IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 10, 2021, accepted March 16, 2021, date of publication March 22, 2021, date of current version March 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3067691

Equity Research Report-Driven Investment Strategy in Korea Using Binary Classification on Stock Price Direction

POONGJIN CHO[1], JI HWAN PARK[2], AND JAE WOOK SONG[2]
[1]KRiBLE, FnGuide, Seoul 07805, Republic of Korea
[2]Department of Industrial Engineering, Hanyang University, Seoul 04763, Republic of Korea
Corresponding author: Jae Wook Song (jwsong@hanyang.ac.kr)

ABSTRACT This research examines and proposes an investment strategy by combining the natural language processing on the equity research reports published in the Korean financial market and machine learning algorithms for binary classification. At first, we deduce the part-of-speech from the report using the KoNLPy and Mecab. Then, we define 33 features as the input variables and perform the binary classification on the price direction of the stocks recommended in the report using various machine learning algorithms. Note that we investigate the model performance in detail by dividing the entire period into three sub-periods, including pre-COVID-19 for the sideways market, COVID-19 for the crashing market, and post-COVID-19 for the extreme bullish market. We confirm that the random forest is the best classifier for all periods, so we utilize its results on positively predicted stocks in the test set as the investment universe for the monthly re-balancing and buy-and-hold investment. The proposed strategy shows a significantly higher return on investment than benchmarks during the pre-COVID-19 and COVID-19 periods, whereas the comparable return during the post-COVID-19.

INDEX TERMS Finance, natural language processing, stock markets, equity research reports, binary classification, investment strategy.

I. INTRODUCTION

Financial companies periodically issue research reports for investors. The report contents include analyzing companies, financial institutions, diplomatic issues between countries, and politics. Among them, this study focuses on the equity research report that recommends a specific stock at a time. Usually, analysts write their perspective on a stock expected to show high returns in the future through various quantitative and qualitative analyses. However, the profit in the future varies in different reports. One reason for such a result is that the person who writes the report may not be equipped with enough analytical skills, extending to low-quality reports. In this study, we assume that the composition of the equity research reports quantified through natural language processing (NLP) can distinguish the stock recommendations' reliability.

In the 2010s, the digital online content volume has exploded, including market analysis reports, news articles, journal texts, online blogs, and social media. Accordingly, research on analyzing public sentiment, especially opinion mining in social media, has become essential. As the market prediction using NLP algorithms has been studied in the financial field, a research field called natural language-based financial forecasting has been gradually established [1]–[4]. In particular, the stock market has received great attention in academia due to its sensitivity to market participants' sentiment. That is, investors' sentiment can change the overall trend of individual stocks and even the market. Many previous studies have analyzed investors' opinions and market sentiment from social media posts regarding financial markets. Some studies extract and analyze the mood of text from social media such as Twitter [5]–[9], news [10], bulletin board [11], [12] and utilize them for market prediction. This research's main objective focuses on a binary classification of positive or negative moods using the text to discover its correlation with the market or company's stock price movement. The methods that have been used for analysis include the Naive Bayes approach [12]–[16], support vector machine [17]–[20], and decision tree [21].

The associate editor coordinating the review of this manuscript and approving it for publication was Chun-Hao Chen.

In addition to the same machine learning algorithm, many studies also have utilized various deep learning algorithms such as Artificial Neural Network [22]–[24] and Recurrent Neural Network [25]–[27]. Depending on how the features of the text are extracted, it may not reflect the movement of stock price well. Hence, the extraction of features representing the characteristics of natural language is an important topic in NLP. The methods of the feature engineering incorporate linguistic feature [28], [29], keyword extraction [30], data reduction with generative probabilistic model [31], and word embedding with n-grams [28], [32], TF-IDF [33], ensemble model [34] and deep learning [35].

Furthermore, there have been many studies regarding the derivation of investment strategies using NLP. Several studies have analyzed market sentiment information using a machine-learning algorithm to construct a portfolio. Such studies either design a neural network with an ensemble of evolving clustering and LSTM [36], or propose a new follow-the-loser portfolio strategy from the post of stock micro-blogs using semi-supervised learning method [37], or establish a trading strategy from new sentiment data using learning-to-rank algorithms [38]. Also, recently, a portfolio investment strategy that considers shareholders' confidence index by combining the existing random forest and sentimental analysis [39] and an investment strategy that encodes external information from financial news using reinforcement learning have been proposed [40].

However, there have been limited efforts on establishing an investment strategy based on the NLP of equity research reports published in the Korean financial market. Therefore, in this paper, we focus on analyzing the report through NLP and investigate if the induced information can be utilized for investment strategy. At first, the NLP element is derived by quantifying the structure of the report in the form of part-of-speech (POS). Then, using NLP elements as input features, a binary classification model that predicts whether the stocks recommended from the report produce the positive or negative return is constructed. The model with the best classification performance is selected for the experiment by applying several machine learning algorithms. Finally, we propose an investment strategy to buy stocks predicted to yield a positive return in future returns through the suggested classification algorithm. To show the superiority of the proposed investment strategy, we compare its investment returns with the strategy of investing all the stocks recommended by the report and the market index as benchmarks. Besides, to investigate whether the proposed investment strategy shows consistent performance in various market conditions, different periods' investment return is analyzed separately.

II. FRAMEWORK OF INVESTMENT STRATEGY

A. EQUITY RESEARCH REPORT

This study utilizes 34,780 equity research reports on stocks traded in Korean financial markets published from 2019-01-01 to 2020-06-12. Note that the securities firm analysts are in charge of writing the reports and provide them in Portable Document Format (PDF) on the firm's website. Each report recommends one stock at a time, providing information on the target price, current status and direction on the underlying company's business, and degree of recommendation on the stock. Note that the number of reports varies in different months as summarized in Table 1.

