LLMs Analyzing the Analysts: Do BERT and GPT Extract More Value from Financial Analyst Reports?

Seonmi Kim[1,*] Seyoung Kim[1,*] Yejin Kim[1,*] Junpyo Park[1] Seongjin Kim[1] Moolkyeol Kim[1]
Chang Hwan Sung[2] Joohwan Hong[1,†] Yongjae Lee[1,†]
{msraask3,abc21389,kimyejin99,jaypark,kimsj7597,ccomo456,joohwanhong,yongjaelee}@unist.ac.kr
ChangHwan.Sung@invesco.com
[1] Ulsan National Institute of Science & Technology
South Korea
[2] Invesco Hong Kong Limited
Hong Kong SAR

ABSTRACT

This paper examines the use of Large Language Models (LLMs), specifically BERT-based models and GPT-3.5, in the sentiment analysis of Korean financial analyst reports. Due to the specialized language in these reports, traditional natural language processing techniques often prove insufficient, making LLMs a better alternative. These models are capable of understanding the complexity and subtlety of the language, allowing for a more nuanced interpretation of the data. We focus our study on the extraction of sentiment scores from these reports, using them to construct and test investment strategies. Given that Korean analyst reports present unique linguistic challenges and a significant 'buy' recommendation bias, we employ LLMs fine-tuned for the Korean language and Korean financial texts. The aim of this study is to investigate and compare the effectiveness of LLMs in enhancing the sentiment analysis of financial reports, and subsequently utilize the sentiment scores to construct and test investment strategies, thereby evaluating these models' potential in extracting valuable insights from the reports. The code is available at https://github.com/msraask3.

ACM Reference Format:
Seonmi Kim[1,*] Seyoung Kim[1,*] Yejin Kim[1,*] Junpyo Park[1] Seongjin Kim[1] Moolkyeol Kim[1], Chang Hwan Sung[2] Joohwan Hong[1,†] Yongjae Lee[1,†]. 2023. LLMs Analyzing the Analysts: Do BERT and GPT Extract More Value from Financial Analyst Reports?. In *4th ACM International Conference on AI in Finance (ICAIF '23), November 27–29, 2023, Brooklyn, NY, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3604237.3627721

1 INTRODUCTION

Over the last few decades, the sheer volume of unstructured data available to financial market participants surged, presenting challenges related to the processing and potential oversight of critical information. Machine learning methods, extensively reviewed and applied in the financial sector as highlighted by [21], offer promising solutions to address these data challenges. The application of Natural Language Processing (NLP) has offered new avenues for data analysis in various domains, including finance. Building upon these NLP foundations, numerous studies have adopted advanced deep learning NLP techniques and Large Language Models (LLMs) to analyze the sentiment-rich data sourced from news and social media. These sentiment scores have subsequently been utilized in various downstream tasks such as price forecasting, risk prediction, and portfolio optimization [3, 4, 22, 28, 36, 37]. However, due to the often high noise-to-signal ratio associated with these sources, critical information related to a corporation's fundamentals can frequently become obfuscated [7, 29, 31]. Within this context, the value of analyst reports as a data source becomes apparent.

According to Jegadeesh et al. [16], analyst reports, underpinned by insights from experienced analysts, serve as valuable resources in financial decision-making, particularly in investments [14, 15, 17]. Compared to news articles or social media posts, these reports have significantly less noise and often contain more refined information about individual stocks. While news articles and social media posts can be useful for understanding overall market sentiment or investor attention, they often fall short in providing detailed, fundamental insights about specific companies. On the other hand, analyst reports are the result of thorough analysis conducted by industry experts, making them a reliable source of information. Therefore, for investors seeking well-reasoned, data-driven investment advice, these reports are often an indispensable resource.

Financial reports are filled with various types of information. On one hand, they offer quantifiable metrics such as target prices and analyst ratings. On the other hand, they delve into textual elements that encompass opinions and market commentary. Although both types of data are crucial, textual elements are increasingly acknowledged as indispensable for financial decision-making, with both academics and practitioners concurring that reliance on numerical data alone is insufficient [19, 23]. Nevertheless, most existing studies have predominantly focused on the analyst's opinion or financial measures within the reports, where they evaluated the value of analyst opinions by assessing the profitability of the portfolio constructed based on their level of recommendations. Therefore, our study aims to expand upon these previous works by examining

[*] These authors contributed equally.
[†] Co-corresponding authors.

ICAIF '23, November 27–29, 2023, Brooklyn, NY, USA

ACM ISBN 979-8-4007-0240-2/23/11…$15.00
https://doi.org/10.1145/3604237.3627721

the value of the explanatory content within analyst reports, doing so through sentiment analysis of their textual content.

