Predicting Stock Market Trends with News Sentiment Analysis and Historical Stock Market Data Analysis Using Data Mining Techniques

Abraham Hotasi, Dicky Prima Satya
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
abrahamhotasi@gmail.com, dicky@itb.ac.id

2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA) | 979-8-3503-2991-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAICTA59291.2023.10390037

***Abstract*—Stock market trends are one of the most difficult topics to predict, as stock prices are influenced by various variables. Several studies have examined various ways to predict stock market trends, but research on this topic has been very rare in Indonesia. Based on a data mining literature review the past 5 years, the methodology proposed by Khedr, Salama, and Yaseen (2017) that combines news sentiment analysis with historical stock market data attributes achieved the highest accuracy (86.21%) in the training and testing phases compared to other studies. Therefore, this paper aims to match the methodology used by Khedr, Salama, and Yaseen (2017) with data from stocks listed on the Indonesia Stock Exchange (IDX) to predict stock trends with different characteristics (BBCA, BUMI, ASII). In addition, this paper will also examine the process of how the stock market trend prediction model works. The first step of this paper is to determine the polarity of financial news with a Naive Bayes classifier, followed by the second step, which is to combine the news polarity with historical stock price data to predict stock market trends using a KNN classifier. The first step of this paper, determining the polarity of financial news, successfully achieved a prediction accuracy of 91.9 - 97.3% in the training and testing phases and 92.3% - 96% in the validation phase. The second step, predicting stock market trends in the future, successfully achieved a prediction accuracy of 72.2% - 90% in the training and testing phases and 43.5% - 64.6% in the validation phase. This research is one of the first studies conducted on stocks listed on the IDX by combining financial news data with historical stock market data attributes.**

Keywords—machine learning, IDX, natural language processing, data mining, stock price trend, sentiment analysis

I. INTRODUCTION

In the modern economy, the stock market has become one of the best ways to invest. According to data from the World Bank, the market capitalization of domestic listed companies or the value of the stock market worldwide at the end of 2020 was $93.7 trillion (increased 38% from 2018). In Indonesia, the performance of the Indonesian capital market has been growing steadily, accompanied by a significant increase in market capitalization value. As of September 2022, the market capitalization value of the Indonesia Stock Exchange (IDX) of the 50 largest stocks in Indonesia is Rp9.2trillion and contributes 50% to Indonesia's GDP. The market capitalization value has increased by 70% compared to the period of October 2020 which was Rp5.6 trillion.

The conditions of the capital market in Indonesia are supported by the surge in investor growth, which is dominated by Gen Z and millennials aged under 30 years. Based on data from the Central Securities Custodian Agency of Indonesia (KSEI), as of September 2022, the number of capital market investors has reached 9.8 million investors. This number has skyrocketed 150% compared to December 2020, which was 3.9 million investors.

From the perspective of issuers or companies listed on the IDX, 43 new issuers have been listed since the beginning of 2022 to August 2022. The total value of the initial public offering (IPO) emissions reached Rp21.6 trillion and the actual number of issuers that carried out IPOs increased 53.5% annually. In fact, from the same period in 2021, only 28 new companies were listed. Not stopping there, the IDX still has 23 potential issuers in the stock listing pipeline as of September 2022, with an estimated total value of IPO emissions of 23 companies reaching Rp9.5 trillion.

If compared to other developing countries, the return on the Indonesia Composite Stock Price Index (IHSG) is also one of the largest. Since the beginning of 2022, the return on IHSG has reached 9.05%, while Malaysia is -4.78%, the Philippines is -6.04%, Thailand is -2.10%, and Vietnam is -14.53%.

These data reflect how the capital market in Indonesia can be a huge opportunity for investors to make a profit and make an impact on Indonesia through the issuers that are invested in. However, to be able to make a profit through capital gains, investors are faced with the risk when buying shares of a company. Deciding which shares to buy and when to sell is very difficult to do because of the complexity of the data available and the instability of the capital market to predict its behavior.

