IJECBE (2024), **2**, **2**, 229–242
Received (4 June 2024) / Revised (5 June 2024)
Accepted (20 June 2024) / Published (30 June 2024)
https://doi.org/10.62146/ijecbe.v2i2.55
https://ijecbe.ui.ac.id
ISSN 3026-5258

International Journal of Electrical, Computer and Biomedical Engineering

RESEARCH ARTICLE

Implementation of Diffusion Variational Autoencoder for Stock Price Prediction with the Integration of Historical and Market Sentiment Data

Ardisurya* and Mia Rizkinia

Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia, Depok, Indonesia
*Corresponding author. Email: ardisurya@ui.ac.id

Abstract

This study aims to predict stock prices using a Diffusion Variational Autoencoder (DVAE) model that integrates technical data and market sentiment. Technical data is obtained from historical stock prices and trading volume, while sentiment data is derived from financial news analyzed using the IndoBERT model for sentiment classification. The research findings indicate that the integration of sentiment data in the D-VAE model enhances the accuracy of stock price predictions compared to a model that uses only technical data. Model evaluation is conducted using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2). The model with sentiment data integration has an MSE of 2753.204, MAE of 42.751, and R^2 of 0.94489, which are better than the model without sentiment data integration. This study demonstrates that the use of sentiment analysis can significantly contribute to improving stock price prediction performance using machine learning technology.

Keywords: Diffusion Variation Autoencoder, Stock Price Prediction, Indonesia Stock Market, Sentiment Analysis, IndoBERT

1. Introduction

The stock market is one of the key pillars of the global economy. Stock price movements can reflect the value of a company, economic conditions, and market sentiment. Stock price prediction is one of the greatest challenges in the financial world due to the volatility and complexity of the data involved.

Historical data and market sentiment are two main components that can be used to improve the accuracy of stock price predictions. Historical data provides information about past patterns and trends, while sentiment analysis offers insights into investor opinions and perceptions that can influence stock price movements. Sentiment analysis using news data has become a popular method in recent years. Market sentiment derived from news can reflect investor reactions to various events and the latest information. Previous studies have shown that integrating sentiment data with historical data can significantly improve the accuracy of stock price predictions [1].

However, the methods and models used still have much room for development and improvement. One promising model for handling the complexity of stock data is the Diffusion Variational Autoencoder (DVAE). DVAE is a generative model capable of capturing complex data distributions and producing accurate predictions even in highly volatile market conditions [2].

By integrating sentiment analysis and historical data, DVAE can leverage both types of information to provide more accurate and reliable predictions. This study aims to explore the use of the DVAE model in predicting stock prices by integrating historical data and sentiment analysis from market news using IndoBERT for sentiment analysis. IndoBERT is an Indonesian language model based on BERT that has proven effective in various natural language processing tasks, including sentiment analysis. Related research shows that combining sentiment analysis with historical data can improve stock prediction accuracy. A study by Al Ridhawi and Al Osman (2023) demonstrated that the combination of financial data and sentiment from social media can enhance stock price prediction performance by up to 74.3% [3].

Thus, this research is expected to provide significant contributions both in theory and practice in the fields of finance and economics, as well as pave the way for further research in the integration of sentiment and historical data for stock market prediction.

2. LITERATURE REVIEW

2.1 Stock Market

The stock market is a complex system where stock prices can be influenced by various factors such as company performance, economic conditions, and market sentiment. Stock price prediction has become a highly sought-after topic in the financial world due to the potential profits that can be obtained from accurate predictions. In this context, two main methods often used for analysis are fundamental analysis and technical analysis.

2.1.1 Fundamental Analysis

Fundamental analysis focuses on the intrinsic valuation of stocks based on company financial data, economic conditions, and other factors that affect company performance. This method evaluates aspects such as financial statements, industry conditions, and macroeconomic conditions to predict stock price trends. A systematic study shows that fundamental analysis, although less dominant compared to technical analysis, has an important contribution to stock market prediction, especially when combined with technical analysis [4].