The task of extracting underlying sentiments from the unstructured data in financial reports is challenging due to the high complexity of the language. This complexity often arises from the manipulation of text intended to obscure true sentiments [26, 30]. For instance, negating positive words to convey negative implications is a common practice, and words like "offset" are often deployed to mask negative information behind a façade of positivity [26]. Moreover, analysts' investment opinions are often skewed towards a 'buy' recommendation, reflecting various underlying interests [17, 24]. For instance, analysts may be influenced by their firm's business relationship with the company in question, the potential impact on their own professional reputation, or the broader market sentiment. The finance field is further characterized by a plethora of specialized terminology and expressions, making the application of standard NLP techniques to this specialized financial language a complex task that may compromise accuracy [2, 38].

The emergence of LLMs, such as BERT [9] and GPT [5], offers promising prospects. These LLMs, trained on extensive textual data, in multiple languages, have shown remarkable capabilities in understanding and inferring the semantics of words and sentences in a wide range of contexts. They extend beyond mere word or sentence-level sentiment analysis, enabling comprehensive understanding and analysis at broader levels such as paragraphs or entire documents. In recent years, there have been developments in specialized iterations of these LLMs that are designed to navigate financial texts. Models like FinBERT [2, 44], FinGPT [43], and BloombergGPT [42] have been developed to decipher the complex financial jargon and unique terminologies, thereby enhancing the applicability of LLMs in the finance domain. However, while these models are often trained with multilingual data, their proficiency in interpreting lower-resourced languages, can be limited [20, 40]. Korean presents unique challenges due to its rich morphological structure, with words frequently comprising multiple morphemes, each carrying distinct information [33]. As a result, accurately tokenizing Korean text proves to be a complex task, requiring specific models designed to manage the intricacies of the Korean language. To address this issue, some Korean specific models, like KoBERT [1], have been fine-tuned with additional Korean language corpora. There are even models specifically fine-tuned for Korean financial text, such as KR-FinBERT [2], which stand as some of the best-performing BERT-based models in this niche field, adept at handling the challenges posed by the intricacies of both the Korean language and financial terminology.

Building upon the pioneering work of Jegadeesh et al. [16], this paper strives to further investigate the inherent value of analyst reports. To this end, we draw on the sophisticated capabilities of LLMs, specifically BERT-based models fine-tuned for Korean language and or financial contexts and GPT-3.5 [3]. Utilizing these models, we perform inference on a substantial corpus of approximately 16,000 Korean analyst reports. From the inferred sentiment scores, we construct simple investment strategies for practical testing, thus gauging the potential of these models to infer actionable information from the reports.

[1]https://github.com/SKTBrain/KoBERT
[2]https://huggingface.co/snunlp/KR-FinBert-SC
[3]https://platform.openai.com/docs/models/gpt-3-5

2 RELATED WORK

2.1 Analysis on Analyst Reports

The seminal work of Jegadeesh et al. [16] underscored the importance of analyst reports, thus inspiring numerous subsequent studies which further affirmed these reports as reliable and insightful resources for discerning stock and market trends [6, 8, 17]. Much of the existing literature focuses on the analyst's opinions or financial measures contained within these reports [11, 16, 17, 25]. For instance, [16, 17] created quintile portfolios based on analyst scores, demonstrating the value of these recommendations. Notably, existing research largely neglects the textual information within these reports. This omission could be significant as, given the potential bias towards buy recommendations in analyst reports, critical information that the analyst intends to imply might be embedded in the textual information. Such information often provides nuanced explanations for recommendation decisions, reflecting market trends, macroeconomic events, or analyses of stock movements - a component that the current literature lacks.

As the volume of analyst reports continues to grow, a handful of studies have started to apply natural language processing (NLP) techniques. [15] advanced the discussion by implementing Naïve Bayes method to classify the reports into either positive, negative, or neutral opinion categories. [8] relied on traditional text processing methods, part-of-speech, and simulated an investment strategy based on the text characteristics of equity research reports.

A growing body of literature emphasizes the need for more advanced language models that can decipher ambiguity in text and perform sentiment analysis in financial contexts [32, 34, 41]. A large portion of these studies involve employing LLMs to perform sentiment analysis of diverse sources such as tweets and news articles, which provides evidence for the effectiveness of LLMs in analyzing textual sentiment [13, 28, 32]. A few studies have applied LLMs to the sentiment analysis of financial reports. For instance, the research conducted by [12] utilizes DistilBERT [35] and FinBERT to process data from corporate annual reports for stock price prediction. DistilBERT is used to create a dataset of forward-looking statements (FLS) from these reports. These sentences then undergo sentiment analysis with FinBERT, yielding sentiment scores that serve as input for the prediction of stock prices. Taking a different approach, [45] proposed a variant of FinBERT, MFF FinBERT, which is proven to be effective for long-term stock price movement prediction by incorporating textual features from analyst reports.

