy can be improved by fine-tuning parameters related to seasonality, holidays, and changepoints based on the data characteristics.

LSTM (Long Short-Term Memory)

LSTM is a type of recurrent neural network (RNN) that is highly effective at handling sequential data and capturing long-term dependencies in time series.

Optimized method: To enhance LSTM performance, several hyperparameters can be tuned, including the number of layers, units per layer, learning rate, batch size, and sequence length for the input data. The process begins by scaling the data using MinMaxScaler to normalize it within the range of 0 to 1. A 5-point lag sequence is then created to predict the next value in the series. The data is reshaped into a 3D format, suitable for LSTM processing, where the input data structure follows the shape (samples, time steps, features). The LSTM model consists of two layers, each with 50 units, and uses the 'tanh' activation function. To reduce the risk of overfitting, two Dropout layers with a rate of 0.2 are included. Finally, a Dense layer is used to output the predicted value. This optimization approach helps ensure accurate predictions while minimizing overfitting.

ARIMA (Autoregressive Integrated Moving Average)

ARIMA is a statistical model used for time series forecasting, consisting of three main components: 1) AR (Autoregressive): This component utilizes the relationship between the current value and past values of the data to predict future values. For example, if today's result is related to the result of the previous day, this relationship will be used to forecast the next value, 2) I (Integrated): This component is used to make the data stationary or reduce its randomness by differencing the data over time. Differencing helps eliminate trends and make the data more stable over time, and 3) MA (Moving Average): This component uses a moving average of past prediction errors or residuals to forecast future values. This helps the model handle volatility in the data by averaging out past errors.

The auto Arima function automatically selects the best ARIMA model by testing different combinations of autoregressive (AR), moving average (MA), and differencing parameters, and choosing the model with the lowest Akaike Information Criterion (AIC). It uses seasonal=False to exclude seasonality and stepwise=True for efficient model selection. Warnings are suppressed with suppress_warnings=True, and errors are ignored during training with error_action="ignore". The model is then trained using arima_model.fit(train['y']), and predictions are made for the test period with arima_model_fit.predict(n_periods=len(test)), comparing the forecasted values with actual data.

Experiment Method

Performance assessment of the NLP-ACC algorithm

The performance of the NLP-ACC algorithm is assessed by examining its capability to process 150 semi-unstructured accounting data samples. The evaluation is based on the percentage of matches, which shows how accurately the algorithm detects relevant data in comparison to the total number of expected elements.

Performance assessment of predictive accounting trend models

The extracted payment accounting dataset from semi-unstructured data is used to assess the performance of time series algorithms. This dataset consists of 365 samples, split into 256 training samples (70%) and 109 testing samples (30%) for building and evaluating the predictive model. Additionally, the best model's performance in predicting data for 10, 20, and 30 days is tested. To assess the predictive model's performance, we use two main metrics:

R^2: Measures the ability of the model to explain the variance in the data. A high R^2 means the model explains the data well, but it does not always mean the model will predict well, as it may be due to overfitting, as shown in Equation 8.

$$R^2 = 1 - \frac{RSS}{TSS} \quad (8)$$

MAE (Mean Absolute Error) : a metric used to evaluate the accuracy of a forecasting model. It is

Table 2. Pseudocode for developing NLP techniques to understand semi-unstructured accounting data

Algorithm: NLP-Based Approach for Semi-Unstructured Accounting Data (NLP-ACC)