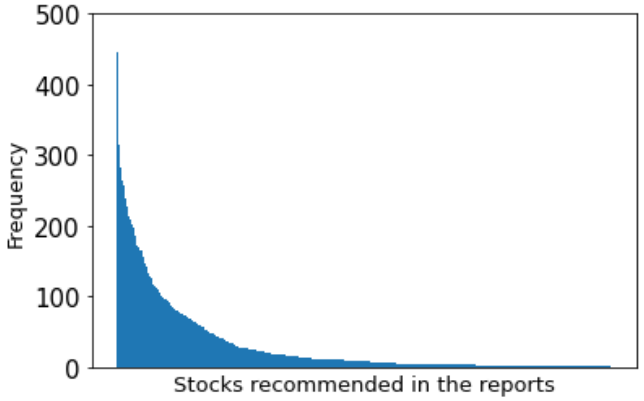


FIGURE 1. Distribution of recommendation frequencies on each stock.

A total of 1,118 stocks were recommended within 34,780 research reports where the reports are concentrated on a limited number of stocks as presented in Figure 1. Specifically, the top 400 stocks account for 90% of all reports. One reason for such concentration could be the investor's preference on stocks with high market capitalization or thematic investing, which can promote readability and click rates on reports in favor of analysts. Due to less diversity among the reports, many investors doubt the report's utility in predicting the future returns on the underlying stock. However, it is also true that most of the reports are carefully written through sufficient research, consisting of the analyst's sentimental but reasonable opinions on the company. In this context, we assume that the valuable reports with rich information are written based on clear facts, whose composition and even sentence structure could differ from the unhelpful report with limited value. At first, we define that a report is valuable if it successfully recommends a stock with a positive return in the near future. A recommended stock from an unhelpful report yields a negative return in the near future. Then, we conduct a binary classification based on NLP and various machine learning algorithms to distinguish the composition of valuable reports.

B. FEATURE ENGINEERING WITH NLP

We utilize the NLP to define the features for binary classification. At first, the contents of each report written in Korean are divided into morpheme units through NLP. English has its meaning decomposed based on spacing, but the Korean can be divided into morphemes containing two or more meanings without spacing. To analyze the Korean language, we employ the KoNLPy [41], a Python package for NLP of the Korean language, and Mecab [42], methods of tagging POS that tags

TABLE 1. Number of equity research report per month.

Months	Jan-2019	Feb-2019	Mar-2019	Apr-2019	May-2019	Jun-2019	Jul-2019	Aug-2019	Sep-2019
Number of reports	600	470	350	607	685	303	602	556	280
Months	Oct-2019	Nov-2019	Dec-2019	Jan-2020	Feb-2020	Mar-2020	Apr-2020	May-2020	Jun-2020
Number of reports	644	702	179	468	492	364	628	593	173

TABLE 2. Selected 10 POS from 45 POS.

Selected POS	Number of Integration	List of Integrated POS
Noun	6	Common, Proper, Bound, Unit, Numerals, Pronoun
Adjective	1	Adjective
Verb	1	Verb
Adverb	2	Common, Connective
Postposition	9	Subjective, Auxiliary, Connective, Complement case, Adnominal case, Objective case, Adverbial case, Vocative case, Citing case
Determiner	1	Determiner
Ending	4	Sentence-closing, Connective, Nominal, Adnominal
Number	1	Number
English	1	English
Others	1	Auxiliary Predicate element
Not-in-use	16	Exclamation, Uninflected prefix, Noun-driven suffix, Verb-driven suffix, Adjective-driven suffix, Root, Period & Marks, Ellipsis, Opening parenthesis, Closing parenthesis, Delimiter, Dash, Symbol, Chinese character Positive copula word, Negative copula word

each morpheme with 43 detailed POS. However, 43 POS divides the sentence in too much detail and has many features, which can cause overfitting. Therefore, in this study, the top 10 most used NLP elements are integrated and selected as summarized in Table 2.

Based on the ten selected POS, we utilize each POS frequency as a feature for the binary classification. Then, we create eight additional features that can represent the characteristic of equity research report: Number of a morpheme (subwords), Average number of morpheme per sentence (mean_subwords_per_sentence), Standard deviation of morpheme per sentence(std_subwords_per_sentence), Number of sentences ending with *da* (da), Number of sentences (sentence), Number of paragraphs (paragraph), Number of pages (page), Number of pages with words (page_with_word). Note that we use optical character recognition (OCR) to count the number of pages with words since there exist pages with only tables or pictures. In Korean, a perfect sentence ends with *da*; otherwise, a sentence has omitted elements. Note that a sentence that does not end with *da* only conveys some financial terms providing limited implication to the investment. Therefore, we assume that the *da* can be a feature representing an equity research report's characteristic. In Figure 2a, the distributions of the ten selected POS and additional features are investigated, showing that all features are skewed. Therefore, we apply log-transformation to all variables as illustrated in Figure 2b. In this context, we successfully obtain variables whose distributions are close to the normal distribution used in machine learning for binary classification.