2.2 Large Language Models

With the advent of the Transformer [39], a plethora of remarkable large language models have emerged in the NLP field. These models leverage massive text data through parallel processing and attention mechanisms, enabling them to comprehend and infer the meanings of words and sentences within extensive and diverse contexts. The recent unsupervised pre-training of language models on large corpora, such as BERT (Bidirectional Encoder Representations from Transformers) [9], and GPT (Generative Pre-training Transformer)

Table 1: Number of Analyst Reports by Source

Sources	Number of Reports	Sources	Number of Reports	Sources	Number of Reports
Kiwoom Securities	4091	NICE D&B	454	Hana Financial Investment	79
SK Securities	2963	Hanwha Investment & Securities	340	Ebest Investment & Securities	50
Yuanta Securities	2505	DB Financial Investment	282	Eugene Investment & Securities	47
Meritz Securities	1664	Korea Corporate Data	217	IBK Investment & Securities	46
HI Investment & Securities	763	Hanyang Securities	202	Daishin Securities	45
DB Securities	656	Korea IR Association	97	Dongbu Securities	26
Korea Investment & Securities	585	NH Investment & Securities	80	Hi Investment & Securities	24

[5] has significantly enhanced performance across numerous NLP tasks, including sentiment analysis and language inference.

BERT employs the encoder part of the Transformer and focuses on understanding the context bidirectionally during training. Typically having between 100 million to 300 million parameters, BERT is pre-trained using two extensive datasets: Wikipedia and BooksCorpus. It is trained by masking a portion of the input data and predicting it, allowing for the development of various pre-trained language models depending on language and application. Recently, GPT, an enormous language model designed for text generation tasks using the decoder part of the Transformer, has garnered attention. Unlike BERT, GPT primarily utilizes a unidirectional structure, processing text from left to right, and is pre-trained using unsupervised learning on a more diverse and larger corpus. For instance, GPT-3.5 boasts roughly 175 billion parameters and exhibits exemplary performance compared to various language models. In the realm of finance, the utility of such a model has been explored extensively [10, 18, 27]. The impressive capabilities of models like GPT-3.5 have set the stage for the next leap in language model development: domain-specific models. These are tailored to specific fields of knowledge, drawing on vast amounts of relevant data for their training. For example, within the financial sector, FinGPT [43], based on the GPT-4 architecture, was developed by training the GPT model with an extensive set of financial data, including financial news, social media, filings, and academic datasets.

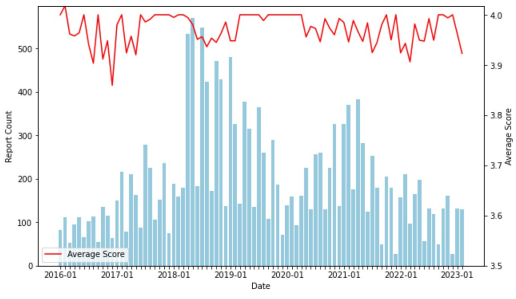
While Large Language Models (LLMs) such as BERT and GPT have been instrumental in advancing NLP, their primary training on general-domain corpora can limit their effectiveness in specialized domains. This is due to the unique language and vocabulary prevalent in such domains [2]. In the financial domain, the unique language and terminology present a significant challenge. To address the limitation, models such as FinBERT [2, 44], BloombergGPT [42], and FinGPT [43] have been introduced. These models have undergone fine-tuning, a process that occurs post pre-training, which involves retraining the pre-existing models on a more specific dataset. This fine-tuning on extensive financial corpora, which includes several years' worth of financial news and reports, is an effort to allow the models to adjust their parameters and improve their understanding of the unique nuances present in the specialized financial language. The results of this process are clear: these specialized models, particularly FinBERT and FinGPT, whose codes are publicly accessible, demonstrate improved performance on tasks related to financial texts when compared with general LLMs [1, 12, 45]. Unfortunately, the specifics and codes related to BloombergGPT are not publicly available, but it's worth noting that this model also demonstrates superior performance in finance-related tasks [42].

While finance-specialized models address the domain-specific challenge, language specificity poses another hurdle. In fact, several studies have indicated that the proficiency of widely used multilingual models in interpreting lower-resourced languages can be limited[20, 40]. In response to this language-specific challenge, several BERT-based models, such as KoBERT and KR-BERT-MEDIUM [4], have been fine-tuned specifically for Korean. KoBERT was trained on Korean Wikipedia data up to 2019, enabling it to handle a maximum input length of 512 tokens. KoBERT supports multiple tasks including text classification, named entity recognition, and question-answering. KR-FinBERT is a Korean-specific, finance-specialized model that builds upon the foundation of the KR-BERT-MEDIUM. The model was further trained on a diverse financial dataset, which included economic news articles from 72 media sources, along with analyst reports from 16 securities companies for sentiment analysis. The securities analyst reports and financial news data used for training spanned from January 2020 to April 2021.