The current state-of-the-art approach to predicting stock price trends that leverages data mining techniques with the analysis of historical stock data (open, high, low, close) and financial news sentiment analysis is being done by many researchers, including Pavithya et al. (2021) and Khedr, Salama, and Yaseen (2017). The research of Pavithya et al. (2021) and Khedr, Salama, and Yaseen (2017) used data mining techniques, financial news datasets, and stock datasets in different stock markets and got an accuracy of 58% and 89.80% respectively. Seeing the success of the prediction research, this paper proposes a stock market prediction methodology by analyzing historical stock data (open, high, low, close) and financial news sentiment analysis to stocks listed on the IDX.

Hence, since Khedr, Salama, and Yaseen's methodology achieved the highest accuracy in this study literature review, the objective of this paper are as follows: 1) To match the methodology proposed by Khedr, Salama, and Yaseen (2017): analysis of historical stock price data and financial news sentiment, with shares listed on the IDX; 2) To show whether the methodology used proposed by Khedr, Salama, and Yaseen (2017) can predict price trends with the same accuracy or even better on shares listed on the IDX.

Other state-of-the-art approaches also include research by Xiao, Ihaini (2017) and Mudinas, Zhang, Levene (2019). Research by Xiao, Ihaini (2017) designed two different time divisions: $0:00_t \sim 0:00_{t+1}$ and $9:30_t \sim 9:30_{t+1}$ to study how tweets and news from the different periods can predict the next-day stock trend. The study selected 260,000 tweets and 6,000 news from Service stocks (Amazon, Netflix) and Technology stocks (Apple, Microsoft). As a result, the experiment shows that opening hours division ($9:30_t \sim 9:30_{t+1}$) outperformed natural hours division (0:00t~0:00t+1). While the research done by Mudinas, Zhang, Levene (2019) investigates the potential of using sentiment attitudes (positive vs negative) and sentiment emotions (joy, sadness, etc) extracted from financial news or tweets to help predict stock price movements. The result of the study that uses Granger-causality revealed that: 1) In general, sentiment attitudes do not seem to Granger-cause stock price change; 2) On some specific occasions, sentiment emotions do seem to Granger-cause stock price changes, the exhibited pattern is not universal and must be analyzed on a case-by-case basis. Moreover, it's also shown that for certain stocks, integrating sentiment emotions as additional features into machine learning based market trend prediction model could improve its accuracy.

II. RELATED WORKS

There have been several studies or approaches that have been conducted to predict the behavior and trend of stock market prices. Some of these studies focused on improving the accuracy of the prediction based on the analysis of news sentiment or tweets along with stock prices, such as the study by Alostad (2015). Others focused on predicting prices with different time periods, such as the study by Uhr & Zenkert (2014). In addition, there are other research approaches that have proven that there is a high correlation between financial news and stock price changes, such as the study by Walter et al. (2013) and the study by Zubair (2015). Finally, there are a few studies that have conducted research to improve the accuracy of prediction, such as the study by Hoang (2014).

Previous research has encountered difficulties due to the intricate nature of handling unstructured data. Several of these methods employ text mining strategies to forecast stock market patterns, while other investigations use textual data in contrast to a stock's closing prices, and some also incorporate textual information alongside stock price charts.

A. Studies Related to Social Media and Stock Price Trend Analysis

In 2014, Bing and Ou conducted research that introduced an algorithm aimed at forecasting stock price movements with a noteworthy accuracy rate of 76.12%. Their approach involved analyzing publicly available social media data in the form of tweets. Bing employed a model designed to examine public tweets alongside hourly stock price trends, integrating natural language processing (NLP) and other data mining techniques to ascertain the patterns that link public sentiment and stock prices. The study delved into the exploration of potential internal associations within a hierarchical data structure, revealing a connection between the inner and outer layers of unstructured data. It's essential to note that this investigation exclusively considered historical stock prices' daily closing values.