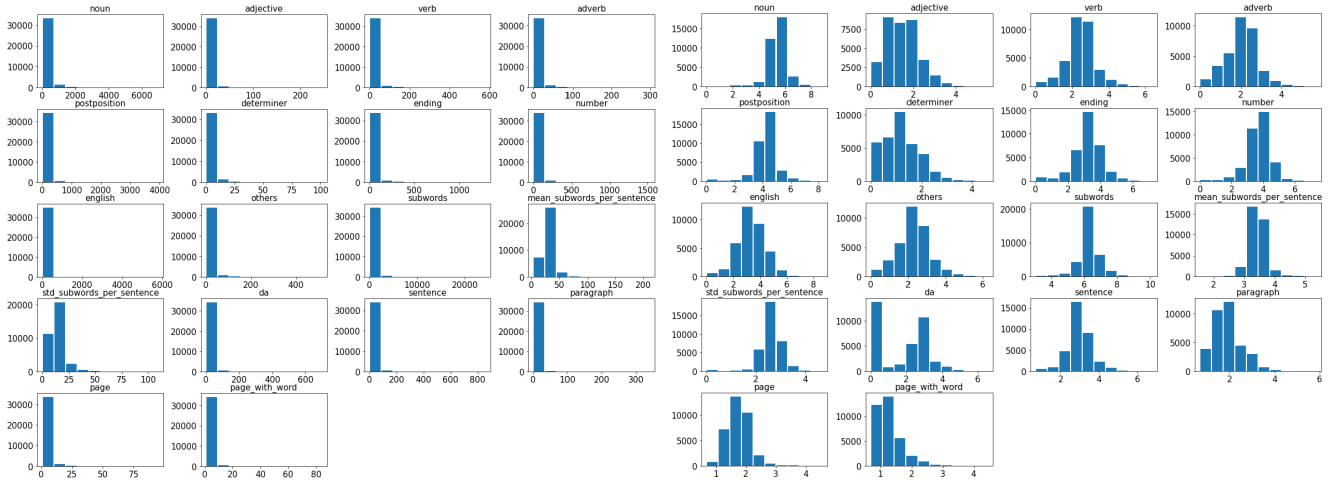
In addition, we include 15 different ratios based on the selected POS and additional features as follows: Noun ratio (noun/subwords), Adjective ratio (adjective/subwords), Verb ratio (verb/subwords), Adverb ratio (adverb/subwords), Postposition ratio (postposition/subwords), Determiner ratio (determiner/subwords), Ending ratio (ending/subwords), Number ratio (number/subwords), English ratio (english/subwords), Others ratio (others/subwords), Ratio of *da* in sentences (da/sentence), Changes in number of morpheme (std_subwords_per_sentence/mean_subwords_per_sentence), Morpheme per page (subwords/page), Morpheme per sentence (subwords/sentence), Ratio of pages with words (page_with_word/page). Note that we also apply the log-transformation to the ratios. Finally, we apply the Min-max scaling to total 33 features to finalize the data pre-processing for the binary classification.

C. BINARY CLASSIFICATION & INVESTMENT STRATEGY

We propose a binary classification based on the pre-processed NLP-driven features, which predicts whether or not the stock suggested in the equity report will show a positive or negative return in the future. Specifically, we utilize five well-known models. At first, we employ the k-Nearest Neighbors (k-NN) classifier. The k-NN algorithm hinges on the assumption that similar data points will be located at close distance [43]. Therefore, it calculates the distance between the test data and the input, which can be obtained as follows:

$$d(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{1}$$

where $\boldsymbol{p}$ and $\boldsymbol{q}$ refer to the data points that have coordinates of $(p_1, p_2, \ldots, p_n)$ and $(q_1, q_2, \ldots, q_n)$ in n dimensions, respectively.



(a) Original features (b) Log-transformed features

FIGURE 2. **Distributions of selected POS and additional features based on NLP.**

Secondly, we utilize the logistic regression using the sigmoid function as follows [44]:

$$cost(W) = \frac{1}{m}\sum c(H(x), y) \tag{2}$$

$$c(H(x), y) = \begin{cases} -\log(H(x)), & : y = 1 \\ -\log(1 - H(x)), & : y = 0 \end{cases} \tag{3}$$

$$H(x) = \frac{1}{1 + e^{-(Wx+b)}} \tag{4}$$

where $H(x)$, W and b correspond to the sigmoid function, weight, and bias, respectively. As a result of approaches 1 or 0, the value of the cost function decreases or increases, respectively.

Thirdly, we utilize the decision tree, which analyzes and represents patterns between data as a combination of possible rules and is built top-down from the root node [45]. To build a decision tree, we use the entropy for an area to which m data points belong can be calculated as follows:

$$Entropy = -\sum_{k=1}^{m} p_k \log_2(p_k) \tag{5}$$

where p_k refers to the percentage of the data points belonging to the category k. It is trained to increase the homogeneity of each area and reduce the impurity or uncertainty as much as possible, which is called information gain.

Fourthly, we utilize the random forest. Since the decision tree has a limitation of overfitting, we employ an ensemble model that generates multiple decision trees and votes on each tree's classification results. It can be obtained through bagging that makes a decision tree with data sampled with replacement from the entire training data [46].

Lastly, we utilize gradient boosting, an ensemble model that produces a robust classifier by combining weak classifiers, typically decision trees [47]. It uses gradient descent to differentiate the loss function as a parameter to obtain the slope and calibrates the parameter so that the loss decreases. The loss function and the negative gradient are expressed as follows.

$$L(y_i, f(x_i)) = \frac{1}{2}(y_i - f(x_i))^2 \tag{6}$$

$$\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} = \frac{\partial[\frac{1}{2}(y_i - f(x_i))^2]}{\partial f(x_i)} = f(x_i) - y_i \tag{7}$$

where L refers to the loss function.

For the experiment, we divide the data into the train(70%) and test(30%) sets. Note that we ensure the partitioned data can carry the equivalent distributional characteristics of the number of equity research reports per month as well as the number of those per stock. Although many prediction problems in financial time-series use the in-sample and out-of-sample on time, our model can utilize random sampling since its explanatory variables are not dependent on time. For 50 different random seeds, we compare the classification performances of five models for different times after the report's release. Based on the model with the best performance, we simulate the backtesting with monthly re-balancing and simple buy-and-hold for different investment horizon investment strategies using the positively predicted stocks in the test set. Then, we compare the investment performance with other benchmarks. A step-by-step scenario of the proposed investment strategy is illustrated in Figure 3

III. EMPIRICAL RESULTS AND DISCUSSIONS

A. BINARY CLASSIFICATION PERFORMANCE

As previously stated, we utilize five machine learning models to predict the price direction of the stock recommended from the equity research reports whose NLP elements are considered as the features. Table 3 summarizes the hyper-parameters for each model. We compare the binary classification performances of each model for a different time in the future in