3 DATA

This section outlines the dataset utilized and elaborates on the preprocessing procedures. We further discuss the conversion of analyst opinions into numerical form and share several revealing statistics gleaned from our dataset.

3.1 Data source and Description



score	count
5	4
4	16274
3	318
2	88
1	22

Figure 1: Monthly Report Activity and Analyst Scoring Patterns

The dataset utilized in this research comprises analyst reports issued by Korean securities firms, each focusing on a single stock, obtained from Paxnet[5]. The reports include tables illustrating financial measures, graphs, and texts. Originally in PDF format, these

[4] https://huggingface.co/snunlp/KR-Medium
[5] http://www.paxnet.co.kr/stock/report/report?menuCode=2222

reports were converted into text files using the PyPDF2 package. The content was filtered to primarily keep the texts that are in Korean, while allowing for certain English terms like company names. The text provides insights into market trends, macroeconomic events, and specific stock information, such as target prices and analysts' buy/hold/sell recommendations. In total, the final dataset comprises 16,706 reports from 21 leading securities companies in South Korea, spanning from January 4, 2016, to February 28, 2023. We have excluded reports missing analyst opinions. On average, about 194 reports are published each month, with a range from a minimum of 27 and a maximum of 570, as represented by the light blue bars in Figure 1. The average number of stocks analyzed in the reports per month is 122, ranging from 24 at the lowest to 273 at the highest.

3.2 Analyst Score

For our analysis, we also quantified the opinions provided in the analyst reports. Analyst opinions were manually translated into numerical values using the following scale: 1 for 'sell', 'sell (maintain)', 'underweight', and 'market underperform', 2 for 'neutral', 'market perform', 3 for 'hold', 4 for 'buy', 'market outperform', 'short-term buy', and 'buy (maintain)', and 5 for 'strong buy'. As shown in the table in Figure 1, most scores skewed towards 4, with very few 'sell' opinions represented by scores of 1. When multiple reports were issued for a single stock on the same day, which occurred in 1,427 instances, we computed the average analyst score. The stocks most frequently covered were CJ Corporation (001040.KS)[6], KT Corporation (030200.KS), Ncsoft Corporation (036570.KS), Samsung Electronics Co., Ltd. (005930.KS), and Hyundai Motor Company (005380.KS), all of which are also the largest companies in South Korea by market capitalization.

Over the entire period, the average monthly analyst score was 3.91. The red line in Figure 1 illustrates the trend of monthly average analyst scores. The month with the lowest average score was December 2016, with a score of 3.86, while the highest average score was observed in February 2016, at 4.02.

4 METHODOLOGY

We use three models and their specific details are provided in Section 4.1. Following this, in section 4.2, we categorize the portfolio construction process into three key perspectives: the choice of score, the manner of score utilization, and the approach to position taking. In Section 4.2.1, we elaborate on the methodology for processing analyst scores and sentiment scores. Section 4.2.2 focuses on the different approaches to score utilization, discussing the direct use of scores and their incremental changes. Finally, in Section 4.2.3, we explore the various approaches to position taking, namely long-only or long-short strategies.

4.1 Model

Our first task was to input the text of analyst reports into various language models and generate corresponding sentiment scores. We utilized models trained in Korean due to the linguistic composition of our dataset.

[6]The ticker notation used is in the format 'CompanyName (Korea ExchangeTicker)'. Korea ExchangeTicker refers to the company's ticker symbol on the Korea Exchange.

Table 2: LLM models used for sentiment analysis

Name	Base Model	Language	Dataset	Level
KoBERT	BERT	Fine-tuned for Korean	General-domain	Sentence
KR-FinBERT	BERT	Fine-tuned for Korean	Finance-specialized	Sentence
GPT-3.5	GPT-3.5	Multilingual	General-domain	Document

We employed three LLMs with distinct characteristics, summarized in Table 2. The first model, KoBERT, was trained on general domain data and fine-tuned for Korean language. The second model, KR-FinBERT, was fine-tuned for both Korean and finance-specific data. For both KoBERT and KR-FinBERT, we utlilized a sentence-level approach to extract the sentiment scores from the analyst reports. The third model, GPT-3.5, supports multilingual processing but lacks fine-tuning specifically for Korean. Unlike KoBERT and KR-FinBERT, GPT-3.5 has the advantage of analyzing sentiment scores for the entire analyst report, due to its ability to handle the longer input token length. Hence, we utilized a document-level approach to extract the sentiment score from GPT-3.5.