Moving to 2015, Y. E. Cakra's study also proposed a model for forecasting the Indonesian stock market by assessing sentiment in tweets. Their model had three primary objectives: predicting price fluctuations, margin percentages, and stock prices. They employed five supervised classification algorithms – support vector machine (SVM), naïve Bayes classification, decision trees, random forests, and neural networks – to predict tweet sentiment. The results demonstrated that random forest and naïve Bayes classifiers surpassed other algorithms, achieving accuracy rates of 60.39% and 56.50%, respectively. Additionally, linear regression demonstrated strong performance in price prediction with an accuracy rate of 67.73%. However, it's important to acknowledge that this study's limitation lay in constructing the prediction model solely based on the last 5 days' price data.

B. Studies Related to News Analysis and Stock Price Trend Analysis

In 2014, Uhr and Zenkert conducted a study employing a combination of text mining techniques to assess market sentiment by incorporating word associations and lexical sources for the analysis of stock market news reports. This research focused on the German language and utilized the sentiWS tool to gauge sentiment across various levels. The study directly compared stock prices on the capital market with the sentiment measurement model to provide investors with recommendations for the upcoming week, assisting them in mitigating investment risks.

On the other hand, in 2017, Khedr, Salama, and Yaseen undertook research to predict the stock prices of Yahoo, Microsoft, and Facebook. Their proposed model delivered an impressive prediction accuracy of 89.90%. This investigation considered two key factors: 1) Analyzing historical stock prices, considering attributes like opening, high, low, and closing (OHLC) prices, and 2) Analyzing stock-related textual content from various sources such as NASDAQ, Reuters, and the Wall Street Journal. To achieve their results, this study employed two data mining classification methods, the KNN algorithm and the Naive Bayes algorithm.

III. PROPOSED METHODS

The proposed model aims to predict whether the stock market will rise, fall, or stay stable. It does this by combining the analysis of stock market news and historical stock prices. The text analysis of stock market news determines the polarity of the news articles, while the historical prices (opening, high, low, and closing) are analyzed to predict the stock market behavior.

A. Data Description

In this paper, there are two groups of data that will be collected: numeric data and financial news. Both groups of data will be divided into two subgroups: 1) Training and testing data; 2) Validation data.The training and testing data will be used to train each sentiment analysis model and the combined model (sentiment analysis and historical stock price data). The validation data will be used to see the accuracy of the model that has been trained with the training and testing data in predicting new data that has never been seen by the

model before. The period of financial news data and historical stock price (numeric) data that is used to become training and testing data starts from January 1, 2021 to February 13, 2023. The period of validation data that is used starts from February 14, 2023 to June 26, 2023.

The stock data used are BBCA, BUMI, and ASII stock from IDX. The financial news data used was taken through the bisnis.com news portal with a Python news scraper program created by the author using BeautifulSoup and requests libraries. The bisnis.com news portal was chosen because bisnis.com almost every day publishes news about the stocks that will be used in this paper (BBCA, BUMI, and ASII). Almost every day, financial news about the stock market or a company is uploaded by bisnis.com. Some of these news can be in the form of an article or news about a company's financial report. The news that will be considered in this paper is news about a company's dividend stocks, stock splitting, news about extraordinary stock changes, and the acquisition or merger of a company with another company, stock merge, and other news that can affect stock price trends.

For the training and testing data, the news data used is from the period January 1, 2021 to February 13, 2023. There are 1450 financial news data for BBCA shares, 350 financial news data for BUMI shares, and 769 financial news data for ASII shares. For the validation data, the news data used starts from February 14, 2023 to June 26, 2023. There are 210 financial news data for BBCA shares, 45 financial news data for BUMI shares, and 121 financial news data for ASII shares.

As for the historical stock price data, the data was taken through the Yahoo Finance website. For the numeric data, the opening, high, low, and closing (OHLC) price attributes of the stock will be used because this data has a direct effect on the prediction of stock price trends in the future. The stock data used are stocks that are traded on the Indonesia Stock Exchange (BEI): BBCA, BUMI, and ASII. For the training and testing data, the news data used is from the period January 4, 2021 to February 13, 2023. There are 526 numeric data for BBCA, BUMI, and ASII stocks. For the validation data, the numeric data used starts from February 13, 2023 to June 26, 2023. There are 87 data for BBCA, BUMI, and ASII stocks.