The primary focus of fundamental analysis is on understanding the underlying factors that influence a company's performance. This involves a thorough examination of a company's financial health through its balance sheet, income statement, and cash flow statement. By assessing these documents, analysts can evaluate key metrics such as revenue, earnings, profit margins, and return on equity, which provide insights into a company's profitability and financial stability [5].

In summary, fundamental analysis provides a comprehensive framework for evaluating the intrinsic value of a security. By integrating financial data, qualitative factors, industry trends, and macroeconomic indicators, investors can make well-informed decisions and identify investment opportunities that may be undervalued or overvalued by the market. This approach helps in constructing a robust investment strategy aimed at achieving longterm financial goals.

2.1.2 Technical Analysis

Technical analysis is a method used to predict stock price movements based on historical price and trading volume data. This technique relies on charting tools and mathematical indicators to identify patterns and trends that can be used to forecast future price movements [6].

Three fundamental principles guide technical analysis: market action discounts everything, prices move in trends, and history tends to repeat itself. These principles imply that all relevant information is already reflected in the price, price movements follow identifiable trends, and historical patterns tend to recur due to consistent market psychology [7].

To implement technical analysis, analysts use a variety of tools and techniques, such as chart patterns and technical indicators. Chart patterns, like head and shoulders or triangles, provide visual cues that help identify potential future price actions [8]. Technical indicators, including moving averages, relative strength index (RSI), and Bollinger Bands, are mathematical calculations based on price and volume data. These indicators help identify trends, overbought or oversold conditions, and potential reversal points [9].

Candlestick patterns, such as doji, hammer, and engulfing patterns, provide further insights into market dynamics based on the shape and color of candlesticks. These patterns help predict future price movements and assist traders in making informed decisions. An example of technical analysis in action might involve using a combination of moving averages and RSI to determine optimal trade times, thereby maximizing gains or minimizing losses [10].

Overall, technical analysis offers a systematic approach to understanding market behavior and identifying trading opportunities, providing valuable insights for investors navigating the complexities of financial markets.

2.2 Machine Learning

Machine learning (ML) has revolutionized stock price prediction by enabling the efficient processing and analysis of large volumes of data. This allows for the identification of complex patterns and trends that may not be visible through traditional analysis methods.

One of the advanced techniques in ML for stock price prediction is the Diffusion Variational Autoencoder (DVAE). Diffusion Variational Autoencoder (D-VAE) is a combination of the principles of variational autoencoders with diffusion processes. D-VAE is designed to capture complex data distributions and generate accurate predictions even in highly volatile and uncertain market conditions. This technique implements a stochastic diffusion process to generate synthetic data similar to the training data, enriching the model with the ability to understand and simulate market dynamics [2].

The main advantage of D-VAE is its ability to model market uncertainty and variability, providing more realistic estimates of risk and potential outcomes. D-VAE can generate synthetic data that extends the model's understanding of market behaviour, enhancing the robustness and accuracy of predictions. The ability of D-VAE to adapt to dynamic changes in market data makes it a highly adaptive tool in stock price prediction.

2.3 Diffusion Variation Autoencoder (D-VAE)

The Diffusion Variational Autoencoder (D-VAE) combines the principles of variational autoencoders with diffusion processes to capture complex data distributions and generate accurate predictions even in highly volatile and uncertain market conditions [2]. This technique implements stochastic diffusion processes to generate synthetic data similar to training data, enriching the model with the ability to understand and simulate market dynamics.

