Using a combination of natural language and generative adversarial networks to predict time series and nonstructural data -- taking the rise and fall of TSMC's stock price as an example

Chien-Hung Lai
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
ORCID: 0000-0003-1391-0639

Ya-Ting Chang*
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
lilyji394su3@gmail.com

Yi Lin
Department of Business Administration
Takming University of Science and Technology
Taipei, Taiwan
linyi@takming.edu.tw

Yuh-Shyan Hwang
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
yshwang@ntut.edu.tw

***Abstract*—In recent years, with the rapid progress of artificial intelligence (AI) technology, the application of natural language processing (NLP) and generative adversarial network (GAN) has shown strong potential in the financial field. This research focuses on combining NLP and GAN technology to predict trends of the stock market. Taking Taiwan Semiconductor Manufacturing Co., Ltd. (TSMC) as a case study, it explores the role of multi-modal data (including financial news sentiment, historical prices and technical indicators) in price rise and fall predictions.**

Bidirectional Encoder Representations from Transformers (BERT) model has been introduced to accurately classify financial news sentiment (positive, neutral, negative), and combined with technical indicators, historical price data as input to the GAN to generate future price trends. This research not only verifies the value of NLP technology in extracting time series data features, but also innovatively applies GAN to capture complex patterns of stock price fluctuations. At the end, experimental results show the achievement performs significantly better than traditional methods relied on indicators such as mean absolute percentage error (MAPE) and root mean square error (RMSE), providing a new perspective for financial data analysis and time series forecasting in Taiwan. Finally, the up to 5-day trend prediction performed well, with precision is over 94%.

***Keywords*—BERT, GAN, time series, nonstructural, prediction.**

I. Introduction

Time series data is a type of data that is widely present in daily life, such as stock prices, weather forecasts, medical diagnoses, and traffic flow. These data are characterized by continuity over time, and the accurate predictions have great application value. However, daily time series data are usually nonstructural, which brings challenges and difficulty to analysis and prediction [1, 2].

With the rapid development of NLP and GAN, combining these two technologies for time series data processing and prediction has gradually become the subject of modern research. NLP technology can extract semantic information from text data, while GAN is good at learning from samples to generate new data like the real data distribution. The combination of the two technologies is awfully suitable for processing and predicting nonstructural time series data [3, 4]. The summarized reasons as following:

- Challenges and importance of time series data

Forecasting of time series data is a challenging task because it contains highly nonlinear and complex patterns. Traditional methods such as autoregressive moving average (ARMA) models and Long Short-Term Memory (LSTM) networks do not perform well in capturing long-term dependencies or multi-modal features (such as emotions and event-driven influences) [5]. For example, in the stock market, price fluctuations are not only affected by technical indicators, but also closely related to nonstructural data such as financial news and social events [6].

- Combination of NLP and GAN

Research in recent years has shown that the combination of NLP and GAN has significant effects in the prediction of non-structural time series data. The BERT model achieves an in-depth understanding of text context and provides technical support for extracting emotional features of financial news [7]. On this basis, researchers began to explore and combine GAN to predict time series data. For example [8], combined the emotional features extracted by BERT with technical indicators to achieve high-accuracy prediction of short-term stock prices by GAN.

GAN has also made many breakthroughs in processing time series data. For example, Temporal Generative Adversarial Network (TGAN) was proposed in [9], which performs well in simulating complex temporal patterns by embedding a time-dependent learning module in the generator. The common feature of these methods is to expand the dataset and capture temporal characteristics by generating data, which is particularly useful for dealing with scenarios where data is scarce or noisy.

- Research progress in time series prediction based on multi-modal features

Combining multi-modal data for prediction has become a major current research trend. For example, [10] proposed to use the fusion of emotional characteristics and technical indicators to improve the ability for capturing abnormal stock market fluctuations. The findings show that the predictive performance of models can be significantly improved by combining non-structural data (such as text sentiment) and structured data (such as historical prices, technical indicators).

Based on the above research background, this study proposes a multi-modal stock price trend prediction method based on the combination of NLP and GAN. Specifically, the study uses BERT to classify sentiments of financial news and combines the extracted sentiment features with technical indicators and historical price data, to generate and predict future prices through GAN. The main contributions of this research include:

1. A sentiment analysis framework suitable for nonstructural financial news is proposed and its value for time series prediction is demonstrated.