The data period is selected because the more historical data that is used to train and test the model will improve the accuracy of the model in predicting stock price trends. Historical stock price data from 2019-2020 will not be used to train or test the model because the volatility of stock price trends is greatly influenced by the force majeure that occurred: COVID-19. Whereas in 2021, with the start of the circulation of vaccines, the stock price trend began to normalize and began to follow the behavior of the stock price trend before COVID-19 occurred.

B. Proposed Method Components

The proposed model architecture is a reference from Khedr, Salama, and Yaseen (2017) that will be matched with the data available in Indonesia. In its execution, several modifications will also be made to process the data and improve the model accuracy, such as: 1) Translating Indonesian financial news into English so it can be processed; 2) Processing Indonesian financial news data with Indonesian stop words and stemmer library before translating it into English; 3) Taking advantage of model hyperparameter tuning using GridSearchCV.

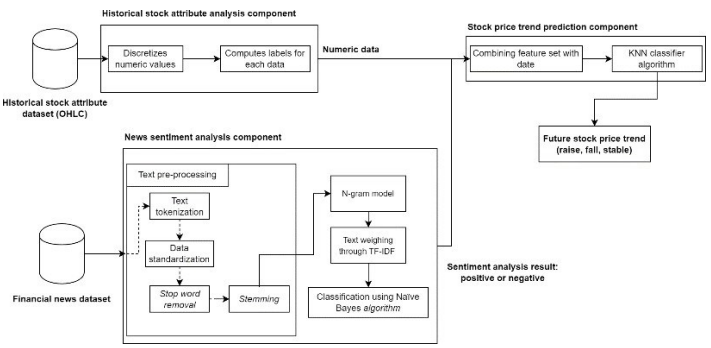

Fig. 1. Proposed stock price trend prediction model

1) News Sentiment Analysis Component

In this component, the analysis of financial news data is carried out with the aim of classifying whether a news has a positive or negative sentiment towards BBCA, BUMI, or ASII shares. To achieve this goal, text pre-processing is first performed on the news text, followed by news classification using the Naïve Bayes classification algorithm.

There are 5 results of this sentiment analysis component: 1) Financial news data with 4 features: title, description, date, polarity, subjectivity, and sentiments; 2) Accuracy results of the Naive Bayes classification model against training and testing or validation results; 3) Confusion matrix from training and testing or validation results; 4) Naive Bayes vectorizer used in the validation phase; 5) Sentiment analysis prediction model used in the validation phase.

Here are the detailed steps on the news sentiment analysis component.

- Text Pre-Processing

The text pre-processing performed in this section on the financial news data is the same as what was explained in the previous chapter. There are 6 steps in performing text pre-processing so that the news sentiment can be analyzed: 1) Tokenization; 2) Data standardization; 3) Stop-word removal; 4) Stemming; 5) Translating financial news into English; 6) Labeling the sentiment on translated financial news.

- N-Gram

Following data pre-processing, the next step involves N-Gram feature extraction, which plays a pivotal role in numerous text mining and natural language processing (NLP) endeavors. N-Gram, in this context, refers to a sequence of tokens or words with a specific length, denoted as 'n.' In the proposed model, N-Grams are created for financial news documents, serving as a means to extract keyword features from the news data corpus. For instance, when 'n' in N-Gram is set to 2, the process generates sequences of two consecutive words for each document. This particular phase substantially enhances the classifier's accuracy, as it furnishes a series of two-word combinations, thereby providing valuable information for analysis.

- TF-IDF (Term Frequency - Inverse Document Frequency)

The proposed model employs TF-IDF, a feature weighting method utilized to assess the significance of individual words within a document or across a corpus. TF-IDF calculates the importance of a word by considering the ratio of IDF in a specific document to the overall occurrences of that word across all documents. Words with high TF-IDF values indicate a strong association between the word and the document in which it is found.