The architecture of D-VAE includes several key components: the encoder, mean and log-variance layers, reparameterization trick, diffusion process with Gaussian noise, and decoder. Each component transforms input data into a more compact and informative latent representation and reconstructs the data from this latent representation while accounting for stochasticity and uncertainty in the data. The encoder transforms input data into a latent representation through several dense layers with ReLU activation. The ReLU (Rectified Linear Unit) activation function is defined as:

$$ReLU(x) = max(0, x) \tag{1}$$

This function helps address the vanishing gradient problem and introduces non-linearity into the model, allowing the neural network to capture more complex patterns in the data. The input data is processed through a series of dense layers to gradually reduce its dimensions. Each dense layer in the encoder compresses important information from the input data, transforming it into a simpler and more compact latent representation. After passing through dense layers in the encoder, the data is split into mean (μ) and logvariance (log σ^2) vectors, describing the latent distribution. The reparameterization trick is used to generate the latent variable $_z$, given by:

$$z = \mu + \sigma.\varepsilon \tag{2}$$

where $\epsilon \sim N(0, 1)$). This process maintains differentiability during backpropagation and captures uncertainty in the data. In D-VAE, a diffusion process is added by incorporating Gaussian noise into the latent representation, expressed as:

$$Z_{\text{diffuzed}} = Z + \varepsilon \tag{3}$$

where ε is Gaussian noise with mean 0 and variance σ^2 . This simulates the stochastic nature of stock prices, helping the model capture variability in market data. The diffused latent variable is then processed through the decoder, which reconstructs the input data from this latent representation using dense layers with ReLU activation. The final layer uses linear activation to produce stock price predictions. D-VAE is optimized by maximizing the Evidence Lower Bound (ELBO), which approximates the true data distribution p(x). The ELBO is given by:

$$L_{ELBO} = E_{q(z|x)} \left[log\ p\,(x|z) \right] - KL\,(q\,(z|x)\,\|p(z)) \tag{4}$$

Where $q\,(z \mid x)$ is the approximate posterior distribution, $p\,(x \mid z)$is the likelihood distribution, and $p\,(z)$ is the prior latent distribution, typically a standard normal distribution N(0,I). The Kullback-Leibler (KL) divergence helps the model learn better latent representations [11]. Denoising Score-Matching (DSM) is used to reduce aleatoric uncertainty by matching gradients, with the objective function given by:

$$L_{DSM,n} = E_{q(\gamma_n \mid \gamma)} \left[\| \gamma - \gamma_n + \bigtriangledown_{\gamma n} E\,(\gamma_n) \|^2 \right] \tag{5}$$

where y is the original data, γ_n is the noisy data, $E\,(\gamma_n)$ is the energy of the noisy data, and $\bigtriangledown_{\gamma n} E\,(\gamma_n)$ is the gradient of that energy. DSM helps reduce uncertainty arising from aleatoric noise, improving latent representations and making the model more robust to noise and variations in the data [12].

2.4 Market Sentiment Analysis

Market sentiment analysis utilizes Natural Language Processing (NLP) technology to evaluate and interpret market opinions contained in various data sources such as news, social media, and financial reports. Market sentiment can influence investor perceptions and decisions, which in turn can affect stock price movements. Natural Language Processing (NLP) is a branch of artificial intelligence focused on the interaction between computers and human language. NLP enables computers to understand, interpret, and generate human language in a useful way. This technology is used to identify patterns, trends, and sentiments related to the stock market, aiding in more informative and data-driven decision-making [13]. The IndoBERT model is an NLP model specifically adapted for the Indonesian language. This model is trained using more than 220 million words from various sources such as Indonesian Wikipedia and news articles. IndoBERT is used for sentiment analysis by assessing and classifying the sentiment of news text into positive, negative, or neutral categories. This model allows for more accurate prediction processes of sentiment contained in Indonesian-language texts [14].

3. IMPLEMENTATION

In this section, we outline the implementation of the Diffusion Variational Autoencoder (D-VAE) model, which is designed to utilize historical stock price data and market sentiment data to produce more accurate stock price predictions. To provide a better understanding of the architecture and working mechanism of the D-VAE model, below is an image illustrating the main components and data flow within the model.