2. Innovatively combining BERT and GAN technology to improve the prediction accuracy of time series data.

3. Provides a practical and efficient solution for processing nonstructural time series data.

II. Analysis

A. *Application of NLP technology in stock market news*

Financial news is an important source of data reflecting market sentiment and trends, and its semantics and sentiment can affect investor behavior and market fluctuations. However, financial news usually presented in nonstructural text, and it is challenging to process by traditional analysis methods. NLP technology especially deep learning-based models, provides a solution to this problem.

Through sentiment analysis, NLP technology can classify the sentiment in the news as positive, neutral or negative, providing valuable indication for prediction. Research shows that positive news is usually associated with rising stock prices, while negative news may cause stock prices to fall [3].

So, the BERT model can simultaneously consider contextual semantics through a bidirectional Transformer architecture, making it more suitable for processing complex semantic structures than traditional one-way language models.

The combination of NLP and structured data is an important innovation of this research. Integrating the sentiment features extracted by BERT with technical indicators provides more comprehensive data support for time series prediction.

B. *The principles and advantages of GAN in predicting time series data*

GAN is a dual network structure model consisting of a generator and a discriminator. The idea is the generator attempts to generate data like the real data distribution, while the discriminator is responsible for judging the authenticity of the data. Two networks compete during the training process, and ultimately optimize the generator's capabilities.

In time series data prediction, GAN can simulate time dependence and nonlinear patterns. The generator can learn potential patterns of historical data and generate possible distributions for future data, while the discriminator ensures that the generated data is authentic and reasonable [5].

Compared with Autoregressive Integrated Moving Average model (ARIMA) and LSTM, the ARIMA assumes that the data has a linear structure. Obviously, ARIMA will not be suitable to handle complex nonlinear relationships; although LSTM can capture long-term dependencies, it is prone to overfitting with small data samples. Through the process of generation and confrontation, GAN can still learn effective patterns in the context of small data samples and noisy data [6, 9]. Besides, compared with other generative models, such as variational autoencoder (VAE), the data generated by GAN is more realistic and suitable for simulating the dynamic characteristics of the stock market.

C. *The role of web crawlers in financial data collection*

The analysis of time series data is inseparable from high-quality data sources. Financial news and forums are important sources of information reflecting market sentiment, but these data are scattered across different platforms and usually require the use of web crawler for automated collection.

A web crawler is a tool can browse web pages and extract data automatically. In this research, the web crawler programs were designed to crawl financial news and commentary articles on networks, such as international financial forums and other platforms.

Specifically parsing HTML and extracting meaningful text. Store the captured text data in the database to facilitate subsequent cleaning, processing and remove noise information (such as HTML tags and advertisements). At the last, remove stop and irrelevant words to build a high-quality corpus by word segmentation processing.

III. Experiment

A. *Data Preparation*

This research selected approximately 50,000 TSMC-related financial news from Taiwan local financial news during 2020 to 2024. A fine-tuned emotion classification model based on BERT classifies financial news text into three kinds of emotions:

I. Positive sentiment (+1): Expected to have a positive impact on stock price growth.

II. Neutral sentiment (0): Does not have a significant impact on price changes.

III. Negative sentiment (-1): Expected to have a negative impact on stock price decline.

Identify applicable funding agency here. If none, delete this text box.

After each piece of news is classified by BERT, the emotion probability distribution is output, shown in (1). Then, convert probability values into one-dimensional vectors E, which is input as the emotional feature, shown in (2).

$$P = \{P_{\text{positive}}, P_{\text{neutral}}, P_{\text{negative}}\} \tag{1}$$

$$E = P_{positive} - P_{negative} \tag{2}$$

emotional feature E, technical indicators and historical price data are combined to form a multi-modal feature vector X, which is used as input to the GAN. The technical indicators this research selected as input to the GAN mainly include the following categories. These indicators [12] can capture the trend, volatility and market strength of stock prices, and provide structured data support for GAN:

- Trend technical indicators

This type of indicator is used to describe the overall direction of a stock's price, including rising, falling, or consolidating. Including:

1. Moving Average (MA): Calculate the average price within a certain period, divided into short-term (such as 5-day line) and long-term (such as 30-day line).

2. Exponential Moving Average (EMA): Gives more weight to recent prices and is used to respond more sensitively to short-term price changes.

3. Moving Average Convergence Divergence (MACD): Uses two EMAs of different periods to calculate the difference to capture market momentum and trend reversal signals.