Step	Action
1	Initialize the PDF reader using PdfReader('doc3.pdf')
2	Get the total number of pages in the PDF using len(reader.pages)
3	Select the desired page by iterating through reader.pages[i]
4	Extract the text from the selected page using text = page.extract_text()
5	Split the extracted text into words using text.split()
6	Store the split words in the list selected_text_split
7	Store the content from the page in the dictionary datax using the page number as the key
8	Initialize the translator using Translator()
9	For each word in selected_text_split, translate it using translator.translate(word, src='th', dest='en'):
10	{
11	Store the translated word in the translated_previous_value variable
12	}
13	Define two lists: pay_money (e.g., "pay", "payment", "settle") and get_money (e.g., "receive", "income", "profit")
14	For each translated word (translated_previous_value), compare it to the dictionary using get_close_matches():
15	{
16	Set the similarity threshold to 70% (cutoff=0.7)
17	If the translated word is similar to a word in the dictionary, categorize the data:
18	{
19	If related to payment:
20	{
21	Add it to pay money
22	}Else {
23	Add it to get_money
24	}
25	}
26	Open a CSV file for each category (corrected_data.csv, pay_money.csv, get_money.csv)
27	Use csv.writer() to create a CSV writer object
28	Write the header in the CSV file using writer.writerow()
29	Write the categorized data into the file using writer.writerows()

calculated by determining the difference between the predicted values and the actual values at each data point and then finding the average of these differences., as shown in Equation 9.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{y}_i) \quad (9)$$

RESULTS AND DISCUSSION

Performance Assessment Results of the NLP-ACC Algorithm

The performance evaluation of the NLP-ACC algorithm in detecting the collected semi-unstructured accounting data from the farmers is as follows.

From Table 3, the NLP-ACC algorithm can process the accounting data recorded by farmers effectively, with an average matching percentage of 90.66%. This indicates that the algorithm performs well in detecting the expected data. Although some data categories have lower matching percentages, the overall results demonstrate the algorithm's strong performance in detecting information from semi-unstructured accounting records.

Performance assessment results of predictive accounting trend models

The extracted payment accounting dataset from semi-unstructured data is used to comparatively assess time series algorithms. This dataset consists of 365 samples, split into 256 training samples (70%) and 109 testing samples (30%) for building and evaluating the predictive model's performance. In assessing the predictive model's performance, we use two main metrics:

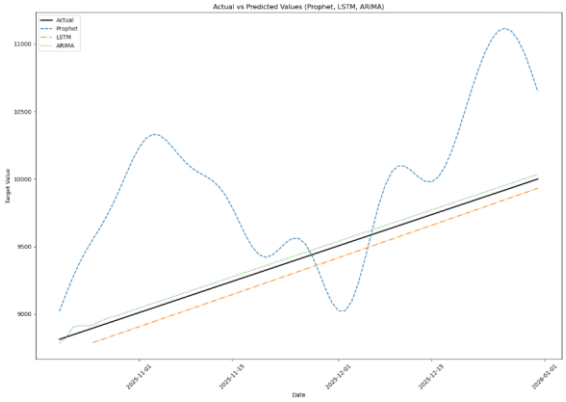

Figure 5. Comparison of predicted and actual values at the date points of the time series algorithms.

Table 3: Performance of the NLP-ACC algorithm in detecting collected semi-unstructured accounting data recordings

Expected Data	Collected Semi-Unstructured Accounting Data Recordings (Number of Detected)	Extracted Data Using NLP-ACC Algorithm (Number of Detected)	Match (%)
Food payment	126	120	95.24
Give money	149	130	87.25
Car fuel	185	150	81.08
Medical expenses	141	135	95.74
Kitchen supplies	157	140	89.17
Tap water	161	150	93.17
Wages	176	165	93.75
Loan	154	130	84.42
Deposit	130	120	92.31
Fertilizers/Pesticides	149	140	93.96
Plowing	159	145	91.19
Average Summarization	**153.36**	**138.64**	**90.66**

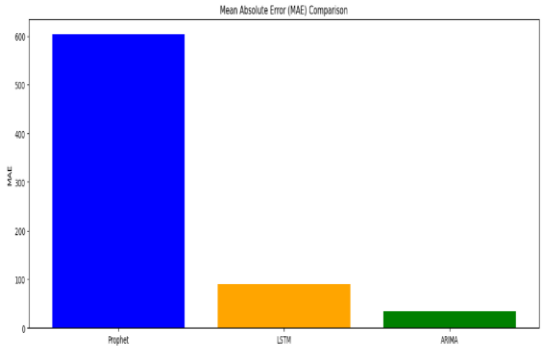

Figure 6. Comparison of MAE of the algorithms.