In this study, we aim to investigate which of these three LLMs is most proficient in deciphering the underlying sentiments embedded within analyst reports. Specifically, we explored the impact of three distinct features of the LLMs used in this study: language specialization (Korean), domain specialization (Korean + finance), and comprehension at a larger context level.

4.2 Portfolio Construction Methods

Our portfolio construction can be categorized into three main perspectives. Firstly, the choice of which score to use (analyst or sentiment). Secondly, the manner in which the score is utilized (score or incremental score). And thirdly, the approach to position taking (long-only or long-short).

4.2.1 Analyst and Sentiment Scores. From the analyst reports, we extracted two scores. One is the analyst score that is directly provided by the analyst, and the other is the sentiment score obtained through inference using pre-trained language models, without additional fine-tuning on our side.

All stocks appearing in the analyst report are denoted by i (where $i = 1, \ldots, N$), securities firms authoring the analyst reports as j (where $j = 1, \ldots, M$), and the time points as t (where $t = 1, \ldots, T$). Then, we can define a set $U_t = \{(i, j) |$ security firm j published analyst report on stock i in month t $\}$.

The analyst score of securities firm j on company i in month t is defined as $A(i, j)_t$, where $(i, j) \in U_t$. For $(i, j) \in U_t$, an analyst report written by securities firm j on stock i in month t has two components: analyst score $A(i, j)_t$ and content of the analyst report $R(i, j)_t$. When the instances where multiple reports were issued for a single stock in the same month, we calculated the mean of the score. So the final analyst score on company i in month t is expressed as follows:

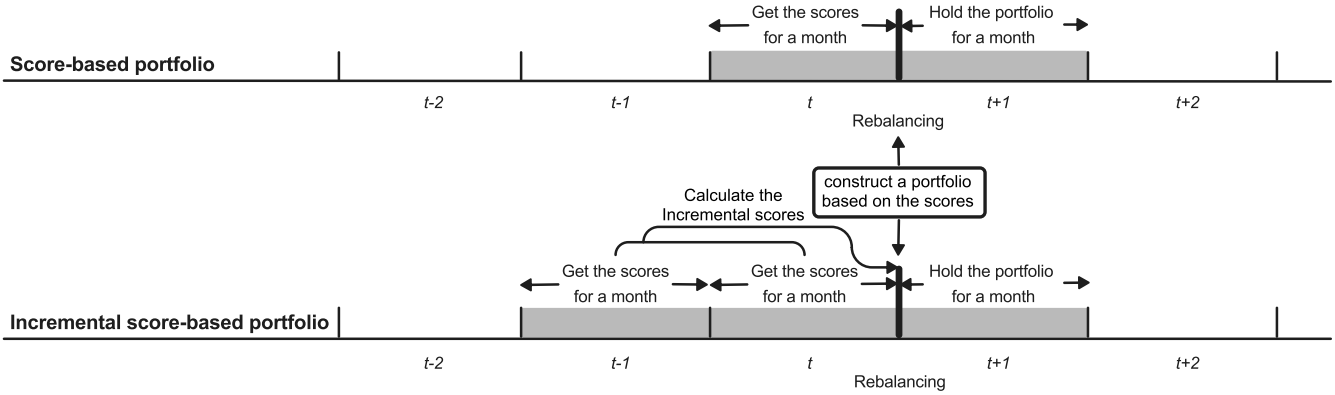


Figure 2: Experiment data accumulation periods relative to portfolio formation data

$$A(i)_t = \frac{\sum_{j \in J_{i,t}} A(i,j)_t}{|J_{i,t}|} \tag{1}$$

where $J_{i,t} = \{j|(i,j) \in U_t\}$ represents the set of securities firms that issued an analysis report for company i in month t, and $|J_{i,t}|$ is the number of such firms.

As we mentioned earlier, the scores can be extracted using two different approaches based on the characteristics of the models: sentence-level and document-level.

For sentence-level experiments, both positive and negative scores were extracted for each sentence in the analyst report $A(i,j)_t$, denoted as pos_s and neg_s respectively. This is achieved by applying an LLM to each individual sentence s in $R(i,j)_t$, represented as $pos_s, neg_s = \mathcal{F}(s|\Theta)$ where $\mathcal{F}$ is model and Θ is the trained model parameters. Here, KoBERT is a binary classification model, so neg_s = 1 - pos_s, and KR-FinBERT is a three-class model, so we can obtain the scores as $pos_s, neu_s, neg_s = \mathcal{F}(s|\Theta)$. However, for the purpose of portfolio construction in this study, we only consider pos_s and neg_s.