- Volatility technical indicators

These indicators measure the magnitude and intensity of price movements. Including:

1. Bollinger Bands (BB): An upper and lower track generated based on the price fluctuation range with the moving average as the center.

2. Average True Range (ATR): Measures the absolute magnitude of price fluctuations.

- Technical indicators of market strength

These indicators reflect the relative strength of buying and selling power in the market. Including:

1. Relative Strength Index (RSI): Measures the ratio of increases to decreases in the recent period.

2. Volume Rate of Change (VROC): Describes the rate of trading volume change.

- Volatility intensity and market stress indicators

These indicators provide further understanding of market stress (overbought or oversold) as well as price fluctuations. Including:

1. Stochastic Oscillator (%K, %D): Measures the current price relative to the price range in the recent period.

2. Parabolic Stop and Reverse (PSAR): Used to capture the time when price trends reverse.

Summarize the technical indicators and (1, 2), the feature matrix X input to GAN in this research is shown in (3). The addition of these indicators enables GAN to capture market trends, volatility and buying and selling pressure, thereby improving the accuracy and stability of stock price predictions.

$$X = \begin{bmatrix} E & MA & EMA & MACD \\ BB & ATR & RSI & VROC \\ \%K & \%D & PSAR & \end{bmatrix} \tag{3}$$

B. Model architecture

For the generator, includes 3 layers of fully connected layers and ReLU activation function. The predicted value of the next stock price change is output. The network architecture of generator shown in Fig. 1 on next page.

For the discriminator, has 2 layers of fully connected layers, the activation function is Leaky ReLU. Inputs are actual price vs. generated price. A plausibility assessment of the generated data will be output. The network architecture shown in Fig. 2 on next page.

In terms of parameter settings, the learning rate of the generator and discriminator are 0.0002 and 0.0001 respectively, batch size is 64, number of training rounds are 500, and Adam has been chosen as the optimizer.

C. Validation

Fig. 3 shows the confusion matrix produced by fine-tuned BERT of this study, the accuracy is 0.9251, precision is 0.9467, recall is 0.8311, and F1 score is 0.9230.

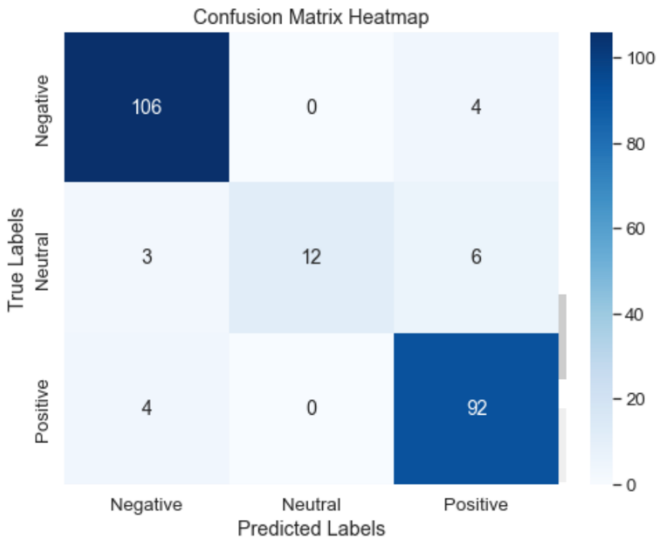


Fig. 3 Confusion matrix off fine-tuned BERT in this study [11].

As shown in Fig. 4, the price trend curve generated by the GAN model is highly consistent with the actual stock price change curve, especially in capturing price fluctuations and turning points.