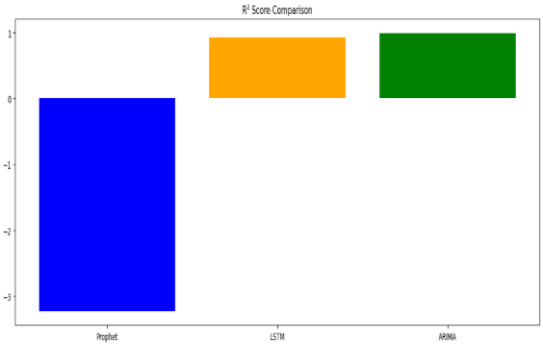

Figure 7. Comparison of R^2 of the algorithms.

Figure 5 – 7, the comparison of prediction results from the Prophet, LSTM, and ARIMA models applied to the payment accounting dataset clearly demonstrates varying performance levels, with Mean Absolute Error (MAE) and R-squared (R^2) used to assess each model's prediction accuracy and ability to explain the data.

MAE, or Mean Absolute Error, indicates the average prediction error of each model. In this case, the ARIMA model has the lowest MAE of 34.1084, showing the highest prediction accuracy as it predicts values closest to the actual data. LSTM has a MAE of 71.0920, indicating good accuracy but with slightly more error than ARIMA. Prophet has the highest MAE of 603.8044, reflecting the greatest prediction error, as its predictions deviate most from the actual values.

Another metric, R-squared (R^2), measures the model's ability to explain data variance. A value closer to 1 indicates a better explanation of the data, while a negative value suggests poor explanatory power. ARIMA has an R^2 of 0.9901, showing that it captures the data characteristics very well and explains the data variance effectively. LSTM has an R^2 of 0.9511, still demonstrating good explanatory power but slightly lower than ARIMA. Prophet has an R^2 of -3.2273, which is a negative value, indicating that it cannot explain the data at all and fails to capture its variance.

From the comparison of both MAE and R^2, it can be concluded that ARIMA is the most effective model for this dataset. Not only does it have the highest accuracy, but it also explains the data well. While LSTM shows good prediction capability, its higher MAE and lower R^2 suggest that it still struggles in comparison to ARIMA. Prophet, on the other hand, shows both high prediction error and poor explanatory power, making it the least suitable model for this dataset. Therefore, ARIMA is the most appropriate model for this data at present.

The performance of ARIMA as the best model is tested for predicting data for 10, 20, and 30 days, as shown in the results in Figure 8.

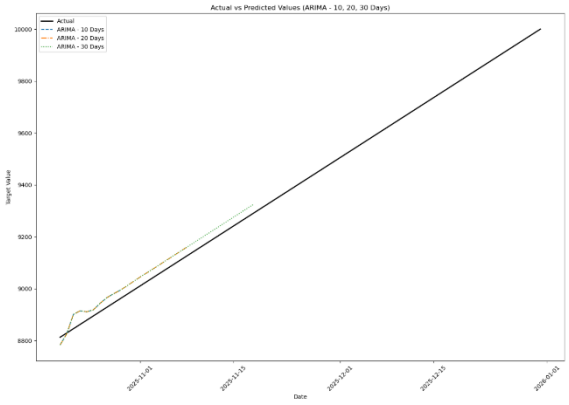

Figure 8. The performance of the ARIMA model in predicting data for different time periods.

From Figure 8, the performance of the ARIMA model in predicting data for 10, 20, and 30 days varies with the prediction period. For the 10-day forecast, the model exhibits a relatively low performance, with an MAE of 33.3129 and an R^2 of 0.4205, indicating limited accuracy and explanatory power.

However, as the forecast period extends to 20 days, the MAE increases slightly to 33.4816, but the R^2 improves to 0.8654, suggesting a better fit and more accurate predictions. The 30-day forecast shows a slight increase in MAE to 33.6254, but the R^2 continues to rise to 0.9414, reflecting the model's highest accuracy and explanatory power for this period. Overall, these results indicate that the ARIMA model performs better with longer prediction periods, particularly in terms of its explanatory capability, as reflected by the increasing R^2 values.