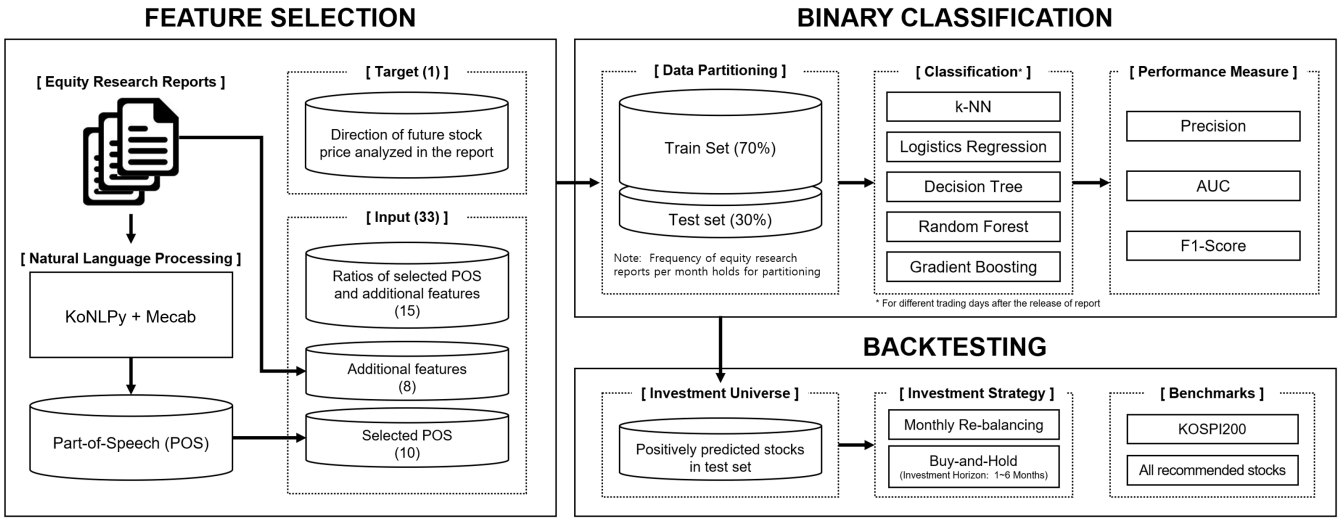


FIGURE 3. Proposed investment framework using NLP-driven features from the equity research report.

TABLE 3. Hyper-parameters of each machine learning algorithms for binary classification.

Machine learning	Hyper-parameters
k-Nearest Neighbors	Metric(Manhattan), Number-of-neighbors(19), Weights(distance)
Logistics Regression	C(100), Penalty(L2), Solver(lbfgs)
Decision Tree	Min-impurity-decrease(0.2), Max-depth(320), Max-features(Sqrt), Min-samples-leaf(3)
Random Forest	Number-of-estimators(1200), Criterion(Gini), Max-depth(460), Max-features(Sqrt), Min-samples-leaf(2)
Gradient Boosting	Number-of-estimators(1000), Learning-rate(0.05), Sub-sample(0.75)

terms of prediction accuracy and area under the receiver operating characteristic curve (AUC). Note that this research's main objective is to examine if the equity research report's NLP elements can be used to construct an investment strategy. Therefore, based on two simple measures, we select a model with the highest classification performance, analyze the classification results in detail using the precision, recall, and F1-score, and utilize it to establish an investment strategy.

The models predict the direction of stock at 30, 60, 90, 120, 150, and 180 trading days after the report's release. We will call this as prediction time. We consider the equity research reports published from 2019-01-01 to 2020-06-12, and the Korean financial market has experienced the sideways period with low volatility (2019-01-01 - 2020-01-20), collapsing period due to the outbreak of COVID-19 (2020-01-21 2020-03-29), and soaring period with the extreme bullish market (2020-03-30 2020-06-30). Specifically, we divide the periods based on the highest and lowest points of KOSPI200, the representative financial market index of Korea, within the entire period. In this regard, the classification performance can be evaluated for different market conditions.

At first, the average classification performances of each model for 50 different random seeds are summarized in Table 4. According to the results, the accuracy and AUC tend to increase as the prediction time increases for all models. It implies that a higher return can be expected when an investment strategy is established based on the long investment horizon's prediction results. Finally, we choose the random forest as the primary classification model since it shows the highest accuracy and AUC for all prediction times.

Detailed classification performance of the random forest is summarized in Table 5. Comparing to the accuracy and AUC, the F1-score is low and invariant for different prediction times, which reduces the utility of the prediction model. Specifically, the low F1-score is caused by the relatively low recall. Note that the precision shares the same pattern as the accuracy and AUC. However, such a result does not affect the random forest's utility since the proposed investment strategy only utilizes the positively predicted stocks, whose return in the future is expected to be positive. A classification model with high precision but low recall in a binary classification indicates relatively lower false positives than false negatives. In this context, the stocks predicted to be positive are likely to be in the actual positive direction, although the model cannot accurately detect all stocks with positive direction. Therefore, we can imply that an investment strategy based on the stocks predicted to be positive returns can produce a high profit.

We further investigate how random forest classification varies in different market conditions as summarized in Table 6. For the reports published during the sideways period, the accuracy increases as the investment period increases, but the AUC remains around 0.5. Hence, the corresponding investment strategy is expected to produce a little

TABLE 4. Average classification performances of each machine learning algorithms for different prediction times.

Machine learning	Prediction time (Trading days) 30	60	90	120	150	180
(a) Accuracy						
k-Nearest Neighbors	0.5504	0.5581	0.5706	0.5959	0.6140	0.6252
Logistics Regression	0.5665	0.5800	0.6080	0.6385	0.6535	0.6506
Decision Tree	0.5279	0.5457	0.5416	0.5507	0.5593	0.5684
Random forest	0.5750	0.5943	0.6196	0.6460	0.6555	0.6609
Gradient Boosting	0.5706	0.5880	0.6147	0.6410	0.6490	0.6532
(b) AUC						
k-Nearest Neighbors	0.5492	0.5553	0.5633	0.5735	0.5860	0.6030
Logistics Regression	0.5619	0.5692	0.5840	0.5941	0.6021	0.6115
Decision Tree	0.5338	0.5337	0.5364	0.5421	0.5462	0.5622
Random Forest	0.5760	0.5913	0.5972	0.6022	0.6097	0.6256
Gradient Boosting	0.5656	0.5752	0.5900	0.5968	0.5974	0.6136

TABLE 5. Average precision and recall of random forest for different prediction times.