The sentiment score $S(i,j)_t$ for $R(i,j)_t$ is defined as the average score of all sentences within $R(i,j)_t$.

$$S(i,j)t = \frac{\sum_{s \in R(i,j)_t} pos_s}{|R(i,j)_t|} \tag{2}$$

where $|R(i,j)_t|$ is the number of sentences in $R(i,j)_t$.

For document-level experiments, the sentiment score can be obtained for each model as $S(i,j)_t = \mathcal{F}(R(i,j)_t|\Theta)$.

Finally, the sentiment score $S(i)_t$ of stock i in month t can be defined as follows:

$$S(i)_t = \frac{\sum_{j \in J_{i,t}} S(i,j)_t}{|J_{i,t}|} \tag{3}$$

4.2.2 Score or Incremental score. To assess the utility of analyst scores and the sentiment scores obtained through various LLMs, we constructed portfolio strategies based on these scores. The objective was to compare their performance and evaluate their effectiveness in portfolio construction. Generally, our portfolio designs were inspired by Jegadeesh et al. [16]. While they employed analyst scores and a quantitative summary score (QScore) in their research, our study focuses on utilizing both analyst scores and sentiment scores.

We split our portfolio construction process into two broad categories, *score-based portfolio*, and *incremental score-based portfolio* as depicted in Figure 2. Each category is distinguished by the specific method employed to utilize the scores for portfolio construction.

The first category is the *score-based portfolio*. This approach relies on the scores garnered during month t to construct the portfolio for the same month, t. The second category, known as the *incremental score-based portfolio*, takes into account the change in scores between the previous month, $t-1$, and the current month, t. This approach stipulates that a stock must possess recorded scores for both $t-1$ and t months to qualify for inclusion in the portfolio construction.

Having outlined the portfolio construction process, it's important to note that our rebalancing period is set on a monthly cycle, with operations taking place on the last day of each month at the closing price. In other words, buying and selling operations coincide with the closing price of the last day of each month.

In the case of sentiment scores, the score-based portfolio sorted the scores during the month t in ascending order. Similarly, the incremental score-based portfolio sorted them based on the magnitude of increase from the month $t-1$ to t, also in ascending order. After sorting the scores, we divided them into quintiles and used them for constructing the Long-only and Long-short portfolios, as explained later.

For the analyst score-based portfolios, we construct a portfolio with which stocks have received 'buy' recommendations in a previous month. Hence, the portfolio takes long positions on stocks with analyst scores of 4 ('buy') or 5 ('strong buy') in the month t. The incremental analyst score-based portfolios, which considers the change in analyst recommendation score, considered stocks with either the score in current month t remains same or increased compared to the previous month, $t-1$.

4.2.3 Long-only or Long-short. The position of the portfolio can be broadly categorized into two approaches: long-only and long-short. The distinction lies in whether to only take long positions on

top-score stocks into equal weight or to simultaneously take short positions on underperforming-score stocks into equal weight.

Specifically, for the sentiment score-based portfolio, we classified score and incremental score into quintiles. The scores were sorted in ascending order as described in Section 4.2.2, and we then identified the top 20% (the fifth quintile) and the bottom 20% (the first quintile) of the scores. In the long-only approach, we select the stocks from the fifth quintile and take long positions on them. In the long-short approach, in addition to the long positions on the fifth quintile, we also take short positions on the stocks from the first quintile.

The construction of an analyst score-based portfolio presents unique challenges due to the heavy skewness of score distribution towards 4, which renders the quintile division less meaningful or insightful. Consequently, our approach with both score and incremental score based portfolios for this case incorporates only long positions, and excludes short positions from this strategy.

5 EXPERIMENT

This section examines the empirical results of investment simulations based on the score-based and incremental score-based portfolios constructed using the analyst and sentiment scores mentioned in the previous section.

5.1 Experiment Settings

Stock data collection For the investment simulation of stocks with available analyst and sentiment scores, we collected the closing price data of 1,365 stocks listed in KOSPI and KOSDAQ (Korea Securities Dealers Automated Quotations) from FinanceDataReader[7]. The data collection period is from January 4, 2016, to February 28, 2023, the same as the start period of the analyst report data in Section 3.1.

Portfolio Construction As described in Section 4.2, we test the investment performance of portfolios constructed using four models: Analyst, KoBERT, KR-FinBERT, and GPT-3.5. In line with the approach of Jegadeesh et al. [16], we also conducted our experiments under the assumption that no transaction costs are incurred.