- Naïve Bayesian Classifier

The Naïve Bayesian classifier is employed to categorize financial news articles into either positive or negative sentiment categories, utilizing the TF-IDF values and labels assigned to them. The Naïve Bayes algorithm operates on the assumption that the impact of an attribute's value on a class is independent of the values of other attributes; this assumption is referred to as class conditional independence. This classifier is chosen for its simplicity and speed in text classification. It excels in predicting the sentiment polarity of each document. The algorithm's task is to assign a label of either positive or negative to each financial news article based on the features extracted and analyzed.

2) Historical Stock Price Data Component

In the context of historical stock price data, the key attributes considered are the open, high, low, and close (OHLC) prices. To process this data effectively, the initial step involves converting these numerical attributes into discrete values, specifically "positive," "negative," or "equal," following the approach outlined by Kim, Jeong, and Ghani in 2014. This conversion hinges on comparing each numeric value within the OHLC attributes to the closing price value of the previous day. How it works is: 1) If the current attribute value (OHLC) is greater than the previous day's closing price, it is replaced with the label "positive."; 2) If the current attribute value is less than the previous day's closing price, it is replaced with the label "negative."; 3) If the current attribute value is equal to the previous day's closing price, it is replaced with the label "equal."

In summary, these actions categorize the OHLC attribute values into "positive," "negative," or "equal" based on a comparison with the previous day's closing price. The subsequent step involves determining a label for each data sample, and this is done by calculating the future stock price trend, which can be categorized as "rise," "fall," or "stable." This categorization is based on the difference between the closing stock price on the current day and the previous day.

The final outcome of this process is a dataset with the following features for each data point: 1) Date; 2) Open; 3) High; 4) Low; 5) Close; 6) Stock Price Trend.

3) Stock Price Trend Prediction Component

The proposed prediction model combines financial news data from the previous sentiment analysis component and historical stock price data to investigate the influence of news releases and historical stock prices on the rise or fall of stock prices. The prediction model component consists of 2 steps, combining the feature set with the date and predicting the stock price trend using the K-nearest neighbor (KNN) classifier. The output of the prediction model component is the KNN classifier model that has been trained with the stored training data and will be used in the validation phase.

- Combining Feature with Date

The two datasets from the previous step will be joined by date. News sentiment and numerical data (OHLC and Stock Price Trend) are merged with the stock price date, which will produce news sentiment and OHLC features for each day. Therefore, the final dataset to be used has the following features: 1) Open; 2) High; 3) Low; 4) Close; 5) Stock Price Trend; 6) Polarity.

- Predicting Stock Price Trend with KNN Classifier

The final phase of this process involves predicting the stock class using the gathered feature set. To accomplish this, the data will be split into training and testing datasets, and the KNN (K-Nearest Neighbors) classifier will be utilized to forecast the stock price trend, which can be categorized as "rise," "fall," or "stable." The KNN classifier operates by classifying objects based on their proximity to the nearest training examples within the feature space. It assigns a class label based on the class of the K-closest instances in the training dataset. KNN belongs to the category of "lazy learner" classifier strategies. This means that the model doesn't make decisions about the training data until it is necessary to classify the test data. In other words, it defers the classification process until a new data point needs to be classified, which can be advantageous in various scenarios.

IV. RESULTS AND DISCUSSION

This section describes the results of experiment that was conducted to predict the stock market behavior using data mining and news sentiment analysis. The experiment consisted of two phases: the first phase analyzed news sentiment to classify news as positive or negative, and the second phase used the results of the sentiment analysis to predict whether the stock market would behave positively, negatively, or not affected. Three companies' data were used in both phases of the experiment: BBCA, BUMI, ASII.

The following sections describe the results of the experiments.