The experimental results show that the ARIMA model achieved the best performance because it was efficiently fine-tuned using the auto_arima function, which automatically selects the best parameters by testing various values based on the Akaike Information Criterion (AIC) (Ali and Masmoudi, 2025). This approach simplifies the parameter selection process and results in more accurate predictions compared to the LSTM and Prophet models. ARIMA exhibited the lowest error and highest accuracy, especially for long-term forecasting. Overall, ARIMA is the most suitable model for this dataset, with potential for further improvement through additional fine-tuning and feature enhancement.

CONCLUSIONS

This study applied AI, specifically Natural Language Processing (NLP), to predict income and expense trends for pineapple farmers in Phitsanulok Province using semi-unstructured accounting data. The NLP-ACC algorithm effectively detected the data, achieving a high matching rate. Time series forecasting models—Prophet, LSTM, and ARIMA—were compared, with ARIMA showing the best performance in terms of accuracy and explanatory power, outperforming both LSTM and Prophet. The performance of the ARIMA model varied with the prediction period. For shorter forecasts, the model showed lower accuracy, but as the forecast period increased, both the accuracy and explanatory power improved, with the best results seen for longer prediction periods. This suggests that ARIMA performs better for longer-term predictions.

ACKNOWLEDGMENTS

We would like to express our heartfelt gratitude to everyone who contributed to the success of this research. Firstly, our sincere thanks go to the pineapple farmers in Ban Yang Subdistrict, Nakhon Thai District, Phitsanulok Province, who generously provided their time, knowledge, and insights into the accounting practices and the use of accounting formulas in their operations. Their participation was vital in enriching the understanding of accounting in agricultural businesses. This research would not have been possible without the contributions of all these individuals. Thank you.

REFERENCES

Ali, A. A. and Masmoudi, A. 2025. Building prediction models for the e-government development index (EGDI) in Iraq and KSA: a comparative ARIMA-based approach. Fusion: Practice and Applications (FPA). 19(2): 134–150.

Bhatt, H. S., Ramakrishnan, S., Raja, S., and Jawahar, C. V. 2024. Unlocking the potential of unstructured data in business documents through document intelligence. In Proceeding of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD) (pp. 505-509).

Blei, D. M., and Lafferty, J. D. 2007. A correlated topic model of Science. The 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Cao, S.S., Jiang, W., Lei, L., and Zhou, Q. 2024. Applied AI for finance and accounting: Alternative data and opportunities. Pacific-Basin Finance Journal. 84: 102307.

Cooperative Audit Department. (2009). Accounting form for pineapple farmers. Retrieved from https://www.cad.go.th/ewtadmin/ewt/cadweb_org/download/e_learning/acc_2009_06.pdf.

Dhawas, P., Ramteke, M. A., Thakur, A., Polshetwar, P. V., Salunkhe, R. V., and Bhagat, D. 2024. Big data analysis techniques: Data preprocessing techniques, data mining techniques, machine learning algorithm, visualization. In Big data analytics techniques for market intelligence (pp. 183-208). IGI Global.

Finkel, J. R., Grenager, T., and Manning, C. D. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In Proceeding of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005).

Mikolov, T., Yih, W., and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In Proceeding of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2013).

Nallapati, R., Zhou, B., and Mahajan, D. 2017. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Proceeding of the 33rd International Conference on Machine Learning (ICML 2017).

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. 2(1-2): 1-135.

Singh, M. 2024. Accounting for a farm business: A conceptual study. Contemporary Social Sciences. 33(1): 66.

Tredinnick, L., and Laybats, C., 2024. Managing unstructured information. Business Information Review. 41(3): 90-93.

VanGessel, F. G., Perry, E., Mohan, S., Barham, O. M., and Cavolowsky, M. 2024. NLP for knowledge discovery and information extraction from energetics corpora. arXiv:2402.06964.