For each portfolio, we set a holding period of one month and calculated the daily returns over a span from February 2016 to February 2023. Note that there could be a potential data leakage issue with KR-FinBERT, considering it was trained on Korean financial news and reports spanning January 2020 to April 2021. Therefore, we carried out an additional experiment that excludes this period from our analysis (i.e., evaluation starts after April 2021). However, we found no significant difference, and thus, we report the results over the entire period.

Evaluation metrics To quantitatively compare the performance of each portfolio, we have set metrics, including cumulative returns, annualized returns, annualized volatility, annualized Sharpe ratio, and maximum drawdown (MDD). For simplicity, we set the risk-free rate to 0 when calculating the annualized Sharpe ratio.

For a more accurate comparison and understanding of each portfolio's performance against the market, we have selected the two versions of KOSPI (Korea Composite Stock Price Index), KOSPI (cap-weighted) and KS200EWI (equal-weighted), as our benchmarks. We have included the equal-weighted version, because we construct equal-weighted portfolios based on different groups of stocks chosen by different models.

[7] https://github.com/financedata-org/FinanceDataReader

5.2 Effectiveness of Score-Based Investment Portfolios

The effectiveness of the score-based portfolios, created according to the previously determined analyst and sentiment scores, were evaluated in real investment scenarios. The entire results from all analyst and sentiment score-based portfolios are illustrated in Table 3, with cumulative returns for Score-based long-only portfolios throughout the testing period shown in Figure 3.

These results reveal higher final cumulative returns for all long-only portfolios in the 'Score-based' section compared to the KOSPI 200 EWI, as shown in Figure 3. Looking more closely at Table 3, while there were minor improvements in MDD when compared with KOSPI 200 EWI and a slight uptick in volatility in some portfolios, it is noteworthy that all models, with the exception of KoBERT in the long-short portfolio, outperformed KOSPI 200 EWI in terms of both Sharpe Ratio and annualized return. These improvements suggest that analyst reports may indeed contain valuable investment insights.

As the results in Table 3 show, the performance levels vary significantly across different models, with GPT-3.5 and KR-FinBERT notably standing out by exhibiting superior performance in terms of both Sharpe Ratio and annualized return. These models excel over others such as KoBERT and Analyst, underscoring the advantages of a comprehensive review of analyst reports and the effectiveness of models fine-tuned specifically for the financial domain.

5.3 Enhanced Performance through Incremental Score-Based Portfolios

The following section discusses the utilization of incremental score-based portfolios and their potential to elevate investment performance as illustrated in the lower half of Table 3. Figure 4 offers a graphical representation of the cumulative returns from these portfolios throughout the test period.

Upon scrutinizing the results in Table 3, we notice a clear outperformance by the LLMs portfolios over the baseline in terms of annualized returns. Moreover, these portfolios, when compared to their score-based counterparts, showcase a significant improvement, underscoring the value of tracking score dynamics in investment strategies.

Among the various models, GPT-3.5 and KR-FinBERT emerge as the top performers in the context of long-only portfolios. GPT-3.5 exhibited the highest annualized return and Sharpe Ratio, further emphasizing the strengths of expansive language models proficient in comprehending texts in broader contexts. Not far behind, KR-FinBERT outperformed both Analyst and KoBERT models, highlighting the benefits of models fine-tuned for both specific language and specific domains.

A noteworthy detail emerges from this comparison: While GPT-3.5 showed strong performance in the long-only strategy, KR-FinBERT excelled in the long-short strategy. This could be a reflection of KR-FinBERT's capability to capture the indirect and subtly negative expressions often used by Korean analysts. This insight underscores

Table 3: Overall Performance of Analyst and LLM Portfolios

Portfolio type	Strategy	Model / Method	Annualized Return	Annualized Volatility	Sharpe Ratio	MDD
Benchmark	Buy-and-Hold	KOSPI	3.24%	17.00%	0.1908	-43.90%
	Buy-and-Hold	KOSPI 200 EWI	0.68%	18.30%	0.0372	-55.24%
Score-based	Long	GPT-3.5	**11.44%**	21.59%	**0.5298**	**-41.81%**
	Long	KR-FinBERT	**10.38%**	22.12%	**0.4690**	**-43.27%**
	Long	KoBERT	**8.40%**	21.00%	**0.3998**	**-47.62%**
	Long	Analyst	**7.31%**	19.92%	**0.3669**	**-44.86%**
	Long-short	GPT-3.5	**1.96%**	**13.00%**	**0.1511**	**-23.31%**
	Long-short	KR-FinBERT	**2.73%**	**14.26%**	**0.1917**	**-22.68%**
	Long-short	KoBERT	-0.37%	**12.10%**	-0.0302	**-26.28%**
Incremental score-based	Long	GPT-3.5	**37.87%**	23.82%	**1.5896**	**-30.08%**
	Long	KR-FinBERT	**33.03%**	22.83%	**1.4467**	**-30.09%**
	Long	KoBERT	**17.61%**	22.49%	**0.7832**	**-44.72%**
	Long	Analyst	**7.52%**	21.22%	**0.3544**	**-41.48%**
	Long-short	GPT-3.5	**36.76%**	23.42%	**1.5695**	**-37.60%**
	Long-short	KR-FinBERT	**60.13%**	23.59%	**2.5487**	**-15.01%**
	Long-short	KoBERT	**2.74%**	18.69%	**0.1466**	**-51.61%**