A. Result of News Sentiment Analysis Component

TABLE I. RESULT OF NEWS SENTIMENT ANALYSIS WITH NAÏVE BAYES CLASSIFIER

Stock	Prediction accuracy		
	Training and testing (without hyperparameter tuning)	Training and testing (with hyperparameter tuning)	Validation
BBCA	76.2%	91.9%	92.3%
BUMI	86.9%	94.6%	93%
ASII	82.2%	97.3%	96%

Table I. shows results of the proposed model achieved slightly higher accuracy than previous studies for stock market news sentiment analysis. The previous studies achieved accuracies in the range of 70% - 86.21% in the training and testing phase, while the experiment achieved 91.9% - 97.3% in the training and testing phase using hyperparameter tuning. As for the validation phase, the experiment achieved 92.3% - 96%. Using hyperparameter tuning was a significant improvement, and it suggests that the model is more accurate at predicting the sentiment of news. Previous studies didn't conduct validation phase to test the model with unseen data, therefore comparison is not doable. The accuracy of this model is quite good compared to previous studies that used Naive Bayes with NLP and TF-IDF techniques. The Naive Bayes algorithm provides quite good performance with high accuracy with text data. Compared to the Khedr, Salama, and Yaseen (2017) study, the accuracy of this sentiment analysis is slightly higher.

TABLE II. TUNED HYPERPARAMETER OF NEWS SENTIMENT ANALYSIS

Metrics	Training and Testing		
	BBCA	BUMI	ASII
Hyperparameter	alpha: 0.1 fit_prior = False	alpha: 0.01 fit_prior = True	alpha: 0.1 fit_prior = False

Table II. shows the tuned hyperparameter used by BBCA, BUMI, and ASII in the training and testing phase. These models were trained with varying hyperparameters, specifically focusing on the alpha parameter and the fit_prior parameter.

For the BBCA model, an alpha value of 0.1 was used, and the fit_prior parameter was set to False during training. This indicates that Laplace smoothing was employed, and prior probabilities were not considered. Conversely, the BUMI model was trained with an alpha value of 0.01, and fit_prior was set to True. This implies a smaller smoothing factor and the inclusion of prior probabilities in the model. Lastly, the ASII model was trained with an alpha value of 0.1 and fit_prior set to False, similar to the BBCA model, suggesting Laplace smoothing without prior probabilities.

B. Result of Stock Price Trend Prediction Component

The training and testing phase for the stock price trend prediction model is almost the same as what is done in the training and testing phase of sentiment analysis. GridSearchCV is used to find the best hyperparameters, but the model used is a KNN classifier. The training and testing phase for the stock price trend prediction model begins by combining the dataset of news polarity results from sentiment analysis with the historical stock price attribute data according to the period that has been predetermined. Then, the historical stock price attribute data will be scaled first using StandardScaler so that the stock price attribute pattern can be recognized by the model that uses a new dataset in the validation phase. After that, the model will be trained with the training data using GridSearchCV to find the best hyperparameters. Then, the model with the best hyperparameters will be obtained and stored for use in the validation phase.

TABLE III. RESULT OF FUTURE STOCK TREND PREDICITON WITH KNN CLASSIFIER

Stock	Prediction accuracy		
	Training and testing (without hyperparameter tuning)	Training and testing (with hyperparameter tuning)	Validation
BBCA	76.9%	86.2%	64.3%
BUMI	66.7%	72.2%	43.5%
ASII	84.4%	90%	64.6%

Table III. shows the accuracy of prediction when using the proposed model, which combines sentiment analysis and numeric data. The results of the study are consistent with the findings of other researchers, who have shown that there is a strong relationship between news and stock price changes. The study found that using sentiment analysis and numeric data for stock trend prediction produces accuracies ranging from 72.2% to 90% in the training and testing phase. However, in the validation testing, the prediction accuracy decreased to 43.5% - 64.6%.

In this step, the model is also tested to only predict the stock price trend using only historical stock price data in both training and testing, as well as validation phase. This step is done to learn and verify the contribution of the sentiment analysis in the stock price prediction method. In the training and testing stage, it is found that the stock price trend prediction without the sentiment analysis results in an average accuracy of 67.5%.