Performance measures	Prediction time (Trading days) 30	60	90	120	150	180
F1-Score	0.5065	0.5111	0.4818	0.4507	0.4667	0.5038
Precision	0.5922	0.6055	0.6242	0.6403	0.6233	0.6501
Recall	0.4424	0.4421	0.3923	0.3477	0.3730	0.4112

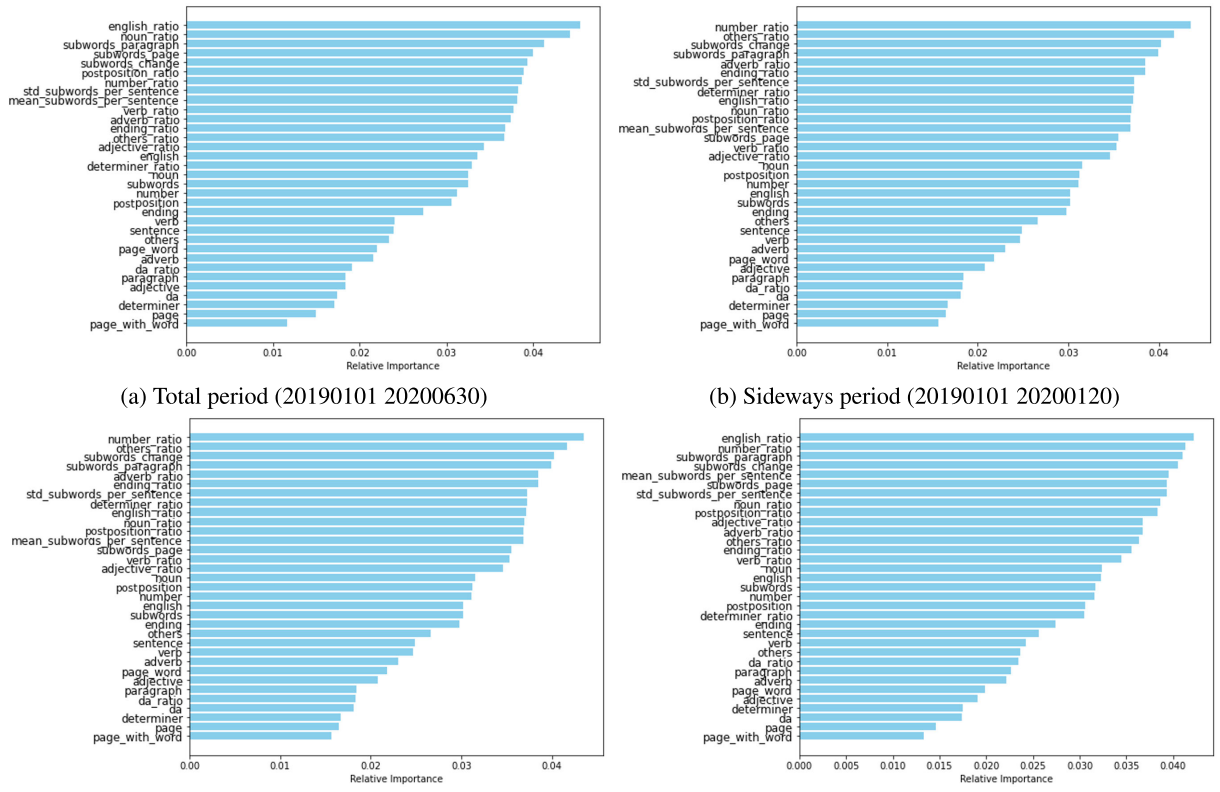
TABLE 6. Average classification performance of random forest for different market periods.

Performance measures	Prediction time (Trading days) 30	60	90	120	150	180
(a) Sideways period						
Accuracy	0.5330	0.5603	0.6026	0.6493	0.6696	0.6680
AUC	0.5178	0.5271	0.5225	0.4146	0.5181	0.5212
F1-score	0.3771	0.3500	0.2597	0.1693	0.1805	0.1912
Precision	0.4955	0.4898	0.4510	0.4095	0.3958	0.4107
Recall	0.3044	0.2723	0.1824	0.1067	0.1169	0.1246
(b) Collapsing period						
Accuracy	0.7099	0.6984	0.6044	0.5665	0.5608	0.5611
AUC	0.5393	0.5467	0.5594	0.5439	0.5537	0.5708
F1-score	0.2624	0.2873	0.3599	0.3636	0.4489	0.5301
Precision	0.2795	0.3533	0.5740	0.5567	0.5652	0.6334
Recall	0.2473	0.2420	0.2622	0.2700	0.3723	0.4557
(c) Soaring period						
Accuracy	0.6452	0.6748	0.6909	0.6645	0.6414	0.6992
AUC	0.4904	0.5104	0.5045	0.5014	0.4885	0.5064
F1-score	0.7723	0.7945	0.8089	0.7869	0.7690	0.8155
Precision	0.7358	0.7611	0.7718	0.7603	0.7356	0.7751
Recall	0.8125	0.8309	0.8497	0.8156	0.8057	0.8602

advantage over investing in all reports. During the collapsing period, both AUC and F1-score increase as the prediction time increases. Hence, the corresponding investment strategy is expected to produce a high profit over investment in all reports. During the soaring period, we observe a high accuracy and F1-score but relatively low AUC values. The high accuracy and F1-score are realized due to the biased target variable on the positive direction during the recovery from the COVID-19 pandemic shock. Therefore, the corresponding investment strategy is expected to show no significantly different profit compared to the investment in all reports.