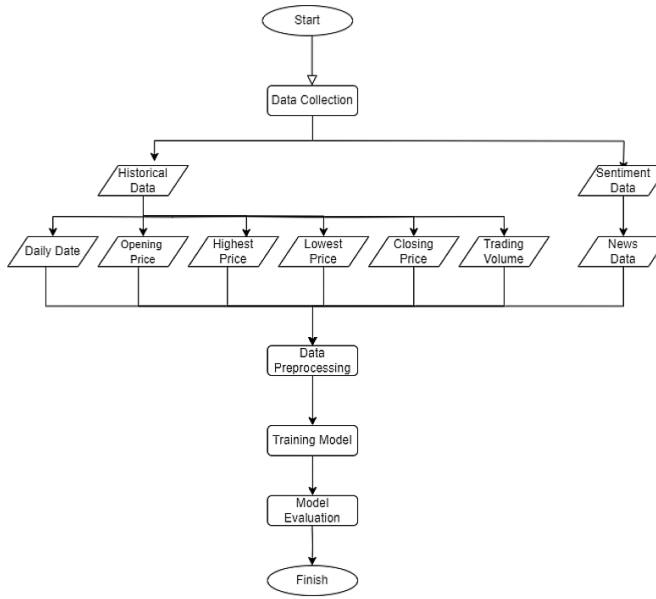

Figure 1. Flowchart Of Method for Stock Price Prediction Using Historical and Sentiment Data

3.1 Data Collection

Data collection is a crucial initial step in this research, as the quality and completeness of the data will directly influence the results and accuracy of the prediction model. In this study, the collected data consists of two main types: historical stock price data and sentiment data from news articles.

3.1.1 Historical Data

Historical stock price data is obtained from Yahoo Finance, one of the reliable sources for financial information. This data includes several key elements used in stock price analysis and prediction: transaction date, opening price, highest price, lowest price, closing price, and trading volume. The transaction date represents the date on which the stock transactions occurred. The opening price is the stock price at the market open. The highest price is the highest price reached by the stock during the trading session. The lowest price is the lowest price reached by the stock during the trading session. The closing price is the stock price at the market close. Lastly, the trading volume is the total number of shares traded during the trading session. For this research, we focused on the stock price data of Bank BCA (Bank Central Asia).

The period taken for this research is from January 1, 2022, to January 1, 2023. This data is downloaded in CSV format and undergoes an initial cleaning stage to remove missing or incomplete values. This process ensures the data integrity for further analysis. In total, the dataset includes 245 trading days of historical stock prices for Bank BCA.

3.1.2 Sentiment Data

Sentiment data is obtained from financial news articles collected through a web scraping process from various online news sources. The sentiment data collection process is carried out systematically to gather information reflecting market sentiment related to the analyzed stocks. The sentiment data collection process involves several steps. First, the news search is conducted using the Google search engine. Relevant keywords, such as "Berita Bank BCA," are used to find news articles related to the stocks being studied. The search is conducted daily to ensure that the collected news is always up-to-date and relevant to the current market conditions. From the search results, only the three most relevant news articles per day are selected for further analysis. Once the relevant news articles are identified, the next step is to collect the news data through web scraping techniques.

Web scraping is an automated process used to extract information from web pages. In this research, web scraping is performed using a Python script. This script is designed to access the identified news web pages, then extract important information such as news titles, full news content, and the URL source of the news. The script works by sending HTTP GET requests to each collected news URL. If the request is successful (status code 200), the HTML content of the web page is retrieved and processed using BeautifulSoup, a Python library for HTML parsing.

The news title is extracted from the <h1> tags, while the news content is extracted from the <p> tags. This information is then stored in a CSV file with columns that include the news title, news content, and URL source. After the news data is successfully collected and stored in CSV format, the next step is to combine these news data into a single DataFrame to facilitate further analysis. The news date is extracted from the file name and added as a column in the news DataFrame. Thus, each news entry will have date information allowing integration with the historical stock price data. This sentiment data collection process ensures that the dataset used in this research includes relevant and up-to-date information about market sentiment.