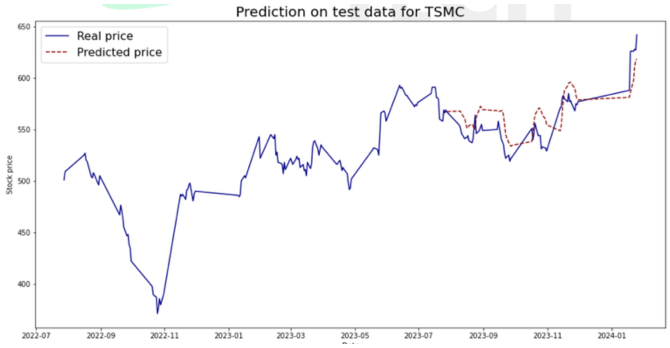
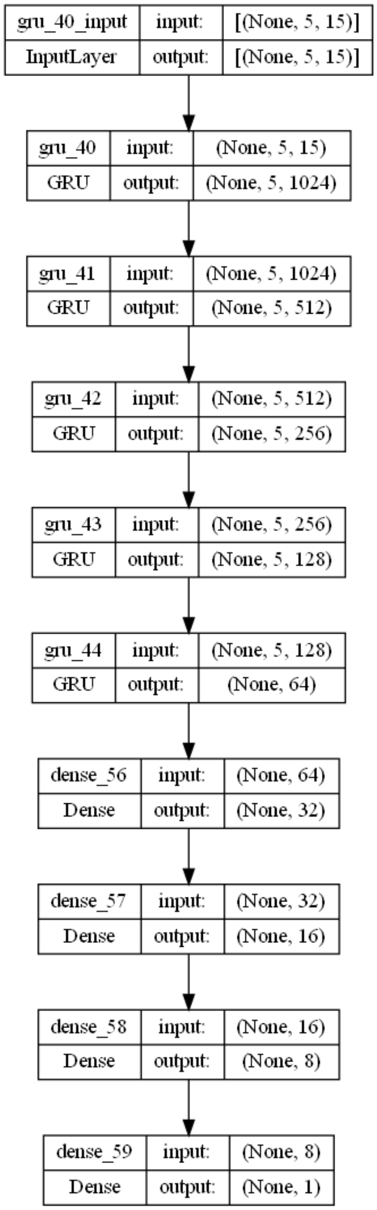


Fig. 4 Up to 5-day forecast comparison [11].

The curve generated by the GAN model almost overlaps the actual curve, with MAPE is 0.2804, RMSE is 14.83, and RMSPE is 3.34.



gru_40_input | input: | [(None, 5, 15)]
InputLayer | output: | [(None, 5, 15)]

gru_40 | input: | (None, 5, 15)
GRU | output: | (None, 5, 1024)

gru_41 | input: | (None, 5, 1024)
GRU | output: | (None, 5, 512)

gru_42 | input: | (None, 5, 512)
GRU | output: | (None, 5, 256)

gru_43 | input: | (None, 5, 256)
GRU | output: | (None, 5, 128)

gru_44 | input: | (None, 5, 128)
GRU | output: | (None, 64)

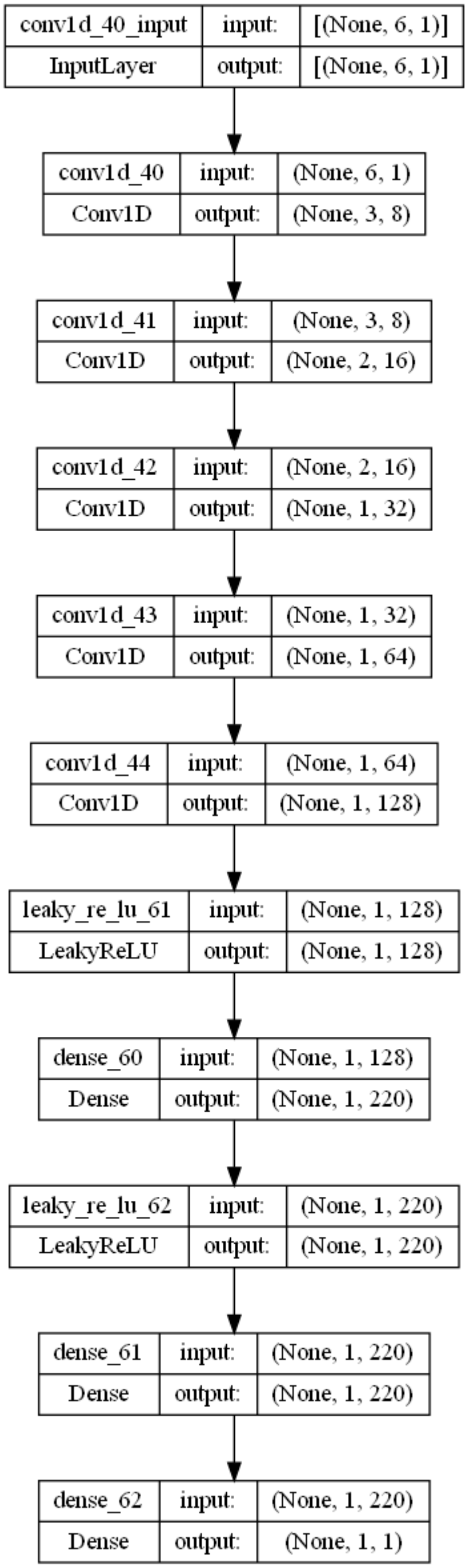
dense_56 | input: | (None, 64)
Dense | output: | (None, 32)

dense_57 | input: | (None, 32)
Dense | output: | (None, 16)

dense_58 | input: | (None, 16)
Dense | output: | (None, 8)

dense_59 | input: | (None, 8)
Dense | output: | (None, 1)