B. FEATURE IMPORTANCE

Prior to utilizing the binary classification into the investment strategy, we investigate the feature importance based on random forest results. The average importance of each NLP element in the random forest is summarized in Figure 4. For the total period in Figure 4a, the most significant feature is the English ratio. Note that the low feature importance indicates no significant influence on predicting the direction of the stock price. Specifically, based on the median of the English ratio, the average investment return for 180 prediction time for all reports with an English ratio lower than the median is -2.3%, whereas that for all reports with an English ratio



(a) Total period (20190101 20200630)

(b) Sideways period (20190101 20200120)

(c) Collapsing period (20200121 20200329)

(d) Soaring period (20200330 20200630)

FIGURE 4. Feature importance for different market periods.

higher than the median is 7.8%, which yields the difference of 10.1% of the return. It implies that a relatively high English ratio report can be expected to show a positive expected return compared to a report that does not have one. Likewise, the noun ratio, the second most crucial variable, shows a 7.2% difference in investment return based on the median. In this context, we discover the NLP-elements that positively affect the investment return, which are English ratio, subwords per page, page word, and subwords, among the top 15 features showing high importance. Otherwise, for most NLP-elements, the lower the value, the higher the investment return. Interestingly, most of the ratios of NLP-elements show high feature importance than selected POS and additional features in Figure 3.

Furthermore, we examine the feature importance of NLP elements for different market conditions in Figures 4b,4c and 4d. Analogous to the total period, the ratios of NLP-elements show high feature importance in all periods. Therefore, we can conclude that the ratios of NLP-elements play a more important role than basic NLP elements regardless of market conditions except for the determiner ratio and ending ratio. Also, subwords per page and number ratio show high feature importance regardless of market conditions. Note that the higher the subwords per page, the higher the investment return, while the lower the number ratio, the higher the investment return.

C. INVESTMENT PERFORMANCE

Finally, we perform the backtesting of two investment strategies based on the positively predicted stocks in the test set from the 50 random seeds as the investment universe. The first strategy is the monthly-rebalancing. At first, we take a long position on the positively predicted stocks on the test set with equal weight. Then, after a month, we sell all the stocks purchased and repeat the process of taking long position. Figure 5 shows the monthly average cumulative rate of return. The proposed strategy is a blue line, and the monthly cumulative returns of KOSPI200 and all stocks recommended from the equity research reports in the test set are provided as benchmarks with sky blue and gray lines, respectively. Note that the vertical lines indicate the three standard deviations of cumulative returns for each month. The result shows that the proposed strategy outperforms the returns of other benchmarks. Besides, the strategy of buying all the stocks recommended by the report slightly exceeds the KOSPI index, which ensures some degree of the reliability of the equity research report on recommending stocks.

In order to compensate for the limitation of the cumulative return, the average return on investment in different market conditions based on a buy-and-hold strategy is summarized in Table 7 for different investment horizons from 30 days to 180 days. The proposed investment strategy yields significantly higher returns for the total period than the benchmarks

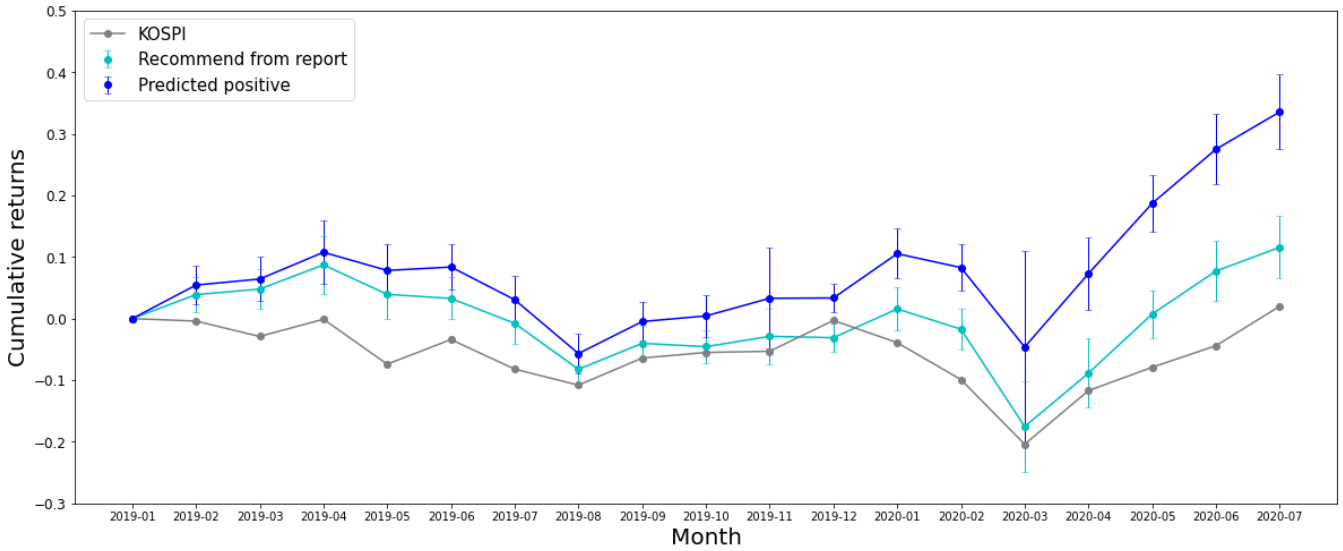


FIGURE 5. Cumulative investment return with monthly re-balancing strategy.

TABLE 7. Investment returns of predicted & all stocks from the equity research reports for different investment horizons.

