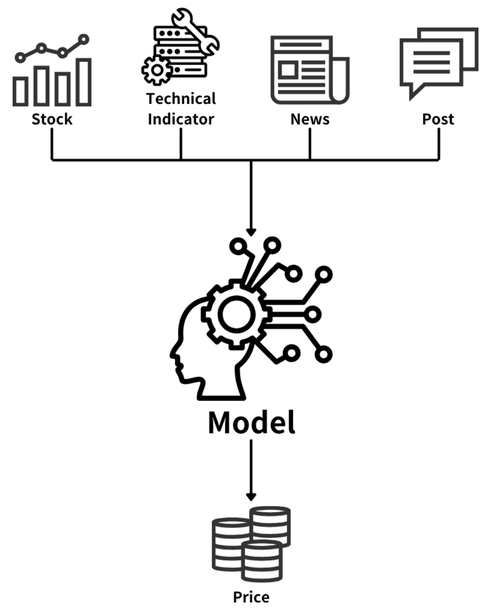
Improving Stock Price Prediction through News and Post Sentiment Analysis

Department of Mathematics, Ahn Tae Geon

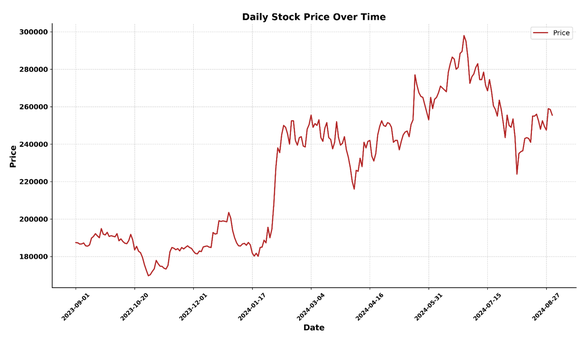
INTRODUCTIONS

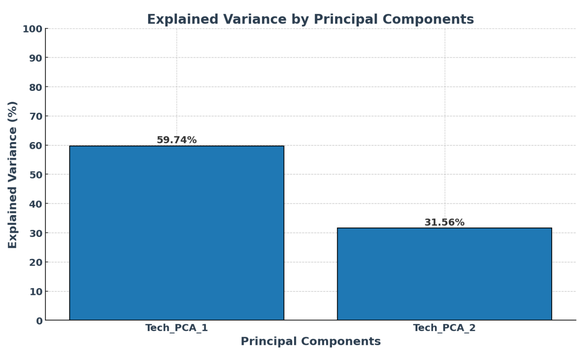


Research on predictive models in the stock market has been actively conducted, as stock prices are influenced by various factors. Particularly, text data such as news and posts reflect market sentiment and external factors, offering insights that are difficult to capture with traditional numerical data (e.g., Open, Close). In this study, we utilized commonly used **stock data** and **technical indicators** (e.g., RSI, EMA) along with **text data** to improve the performance of stock price prediction models. The primary objective of this research is to evaluate whether the inclusion of **text data** enhances model performance compared to using traditional variables alone. For this purpose, we compared three models: **Logistic Regression**, **Random Forest**, and **XGBoost**. **Text data** were summarized using **KoBERT**, a model developed by SKTBrain, and sentiment analysis was conducted using a fine-tuned **KoBERT** model trained on financial news articles.

METHODOLOGY

This study focuses on Hyundai Motor's stock, with data collected from September 1, 2023, to August 31, 2024. The data used is divided into four main categories: **Stock data**, **Technical Indicators**, **News data**, and **Post data**. **Stock data** included Open, Close, and Volume, with the dependent variable being the price change (**Change**) based on the closing price. The Change was labeled as 2 for a rise of 2% or more, 0 for a drop of 2% or more, and 1for all other cases, thus framing it as a multiclass classification problem. **Technical indicators** included RSI, H-band, L-band, EMA5, EMA20, MACD, and Signal. These indicators were reduced to two principal components(**Tech_PCA_1** and **Tech_PCA_2**) using PCA(Principal Component Analysis), explaining approximately 91% of the total variance.