The sentiment data obtained from the news will then be further analyzed using Natural Language Processing (NLP) models to assess the market sentiment contained in each news article. Integrating sentiment data with historical stock price data will provide deeper insights into how market sentiment affects stock price movements.

3.2 Data Preprocessing

3.2.1 Historical Data Preprocessing

Historical data preprocessing is a crucial step in ensuring that the data is ready for use in prediction models. In this study, historical stock price data is processed to clean, format, and normalize the data so it can be integrated with sentiment data and used in model training. Historical stock price data often contains missing values due to incomplete data or errors in data collection. The first step in preprocessing is to remove rows with missing values to ensure data integrity. This is done using the dropna() function from the pandas library, which removes all rows that contain NaN (Not a Number) values. The dates in the historical stock price data are converted to datetime format to facilitate merging with sentiment data. The date column (Date) is converted using the pd.to_datetime() function from the pandas library.

This conversion ensures that date data can be correctly used in time series analysis and data merging. To ensure that all features are within the same range, data is normalized using Min-Max Scaling. Normalization is performed using the MinMaxScaler from scikit-learn, which scales data values to a range of 0 to 1. The normalized features include opening price (Open), highest price (High), lowest price (Low), trading volume (Volume), and closing price (Close). Normalizing the data is important to reduce the different scales of these features, allowing the model to be trained more effectively and efficiently.

3.2.2 Sentiment Data Preprocessing

Sentiment data preprocessing is an essential step to transform raw data from financial news into data that can be used in prediction models. Sentiment data is obtained from financial news articles collected through web scraping from various online news sources. The process begins with collecting news data from various online sources that provide financial news. Each news item collected includes the news title and publication date. After gathering the news data, the next step is to preprocess the text to remove unnecessary elements and prepare the data for sentiment analysis. During the text preprocessing stage, punctuation is first removed using the function $re.sub\left(\acute{r}\left[\hat{}\setminus w\setminus s\right]',\text{''},text\right)$ to ensure that the text is free from irrelevant characters.

Next, the text is converted to lowercase for consistency, and common words that do not carry significant meaning (stop words) are removed using the NLTK library. These steps help reduce noise in the data and improve the accuracy of sentiment analysis. After the news text is processed, sentiment analysis is performed to determine the sentiment of each news item. The IndoBERT model, a pre-trained transformer model specifically designed for the Indonesian language, is utilized for this purpose. First, the IndoBERT model and its tokenizer are loaded to convert the news text into a format suitable for the model. The preprocessed news text is then tokenized using the IndoBERT tokenizer, converting the text into tokens that the model can process. The tokenized text is fed into the IndoBERT model to obtain sentiment scores.

The model outputs a sentiment score for each news item, indicating the sentiment's polarity. The sentiment scores range from 0 to 10, where 0 indicates extremely negative sentiment and 10 indicates extremely positive sentiment. After obtaining the sentiment scores for each news item, the daily sentiment score is calculated. This involves aggregating the sentiment scores of all news articles published on the same day.

The aggregation is done by computing the average sentiment score of all articles for each day. This daily sentiment score provides an overall view of market sentiment for each day and allows for the integration of sentiment data with historical stock price data.

3.2.3 Merging Historical Data with Sentiment Data

After processing the sentiment data and historical stock price data separately, the next step is to merge these two types of data to form a comprehensive dataset that will be used in the prediction model training. The purpose of merging the data is to utilize market sentiment information along with historical stock price data to improve the accuracy of stock price predictions.

Both the historical stock price data and sentiment data have date columns used as the merging key. The historical data includes opening price (Open), highest price (High), lowest price (Low), closing price (Close), and trading volume (Volume) for each date. Meanwhile, the sentiment data includes daily sentiment scores generated from the financial news analysis. The merging process is performed using the pd.merge() function from the pandas library. This function merges two DataFrames based on the same column, which is the date column (Date) in the historical data and the date column (Date) in the sentiment data. The merging is done using the left join method so that all rows from the historical data remain, and the rows from the sentiment data that correspond to the dates in the historical data are added.