Investment returns	Investment horizon (Trading days) 30	60	90	120	150	180
(a) Total period						
Return of recommended from report (%)	0.60	0.98	2.00	1.58	1.26	2.57
Return of predicted positive (%)	3.82	6.93	11.87	17.53	18.74	21.21
(b) Sideways period						
Return of recommended from report (%)	-0.13	-1.05	-2.61	-5.34	-6.56	-6.35
Return of predicted positive (%)	0.80	0.69	0.11	-2.60	0.05	0.51
(c) Collapsing period						
Return of recommended from report (%)	-9.95	-9.93	0.62	7.70	12.42	19.55
Return of predicted positive (%)	-4.78	-2.53	12.05	21.53	23.76	30.74
(d) Soaring period						
Return of recommended from report (%)	9.44	15.29	21.33	26.05	26.51	28.99
Return of predicted positive (%)	8.71	14.91	20.99	26.51	26.72	29.29

invested in all stocks recommended by the report for all investment horizons. Also, the difference in returns between the two investment strategies increases as the investment period increases. During the sideways period, the proposed investment strategy shows slightly better returns than the benchmark. However, the equity research report published in the sideways period includes a collapsing period on the long-term investment horizon. Despite the sharp decline in the market, the proposed strategy does not record negative returns except for the investment horizon of 120 trading days, which is very encouraging. During the collapsing period, it yields significantly higher returns than the benchmark for all investment horizons. In particular, since the long-term investment horizon includes a soaring period, the proposed investment strategy can be considered to possess an ability to detect stocks whose prices will rise rapidly during the recovery of a financial market after the market crash. Finally, during the soaring period, the presented model shows a similar investment return as the benchmark.

IV. CONCLUSION

Throughout this research, we explore the possibility of developing an investment framework using a binary classification based on NLP-elements of the equity research report. To the best of our knowledge, this is the first attempt to utilize the NLP-elements of the equity research report in Korea to establish investment strategies. Therefore, this research's novelty lies in providing the possible integration of NLP-elements of the equity research report in stock investment. Through the experiments, the random forest shows the best classification performance whose AUC of the random forest during the sideways period and the collapsing period is higher than 0.5. Therefore, we select the random forest as the binary classification algorithm. Then, we perform the backtesting based on classification results for monthly re-balancing and buy-and-hold for different investment horizons. As a result, we confirm that the proposed investment strategy generates higher returns than the benchmark during the sideways period and collapsing period. In an extreme bull market, selecting

stocks with high expected return does not make much of a difference since any stock an investor chooses will yield a high return. However, an investment strategy that helps select stocks with a high return in the future during sideways or bearish markets has a significant implication in real-world investment practice. Therefore, for further research, we plan to utilize various portfolio theories in constructing efficient investment strategies rather than simple buy-and-hold by using the positively predicted stocks from the binary classification.

REFERENCES

[1] P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy, "More than words: Quantifying language to measure Firms' fundamentals," *J. Finance*, vol. 63, no. 3, pp. 1437–1467, Jun. 2008.

[2] R. P. Schumaker and H. Chen, "A discrete stock price prediction engine based on financial news," *Computer*, vol. 43, no. 1, pp. 51–56, Jan. 2010.

[3] R. P. Schumaker, "Analyzing parts of speech and their impact on stock price," *Commun. IIMA*, vol. 10, no. 3, p. 1, 2010.

[4] F. Z. Xing, E. Cambria, and R. E. Welsch, "Natural language based financial forecasting: A survey," *Artif. Intell. Rev.*, vol. 50, no. 1, pp. 49–73, Jun. 2018.

[5] J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng, "Exploiting topic based Twitter sentiment for stock prediction," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2013, pp. 24–29.

[6] J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič, "Stream-based active learning for sentiment analysis in the financial domain," *Inf. Sci.*, vol. 285, pp. 181–203, Nov. 2014.

[7] G. Ranco, D. Aleksovski, G. Caldarelli, M. Grčar, and I. Mozetič, "The effects of Twitter sentiment on stock price returns," *PLoS ONE*, vol. 10, no. 9, Sep. 2015, Art. no. e0138441.

[8] A. Papana, C. Kyrtsou, D. Kugiumtzis, and C. Diks, "Detecting causality in non-stationary time series using partial symbolic transfer entropy: Evidence in financial data," *Comput. Econ.*, vol. 47, no. 3, pp. 341–365, Mar. 2016.

[9] A. Tafti, R. Zotti, and W. Jank, "Real-time diffusion of information on Twitter and the financial markets," *PLoS ONE*, vol. 11, no. 8, Aug. 2016, Art. no. e0159226.

[10] R. Akita, A. Yoshihara, T. Matsubara, and K. Uehara, "Deep learning for stock prediction using numerical and textual information," in *Proc. IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2016, pp. 1–6.

[11] S. R. Das and M. Y. Chen, "Yahoo! For Amazon: Sentiment extraction from small talk on the Web," *Manage. Sci.*, vol. 53, no. 9, pp. 1375–1388, Sep. 2007.

[12] T. H. Nguyen and K. Shirai, "Topic modeling based sentiment analysis on social media for stock market prediction," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process. (Long Papers)*, vol. 1, 2015, pp. 1354–1364.

[13] W. Antweiler and M. Z. Frank, "Is all that talk just noise? The information content of Internet stock message boards," *J. Finance*, vol. 59, no. 3, pp. 1259–1294, Jun. 2004.

[14] F. Li, "The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach," *J. Accounting Res.*, vol. 48, no. 5, pp. 1049–1102, 2010.

[15] N. Jegadeesh and D. Wu, "Word power: A new approach for content analysis," *J. Financial Econ.*, vol. 110, no. 3, pp. 712–729, Dec. 2013.

[16] A. H. Huang, A. Y. Zang, and R. Zheng, "Evidence on the information content of text in analyst reports," *Accounting Rev.*, vol. 89, no. 6, pp. 2151–2180, Nov. 2014.

[17] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," *Knowl.-Based Syst.*, vol. 69, pp. 14–23, Oct. 2014.

[18] F. Xu and V. Keelj, "Collective sentiment mining of microblogs in 24-hour stock price movement prediction," in *Proc. IEEE 16th Conf. Bus. Informat.*, Jul. 2014, pp. 60–67.

[19] Y. Xie and H. Jiang, "Stock market forecasting based on text mining technology: A support vector machine method," 2019, *arXiv:1909.12789*. [Online]. Available: http://arxiv.org/abs/1909.12789

[20] W. Long, L. Song, and Y. Tian, "A new graphic kernel method of stock price trend prediction based on financial news semantic and structural similarity," *Expert Syst. Appl.*, vol. 118, pp. 411–424, Mar. 2019.