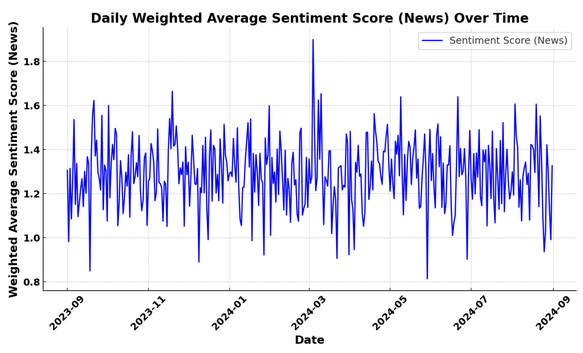


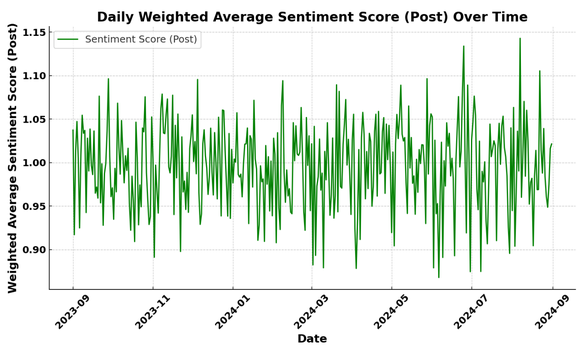


News data consisted of Hyundai Motor-related news articles uploaded on Naver, summarized using the KoBERT model developed by SKTBrain with max_length=128. **Post data** comprised titles of posts related to Hyundai Motor on Naver's stock discussion board. The **KoBERT model**, pre-trained on Korean Wikipedia and news sentences, was further fine-tuned with 4,800 financial news articles to classify text into three categories: positive, neutral, and negative. Sentiment analysis results were produced as Labels (0: negative, 1: neutral, 2: positive) and Logit values (model scores), which were aggregated daily using the following formula:

$$Weighted\ Average = \frac{\sum(Label \times Logit)}{\sum Logit}$$

The resulting **Weighted Average** value indicates sentiment, where a value close to 0 represents negative sentiment, and a value close to 2 represents positive sentiment.





All preprocessed data, including stock data, technical indicators, and sentiment scores, were consolidated, and the sentiment scores from weekends and holidays were merged with the closest subsequent trading day to maintain consistency. The consolidated data was standardized using **StandardScaler** to ensure a normal distribution. For modeling, we used three multi-class classification models: **Logistic Regression**, **Random Forest**, and **XGBoost**. The primary objectives of this study were (1) to compare which model achieves the highest performance by conducting hyperparameter optimization for each model, and (2) to evaluate whether including news and post sentiment data enhances model performance. For the second goal, the optimized parameters were kept the same while varying only the inclusion or exclusion of sentiment variables to compare their contributions.

CONCLUSIONS

In this study, we conducted a comprehensive parameter optimization for three distinct machine learning models : **Logistic Regression**, **Random Forest**, and **XGBoost**. The optimal hyperparameters for each model were determined as follows:

Model	Parameter	Value	Description
Logistic Regression	C	100	Regularization strength; smaller values specify stronger regularization.
	solver	lbfgs	Algorithm to use in the optimization process.
Random Forest	n_estimators	200	Number of trees in the forest.
	min_samples_split	5	Minimum number of samples required to split an internal node.
	min_samples_leaf	2	Minimum number of samples required to be at a leaf node.
	max_depth	None	Maximum depth of the tree; None means nodes are expanded until all leaves are pure.
XGBoost	subsample	0.6	Fraction of samples used for training each tree
	n_estimators	200	Number of boosting rounds (trees).
	max_depth	10	Maximum depth of a tree to control model complexity.
	learning_rate	0.01	Step size shrinkage to prevent overfitting.
	colsample_bytree	0.6	Fraction of features to use for building each tree.

The models were evaluated both with and without the inclusion of sentiment indicators, using accuracy and weighted average (W) as the evaluation metrics. Incorporating sentiment indicators yielded an overall improvement in model performance.

Sentiment Indicators	Metric	Logistc Regression	Random Forest	XGBoost
With	Accuracy	0.87	0.70	0.74
	Weighted Average	0.87	0.67	0.70
Without	Accuracy	0.82	0.72	0.72
	Weighted Average	0.82	0.69	0.69

The findings indicate that the inclusion of sentiment indicators significantly enhanced the predictive performance across all models, with **Logistic Regression** demonstrating the highest overall accuracy and weighted average (Accuracy: 0.87, W: 0.87) when sentiment indicators were included. In comparison, the **Random Forest** model did not show improvement, while the **XGBoost** model exhibited less pronounced improvement, although both benefitted from the sentiment indicators' inclusion. These results suggest that **sentiment information** provides a meaningful contribution to the predictive capabilities of the models, particularly for **Logistic Regression**, which emerged as the most effective approach in this context.

LIMITATIONS

In previous studies, attempts have been made to analyze market sentiment using news data and apply it to stock price prediction models. However, news data often has a high proportion of neutral articles, which limits the accuracy of sentiment analysis and the predictive performance of the models. To address this limitation, this study simultaneously utilizes news data and post data that reflect the direct reactions of investors. While news data captures macroeconomic external factors, post data includes the immediate emotions and opinions of investors. The combination of these two data sources creates a synergistic effect, enhancing the granularity of sentiment analysis and improving the performance of the prediction models. Despite these promising results, this study has certain limitations that warrant further exploration. Firstly, only one year of data was used in this study, but extending the period to utilize more data is expected to yield better results. Secondly, focusing on multiple stocks instead of a single stock could produce more statistically significant findings. Selecting stocks from various sectors would be particularly beneficial in capturing a broader range of market dynamics. Additionally, increasing the data volume to incorporate deep learning-based models beyond machine learning models could further enhance the predictive performance and provide deeper insights.