After merging, the resulting dataset will have all the features from the historical data plus the market sentiment features. These features are then used as input in the prediction model. This combined data allows the model to utilize additional information from market sentiment that may influence stock prices.

3.3 DVAE Model Implementation

The implementation of the model in this study focuses on using the Diffusion Variational Autoencoder (D-VAE) to predict stock prices. D-VAE was chosen for its ability to capture complex data distributions and generate accurate predictions in volatile market conditions. This model is tested with market sentiment data obtained from financial news analysis to assess the impact of sentiment on prediction accuracy. The flowchart in Figure 1 illustrates the method for stock price prediction using these data types, and Figure 2 outlines the architecture of the D-VAE model.

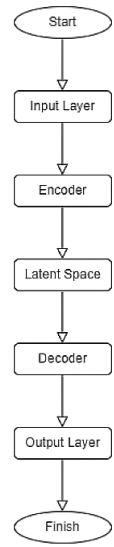

Figure 2. Flowchart D-VAE architecture

The encoder component of the D-VAE model consists of several dense layers, beginning with 512 neurons and progressively reducing to 8 neurons in the final layer. The ReLU activation function is used to introduce nonlinearity into the encoding process, ensuring that complex patterns in the input data are captured. The encoder transforms the input data into a latent space representation, which is a condensed form of the data containing essential information. In the latent space, the model learns a probabilistic representation of the input data. This space is modeled to follow a Gaussian distribution, with the encoder outputting the parameters of this distribution (mean and variance).

The model samples latent variables from this distribution during the decoding process. The decoder component reconstructs the input data from the latent space representation. It inversely processes the latent variables, gradually increasing the dimensionality of the data until it matches the original input dimensions. The final layer of the decoder uses a linear activation function to produce stock price predictions. Training the D-VAE model involves optimizing a composite loss function that includes both reconstruction loss and KL-divergence loss. The reconstruction loss measures the accuracy of the decoded output compared to the original input, while the KL-divergence loss ensures that the latent distribution closely approximates the prior Gaussian distribution. This dual-objective optimization promotes both accurate reconstruction and good generalization.

The integration of sentiment data, obtained from financial news analysis using the IndoBERT model, adds a significant layer of contextual information to the prediction process. IndoBERT processes the preprocessed news text to generate sentiment scores ranging from 0 (extremely negative) to 10 (extremely positive). These scores are aggregated daily and combined with historical stock price data to form a comprehensive feature set. This integration allows the D-VAE model to leverage both historical

trends and current market sentiment, enhancing its predictive power. The D-VAE model is trained using a dataset that includes both historical stock prices and sentiment scores. The dataset is divided into training and validation sets to evaluate the model's performance.

An optimizer such as Adam is used to minimize the loss function, and hyperparameter tuning is conducted to identify the best configuration for the model. The training process is iterative, with the model learning to accurately predict stock prices based on the combined data inputs. The model's performance is assessed using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2). These metrics provide a quantitative measure of the model's accuracy and reliability in predicting stock prices.

The effectiveness of the D-VAE model is further validated by comparing its performance with and without the inclusion of sentiment data. The results of the D-VAE model demonstrate its capability to predict stock prices more accurately when sentiment data is incorporated. The comparison between models with and without sentiment data shows a significant improvement in prediction accuracy, highlighting the added value of integrating market sentiment into financial prediction models.

4. RESULT AND DISCUSSION

This study aims to evaluate the effectiveness of using the Diffusion Variational Autoencoder (D-VAE) in predicting stock prices by considering market sentiment data. In this chapter, we will present the results of the D-VAE model both with and without sentiment data, as well as discuss the implications of these results on the model's performance in predicting stock prices.