[21] B. Weng, M. A. Ahmed, and F. M. Megahed, "Stock market one-day ahead movement prediction using disparate data sources," *Expert Syst. Appl.*, vol. 79, pp. 153–163, 2017.

[22] S. K. Khatri and A. Srivastava, "Using sentimental analysis in prediction of stock market investment," in *Proc. 5th Int. Conf. Rel., Infocom Technol. Optim. (Trends Future Directions) (ICRITO)*, Sep. 2016, pp. 566–569.

[23] X. Zhang, S. Qu, J. Huang, B. Fang, and P. Yu, "Stock market prediction via multi-source multiple instance learning," *IEEE Access*, vol. 6, pp. 50720–50728, 2018.

[24] M. Shastri, S. Roy, and M. Mittal, "Stock price prediction using artificial neural model: An application of big data," *ICST Trans. Scalable Inf. Syst.*, vol. 6, no. 20, Jul. 2018, Art. no. 156085.

[25] J. Li, H. Bu, and J. Wu, "Sentiment-aware stock market prediction: A deep learning method," in *Proc. Int. Conf. Service Syst. Service Manage.*, Jun. 2017, pp. 1–6.

[26] M. Kraus and S. Feuerriegel, "Decision support from financial disclosures with deep neural networks and transfer learning," *Decis. Support Syst.*, vol. 104, pp. 38–48, Dec. 2017.

[27] M.-Y. Chen, C.-H. Liao, and R.-P. Hsieh, "Modeling public mood and emotion: Stock market trend prediction with anticipatory computing approach," *Comput. Hum. Behav.*, vol. 101, pp. 402–408, Dec. 2019.

[28] A. Onan, "An ensemble scheme based on language function analysis and feature engineering for text genre classification," *J. Inf. Sci.*, vol. 44, no. 1, pp. 28–47, 2018.

[29] O. Aytuğ, "Sentiment analysis on Twitter based on ensemble of psychological and linguistic feature sets," *Balkan J. Electr. Comput. Eng.*, vol. 6, no. 2, pp. 69–77, 2018.

[30] A. Onan, S. Korukoğlu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Syst. Appl.*, vol. 57, pp. 232–247, Sep. 2016.

[31] A. Onan, S. Korukoglu, and H. Bulut, "Lda-based topic modelling in text sentiment classification: An empirical analysis," *Int. J. Comput. Linguistics Appl.*, vol. 7, no. 1, pp. 101–119, 2016.

[32] A. Onan and M. A. Toçoğlu, "A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification," *IEEE Access*, vol. 9, pp. 7701–7722, 2021.

[33] A. Onan, "Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks," *Concurrency Comput., Pract. Exper.*, Jun. 2020, Art. no. e5909.

[34] A. Onan, "Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering," *IEEE Access*, vol. 7, pp. 145614–145633, 2019.

[35] A. Onan, "Topic-enriched word embeddings for sarcasm identification," in *Proc. Comput. Sci. Line Conf.*, Springer, 2019, pp. 293–304.

[36] F. Z. Xing, E. Cambria, L. Malandri, and C. Vercellis, "Discovering Bayesian market views for intelligent asset allocation," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Springer, 2018, pp. 120–135.

[37] S. Koyano and K. Ikeda, "Online portfolio selection based on the posts of winners and losers in stock microblogs," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2017, pp. 1–4.

[38] Q. Song, A. Liu, and S. Y. Yang, "Stock portfolio selection using learning-to-rank algorithms with news sentiment," *Neurocomputing*, vol. 264, pp. 20–28, Nov. 2017.

[39] Y. Ye, H. Pei, B. Wang, P.-Y. Chen, Y. Zhu, J. Xiao, and B. Li, "Reinforcement-learning based portfolio management with augmented asset movement prediction states," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 1, 2020, pp. 1112–1119.

[40] M. Chen, Z. Zhang, J. Shen, Z. Deng, J. He, and S. Huang, "A quantitative investment model based on random forest and sentiment analysis," *J. Phys., Conf. Ser.*, vol. 1575, Jun. 2020, Art. no. 012083.

[41] E. L. Park and S. Cho, "KoNLPy: Korean natural language processing in python," in *Proc. 26th Annu. Conf. Hum. Cognit. Lang. Technol.*, Chuncheon, South Korea, Oct. 2014, pp. 133–136.

[42] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to Japanese morphological analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2004, pp. 230–237.

[43] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Statistician*, vol. 46, no. 3, pp. 175–185, Aug. 1992.

[44] J. S. Cramer, "The origins of logistic regression," Tech. Rep., 2002.
[45] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
[46] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[47] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, pp. 1189–1232, Oct. 2001.

POONGJIN CHO received the B.Sc. degree in industrial and management engineering, and mathematics from POSTECH, in 2013, and the Ph.D. degree in industrial engineering from Seoul National University, in 2019. He is currently working as a Senior Data Scientist with FnGuide. His research interests include pattern recognition, time-series analysis, econophysics, operations research, and the applications of these areas in crisis management, risk assessment, fraud detection, AI trading, demand forecasting, trend prediction, fintech, and other topics related to data-driven analytics.

JI HWAN PARK received the B.Sc. and Ph.D. degrees in industrial engineering from Seoul National University, in 2014 and 2020, respectively. He is currently working as a Postdoctoral Researcher with Hanyang University. His research interests include complex network analysis in financial markets, link prediction, portfolio management, anomaly detection, machine/deep learning applications in financial time-series, and other topics related to business analytics.

JAE WOOK SONG received the B.Sc. degree in industrial engineering from the Georgia Institute of Technology, in 2010, and the Ph.D. degree in industrial engineering from Seoul National University, in 2016. Then, he had been working as a Senior Data Scientist with Samsung Electronics and an Assistant Professor of data science with Sejong University. In 2019, he joined the Department of Industrial Engineering, Hanyang University, where he is currently an Assistant Professor running the Financial Innovation & AnalytiX Laboratory. His research interests include inter- and multi-disciplinary approaches for data-driven innovations in financial markets, risk management, investment strategies, portfolio management, and business analytics.