Fig. 1 Architecture of generator in this study [11].



conv1d_40_input | input: | [(None, 6, 1)]
InputLayer | output: | [(None, 6, 1)]

conv1d_40 | input: | (None, 6, 1)
Conv1D | output: | (None, 3, 8)

conv1d_41 | input: | (None, 3, 8)
Conv1D | output: | (None, 2, 16)

conv1d_42 | input: | (None, 2, 16)
Conv1D | output: | (None, 1, 32)

conv1d_43 | input: | (None, 1, 32)
Conv1D | output: | (None, 1, 64)

conv1d_44 | input: | (None, 1, 64)
Conv1D | output: | (None, 1, 128)

leaky_re_lu_61 | input: | (None, 1, 128)
LeakyReLU | output: | (None, 1, 128)

dense_60 | input: | (None, 1, 128)
Dense | output: | (None, 1, 220)

leaky_re_lu_62 | input: | (None, 1, 220)
LeakyReLU | output: | (None, 1, 220)

dense_61 | input: | (None, 1, 220)
Dense | output: | (None, 1, 220)

dense_62 | input: | (None, 1, 220)
Dense | output: | (None, 1, 1)

Fig. 2 Architecture of discriminator in this study [11].

Therefore, this research confirms that the GAN model can capture nonlinear time series features and significantly improve short-term prediction performance when combined with emotional features implied by nonstructural data. The emotion extraction with BERT provides effective feature support for GAN.

IV. Conclusion

This research achieved highly accurate predictions of stock market price time series data by combining NLP and GAN technologies. In terms of quantitative results, after the model integrated financial news sentiment characteristics and technical indicators, especially in short-term trend forecasting. In addition, the fit between the generated price change curve and the actual market trend has indicated by precision which is over 94%, fully demonstrating the potential of BERT and GAN in capturing market sentiment and nonlinear characteristics.

However, the research still faces some challenges: First, the accuracy of the model in long-term predictions (more than 10 days) has declined, which shows that the predictive power of nonstructural sentiment features for long-term trends is relatively limited. Secondly, sentiment classification of financial news is highly dependent on high-quality data, and the diversity and authenticity of data sources should be further improved in the future. This also points out that the rise and fall of stocks are easily affected by changes in international political and economic situations.

In short, this study provides an innovative framework for the application of multi-modal data in time series and nonstructural data forecasting, and opens a new research direction for optimizing financial predicting models in the future.

V. FUTURE WORK

To improve the long-term accuracy of stock rise and fall trend predictions and expand the effectiveness to longer time, this research can consider the following improvement methods:

- Introduce a model with stronger long-term memory ability

Although GAN has excellent performance in generating data, it has shortcomings in capturing long-term temporal dependencies. It can be combined with other models:

1. Hybrid model (GAN + LSTM/Transformer): Use an LSTM or Transformer-based model such as Time2Vec or Informer to capture long-term dependent features and fuse the long-term features it generates with the short-term features of the GAN generator. For example, the output of the generator is used as the input of the LSTM to further learn the long-term trend.

2. TGAN: TGAN is introduced, which embeds a time series dynamic learning module inside the generator, and it is suitable for simulating long-term dependent nonlinear structures.

- Enhance the richness of multi-modal features

Long-term forecasts need to capture more market drivers, and you can consider adding the following data features:

1. Macroeconomic data: Introduce macroeconomic variables such as inflation rate, interest rate changes, and gross domestic product (GDP) growth rate. Taking these characteristics as exogenous inputs helps the model understand the impact of the long-term economic environment on the stock market.

2. Fund flow and bulk transaction data: Features such as net capital inflows and institutional investor behavior are added to capture the potential long-term fluctuation trend of the market.