Table 1. input data between historical data and sentiment score

Date	Open	High	Low	Close	Volume	Sentiment Score
03/01/2022	Rp 7325	Rp 7400	Rp 7300	Rp 7325	54287400	0
04/01/2022	Rp 7325	Rp 7450	Rp 7325	Rp 7400	70624000	6
05/01/2022	Rp 7450	Rp 7525	Rp 7375	Rp 7450	76164900	6
06/01/2022	Rp 7500	Rp 7525	Rp 7425	Rp 7475	63657100	10
07/01/2022	Rp 7550	Rp 7700	Rp 7500	Rp 7650	143433300	3

Table 1 represents a slice of the data used in this analysis. The data includes daily stock prices consisting of opening (Open), highest (High), lowest (Low), closing (Close) prices, trading volume (Volume), and sentiment scores calculated based on financial news analysis. This data is crucial for understanding how market sentiment can influence stock price movements.

Table 2. Comparison Of Model Evaluation For Each Data

Model Evaluation	Integration With Sentiment Data	Integration Without Sentiment Data
Mean Squared Error	Rp 2753,204	Rp 3490,819
Mean Absolute Error	Rp 42,751	Rp 46,220
R-Squared	0,94489	0,93013

Table 2 compares the performance of the D-VAE model with and without sentiment data. The results show that integrating sentiment data significantly improves the model's prediction accuracy. The integration of sentiment data provides additional context to market movements that pure technical analysis might miss. This added layer of information helps the model to make more nuanced predictions, particularly in capturing market sentiment-driven fluctuations.

Table 3. Paired t-Test Results Comparing Prediction Errors (MSE and MAE) with and without Sentiment Data

Metric	p-value
Paired t-test for MSE	0.0042
Paired t-test for MAE	0.0007

The p-values obtained from the paired t-tests are far less than the significance level of 0.05, indicating that the differences in prediction errors between the models with and without sentiment data are statistically significant. This confirms that the integration of sentiment data significantly improves the model's prediction accuracy. The low p-values indicate that the improvement is not due to random chance, and the sentiment data provides meaningful additional information that enhances the model's performance.

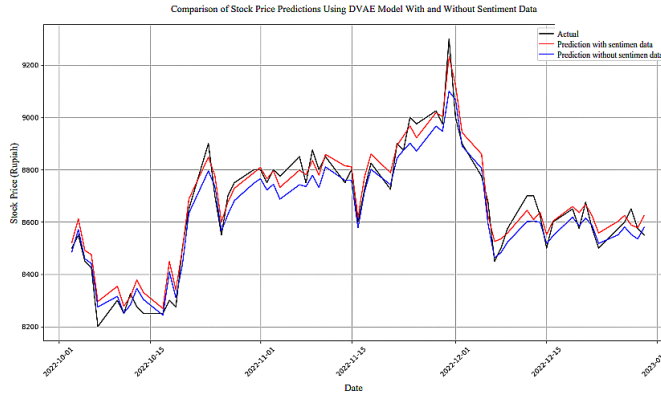

Figure 3. Chart Of Comparison of Actual Data, Prediction with Sentiment data, and Prediction Without Sentiment Data

The comparison graph between actual stock prices and the prices predicted by the D-VAE model with sentiment shows that this model is more capable of following stock price trends, particularly in volatile market conditions. Conversely, the model without sentiment shows limitations in capturing sudden changes in stock prices, indicating that historical stock price data alone is not sufficient for predicting stock prices with high accuracy. Furthermore, the evaluation results with metrics such as MSE, MAE, and R^2 support these findings.

Lower MSE and MAE values in the model with sentiment indicate smaller prediction errors, while a higher R^2 value shows that this model is better at explaining the variability in stock price data. In a practical implementation context, these findings suggest that integrating market sentiment data into stock price prediction models can provide significant benefits for investors and market analysts. By leveraging information from financial news, the prediction model can become a more reliable tool in aiding investment decision-making.