TABLE IV. TUNED HYPERPARAMETER AND KAPPA SCORE OF FUTURE STOCK TREND PREDICTION

Metrics	Training and Testing		
	BBCA	BUMI	ASII
Kappa	0.1709	0.0807	0.3969
Hyperparameter	n_neighbors: 3 p: 1 weights: distance	n_neighbors: 8 p: 1 weights: distance	n_neighbors: 5 p: 1 weights: distance
Metrics	**Validation**		
Kappa	0.491	0.470	**0**.342

Table IV. provides a comparison of metrics for three different models: BBCA, BUMI, and ASII. These models have been evaluated using various hyperparameter configurations during training and their performance on both training/testing and validation data sets.

The Kappa statistics in the training and testing phase is < 0.4. This means that the KNN algorithm shows moderate degrees of acceptance for stock trend prediction as 0.1709, 0.807, and 0.3969 are shown in the table. The validation phase, slightly better than previous phase, shows moderate-to-high degrees of acceptance as 0.491, 0.470, and 0.342 is shown on the table.

TABLE V. RESULT OF FUTURE STOCK TREND PREDICITON WITH KNN CLASSIFIER WITHOUT SENTIMENT ANALYSIS

Stock	Stock trend prediction accuracy, without sentiment analysis	
	Training and testing	Validation
BBCA	64.7%	60.7%
BUMI	63.8%	54.1%
ASII	74.2%	63.5%

Table V. shows the accuracy of prediction when using the proposed model, without sentiment analysis, and only historical numeric data. The results of the study proved that sentiment analysis contributes fairly to the stock trend prediction. The study found that using sentiment analysis alone for stock trend prediction produces accuracies ranging from 63.8% to 74.2% in the training and testing phase. However, in the validation testing, the prediction accuracy decreased to 54.1% - 63.5%.

However, the results of the study still demonstrate that the proposed model is an effective way to improve the prediction accuracy of stock trend prediction. The study also compared the accuracy of the proposed model to the accuracies of other studies that have used sentiment analysis to predict stock trends. The accuracy of this model in the training and testing phase is very good compared to previous studies, even the study by Khedr, Salama, and Yaseen (2017). The KNN algorithm provides quite good performance with high

accuracy with combined text and historical stock price data. This high accuracy is possible because the historical stock price data is scaled with StandardScaler so that the stock price values are in a certain range and make it easier for the model to group the data. Stocks like BUMI and ASII might not achieved the accuracy as high as the BBCA stock due to the number of financial news data that are much fewer than BBCA which are very active in the stock market and news.

V. CONCLUSION

In relation to the study objectives, this paper was successful in matching the methodology that was developed by Khedr, Salama, and Yaseen (2017), the analysis of historical stock price data and financial news sentiment, with stocks listed on the Indonesia Stock Exchange (IDX). In addition, this paper also managed to show that the methodology that was developed by Khedr, Salama, and Yaseen (2017) can predict price trends with approximately the same accuracy for stocks listed on the IDX.

There are two types of data that were used in this paper, daily financial news data and historical stock price data attributes. The daily financial news data that was used was taken from the bisnis.com news portal and the types are diverse. Starting from news that is relevant to stocks on the IDX, news about companies, or corporate actions that are carried out by the shares of companies that were analyzed in this paper (BBCA, BUMI, and ASII). Whereas the historical stock price data attributes were taken from Yahoo Finance. The data period that was used to become training and testing data started from January 1, 2021, to February 13, 2023. Whereas the data validation period that was used started from February 14, 2023, to June 26, 2023.

There are two main stages of the model processing that was proposed: 1) Training and testing stage; 2) Validation stage. The training and testing stage consists of 4 stages: 1) Determining the polarity of the news data, whether it is positive or negative using Naive Bayes classification; 2) Processing the historical stock price data attributes to determine the price pattern from the difference in closing prices each day; 3) Combining the results of the polarity from the first stage with the processed historical stock price data attributes to predict stock price trends with the KNN algorithm; 4) Performing hyperparameter tuning on the predictive model. After the model was trained with the training data, the model was stored and validated with data that had never been seen before. This study found that a method used to predict stock trends in the US (NASDAQ) can also be used to predict stock trends in Indonesia (IDX). This suggests that the method is generalizable and could be used to develop more accurate stock price prediction models for companies in other countries and industries.