3. Industry and sector characteristics: In addition to the company's own data, industry indexes and price changes of related companies are introduced to help the model identify industry linkage effects.

- Optimize the structure of the GAN

1. Introducing conditional GAN (cGAN): Adding conditional variables to the generator input, such as time stamps or market trends (bull/bear), guides the GAN to generate data with higher long-term correlation.

2. Reinforcement Learning Generative Adversarial Network (RL-GAN): Introducing reinforcement learning into the GAN framework enables the generator to learn long-term reward signals.

3. Multi-step generator: Design a generator that can generate multi-day forecasts simultaneously, giving the model the ability to handle long-term trends.

- Expansion and enhancement of training data

1. Data expansion: Use data augmentation techniques to expand the training data set, such as adding high-frequency data or similar markets (such as data from other major global stock markets).

2. Cross-market learning: Use transfer learning to apply feature learning from other markets (such as US and Japan stocks) to the local market for improving the generalization ability of the model.

- Evaluate and calibrate the long-term stability of the model

1. Multi-objective loss function: Add weights to the training so that the loss function focuses on both short-term and long-term prediction accuracy. For example, like (4):

$$\mathcal{L} = \alpha \times \mathcal{L}_{short-term} + (1 - \alpha) \times \mathcal{L}_{long-term} \qquad (4)$$

where α is the weight coefficient, which can be adjusted according to long-term and short-term needs.

2. Hierarchical forecasting architecture: Use hierarchical GAN to process short-term and long-term predictions separately and fuse the results through weighted average to consider both short-term accuracy and long-term stability.

- Add event-driven long-term simulation

Long-term market fluctuations are often driven by major events (such as policy changes, macroeconomic data releases). To improve long-term forecasting capabilities, event-driven simulation technology can be combined:

1. Event impact analysis: Use event-based models to analyze the impact of events on long-term market trends.

2. Simulate future scenarios: Incorporate simulated data generation techniques (such as Monte Carlo simulation) to assess possible market movements and supplement the stability of long-term forecasts.

Through the improvement of the above methods, the model is expected to maintain high prediction accuracy time longer and can effectively cope with the challenges brought by nonstructural data and market fluctuations. These improvements can also enhance the model's interpretability and generalization capabilities, making the research more widely applicable in different markets and scenarios.

References

[1] Yan, X., Zhang, T., & Liu, H. (2020). Applications of time-series data in real-world scenarios: Challenges and solutions. IEEE Transactions on Knowledge and Data Engineering, 32(5), 1015–1030.

[2] Wang, J., & Li, K. (2021). Non-structured time-series data and predictive models: A review. Journal of Data Science Research, 14(3), 89–102.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186.

[4] Xu, L., Sun, J., & Zhang, T. (2021). Enhancing stock price prediction using sentiment features: Evidence from financial news. Financial Data Science, 18(4), 85–98.

[5] Yoon, J., Jarrett, D., & van der Schaar, M. (2019). Time-series generative adversarial networks. Neural Information Processing Systems, 32, 550–562.

[6] Zhang, Y., Chen, M., & Liu, D. (2022). Combining sentiment analysis and technical indicators for anomaly detection in financial markets. Journal of Financial Technology, 10(1), 50–65.

[7] Liu, X., Chen, Y., & Xu, F. (2023). Predicting stock trends using multimodal data: A BERT and GAN-based approach. Journal of Computational Finance, 19(2), 130–145.

[8] Xu, Z., & Huang, L. (2020). Financial news sentiment analysis and market prediction. International Journal of Financial Engineering, 7(4), 105–120.

[9] Goodfellow, I., Bengio, Y., & Courville, A. (2020). Generative adversarial networks for time-series analysis: A practical guide. Deep Learning and Applications, 22(5), 345–370.

[10] Zhang, T., Xu, Z., & Sun, J. (2021). Predictive modeling of non-linear time-series with attention mechanisms. IEEE Transactions on Neural Networks and Learning Systems, 32(12), 5250–5263.

[11] Ya-Ting Chang, "Predicting Stock Price Trends using Natural Language Processing Techniques and Generative Adversarial Network Models — A Case Study of TSMC", Master Program in Artificial Intelligence Technology, National Taipei University of Technology, master's degree, 2024, Taiwan.

[12] Brian Hale, "The Only Technical Analysis Book You Will Ever Need", https://www.amazon.com/Only-Technical-Analysis-Book-Will/dp/B0CCX9DFJK