5. CONCLUSION

This study aimed to evaluate the effectiveness of using Diffusion Variational Autoencoder (D-VAE) in predicting stock prices by incorporating market sentiment data obtained from financial news analysis. Based on the results and discussions presented in the previous chapters, several key conclusions can be drawn from this research. First, the study results indicate that integrating market sentiment data into the D-VAE prediction model improves the accuracy of stock price predictions. This is evidenced by the better evaluation metric values such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2) in the model with sentiment compared to the model without sentiment. Lower MSE and MAE values, along with higher R^2 values, demonstrate that the model with sentiment can predict stock prices with smaller errors and better explain the variability in the data. Second, market sentiment data obtained from financial news analysis provides important additional information about market conditions and investor perceptions. This information cannot be

obtained solely from historical stock price data. By integrating sentiment data, the prediction model can capture more complex market dynamics and be more responsive to rapid changes in market conditions. Overall, this study successfully demonstrates that using the D-VAE model with market sentiment data is an effective approach to predicting stock prices. The findings of this study are expected to contribute to the development of more sophisticated and effective stock price prediction models in the future. Further research can be conducted to explore the use of various sources of sentiment data and more complex analysis methods to enhance the accuracy of stock price predictions.

References

[1] L. Owen and F. Oktariani. "SENN: Stock Ensemble-based Neural Network for Stock Market Prediction using Historical Stock Data and Sentiment Analysis". In: *2020 International Conference on Data Science and Its Applications (ICoDSA)*. IEEE, 2020, pp. 1–7.

[2] K. J. L. Koa et al. "Diffusion Variational Autoencoder for Tackling Stochasticity in Multi-Step Regression Stock Price Prediction". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. ACM, 2023, pp. 1087–1096.

[3] M. Al Ridhawi and H. Al Osman. "Stock Market Prediction from Sentiment and Financial Stock Data Using Machine Learning". In: *Proceedings of the Canadian Conference on Artificial Intelligence*. 2023, pp. 1–6.

[4] I. K. Nti, A. F. Adekoya, and B. A. Weyori. "A systematic review of fundamental and technical analysis of stock market predictions". In: *Artificial Intelligence Review* (2020), pp. 3007–3057.

[5] K. Priyanka. "The Study of Fundamental & Technical Analysis". In: *International Journal of Scientific Research in Engineering and Management (IJSREM)* 6.5 (2022), pp. 1–20. DOI: 10.55041/IJSREM13093.

[6] S. Zhong and D. B. Hitchcock. "S&P 500 Stock Price Prediction Using Technical, Fundamental and Text Data". In: *arXiv preprint arXiv:2108.10826* (2021), pp. 1–19.

[7] Y. Han et al. "Technical Analysis in the Stock Market: A Review". In: *Journal of Financial Markets* 33 (2021), pp. 2326–2377. URL: https://ssrn.com/abstract=3850494.

[8] M. F. Dicle. "Candle charts for financial technical analysis". In: *The Stata Journal* 19.1 (2019), pp. 200–209. DOI: 10.1177/1536867X19830918.

[9] P. Oktaba and M. Grzywińska-Rąpca. "Modification of technical analysis indicators and increasing the rate of return on investment". In: *Central European Economic Journal* 10.57 (2023), pp. 148–162. DOI: 10.2478/ceej-2023-0009.

[10] A. Ratto et al. "Technical analysis and sentiment embeddings for market trend prediction". In: *Expert Systems with Applications* 135 (2019), pp. 60–70. DOI: 10.1016/j.eswa.2019.06.014.

[11] D. P. Kingma and M. Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML].

[12] Chao et al. "Denoising Likelihood Score Matching for Conditional Score-Based Data Generation". In: *International Conference on Learning Representations*. 2022, pp. 1–24.

[13] M. Napizahni. *dewaweb*. https://www.dewaweb.com/blog/nlp-adalah/. [Online]. June 2022.

[14] F. Koto et al. "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP". In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020, pp. 1–14.