The accuracy results from the training and testing stage of the third stock models that were used in this paper (BBCA, BUMI, and ASII) for analyzing news sentiment using Naive Bayes classification reached 97.3%, while at the validation stage it reached 96%. In the stock price trend prediction model, the training and testing stage produced an accuracy of 90%, while the validation stage produced an accuracy of 64.6% to predict the future behavior of the stock market.

The results of the model that was proposed show that there is a strong relationship between financial news that is related to stocks and changes in stock price attributes. The model that was proposed in this paper can be developed in the future by: 1) Combining the component analysis of stock technical indicators; 2) Using several news portals to collect financial news data; 3) Using training data from news or stock price attributes before COVID-19; 4) Using other machine learning models, such as neural networks; 5) Considering emotional sentences in determining news polarity and the influence of news that appears on social media; 6) Using a tool to determine the polarity of financial news without having to translate the news into English first.

REFERENCES

[1] Bing, Li, Keith C.C. Chan, and Carol Ou. “Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movements.” 2014 IEEE 11th International Conference on e-Business Engineering, 2014. https://doi.org/10.1109/icebe.2014.47.

[2] Cakra, Yahya Eru, and Bayu Distiawan Trisedya. “Stock Price Prediction Using Linear Regression Based on Sentiment Analysis.” 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2015. https://doi.org/10.1109/icacsis.2015.7415179.

[3] Khedr, Ayman E., S.E.Salama, and Nagwa Yaseen. “Predicting Stock Market Behavior Using Data Mining Technique and News Sentiment Analysis.” International Journal of Intelligent Systems and Applications 9, no. 7 (2017): 22–30. https://doi.org/10.5815/ijisa.2017.07.03.

[4] Pring, Martin J. Technical Analysis Explained. New York: McGraw-Hill Education, 2014.

[5] Bakshi, Rushlene Kaur, Kaur, Navneet, Kaur, Ravneet, Kaur, Gurpreet. "Opinion mining and sentiment analysis," 2016 IEEE 3rd International Conference on Computing for Sustainable Global Development (2016).

[6] Li, Xiaodong, Wang, Chao, Dong, Jiawei, Wang, Feng, Deng, Xiaotie, Zhu, Shanfeng. “Improving stock market prediction by integrating both market news and stock prices.” Springer 22nd International Conference on Database and Expert Systems Applications (2011).

[7] Liu, Bing. “Sentiment Analysis and Opinion Mining.” Morgan & Claypool Publishers (May 2012).

[8] Kim, Yoosin, Jeong, Seung Ryul, Ghani Imran. “Text Opinion Mining to Analyze News for Stock Market Prediction.” Int. J. Advance. Soft Comput. Appl., Vol. 6, No. 1 (March 2014)

[9] Wang, Wanbin (Walter), Ho, Kin-Yip, Liu, Wai-Man (Raymond), Wang, Kun (Tracy). “The relation between news events and stock price jump: an analysis based on neural network.” 20th International Congress on Modelling and Simulation (2013).

[10] Data Services Division, Indonesia Stock Exchange. “IDX Monthly Statistics.” Accessed October 29, 2022. https://www.idx.co.id/media/20221139/idx-monthly-september2022.pdf.

[11] Market capitalization of listed domestic companies (current US$). World Federation of Exchanges database. Accessed October 29, 2022. https://data.worldbank.org/indicator/CM.MKT.LCAP.CD.

[12] A. Mudinas, D. Zhang, and M. Levene, Market Trend Prediction using Sentiment Analysis: Lessons Learned and Paths Forward, Mar. 2019. doi:https://doi.org/10.48550/arXiv.1903.05440

[13] Q. Xiao and B. Ihnaini, “Stock trend prediction using sentiment analysis,” PeerJ Computer Science, vol. 9, 2023. doi:10.7717/peerj-cs.1293

[14] KSEI. “Statistik Pasar Modal Indonesia September 2022.” October 29, 2022.https://www.ksei.co.id/files/Statistik_Publik_-_September_2022_v5.pdf