*Metrics that outperform the benchmarks are highlighted in bold, and the best performing instance within each portfolio type is underlined.
*Evaluation period: 02/01/2016 ~ 02/28/2023

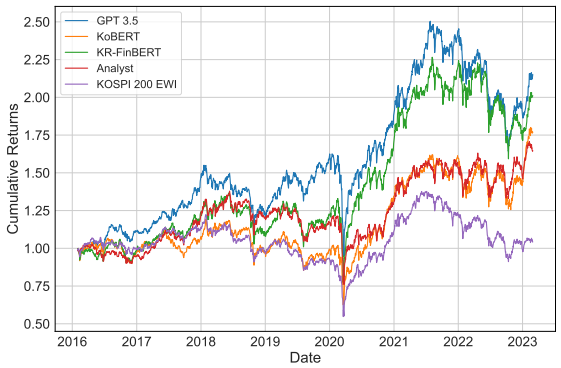


Figure 3: Cumulative Returns of Score-based Portfolios with Long-only Strategies

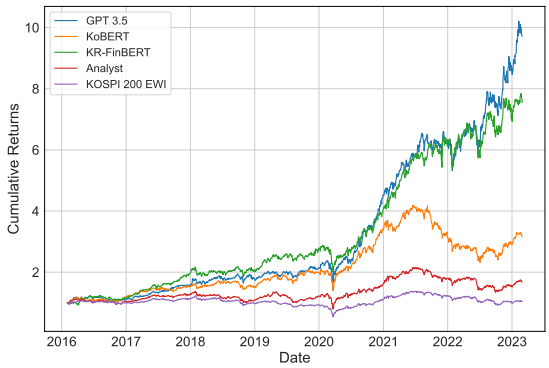


Figure 4: Cumulative Returns of Incremental Score-based Portfolios with Long-only Strategies

the utility of models fine-tuned for both language and financial domains in the context of incremental score-based portfolios.

In contrast, KoBERT displayed lower performance across all scenarios. However, when score dynamics were integrated into the investment strategy, even KoBERT managed to surpass the benchmark. This result serves to highlight the potential benefits of considering score dynamics in adaptive investment strategies.

6 CONCLUSION AND DISCUSSION

We have shown that LLMs can be useful in extracting value from Korean analyst reports. The portfolios constructed based on sentiment analysis of LLMs outperformed the benchmarks in the majority of the scenarios. This demonstrates LLMs' ability to decipher the subtle and often obscure sentiments present in analyst reports—nuances that analysts may feel reluctant or avoid revealing outright. Thus, our study suggests that the application of LLMs can play a pivotal role in effectively extracting and capitalizing on such implicit sentiment information in investment strategies.

Among various models tested, GPT-3.5 and KR-FinBERT stood out by significantly outperforming other models in all instances. It shows that the full potential of these reports is more effectively harnessed when analyzed through advanced LLMs. Their distinct performance not only highlights GPT-3.5's proficiency at comprehending context at the document level, but also underscores the potential of fine-tuning LLMs for specific domains, as evidenced by the results of KR-FinBERT.

In light of the promising results, it is important to address potential constraints that may affect the real-world applicability of our findings. Similar to the study by Jegadeesh et al. [16], our portfolio construction and overall experiments were conducted without considering transaction costs, which can influence outcomes in

practical investment scenarios. However, we note that the incremental score-based portfolios outperformed the benchmarks even with transaction costs of 0.5%. While this suggests the robustness of the methodologies proposed in this study, it also underscores the need for additional refinement for real-world applications. Given these findings, our study suggests a noteworthy potential for incorporating advanced NLP techniques such as LLMs into financial analysis and portfolio management.

ACKNOWLEDGMENTS

This work was supported by the 2023 Project Fund (1.230064.01) of UNIST (Ulsan National Institute of Science & Technology).
The views expressed herein are those of the authors and do not represent the views of Invesco Hong Kong Limited.

REFERENCES

[1] AI4Finance-Foundation. 2023. FinGPT Repository. https://github.com/AI4Finance-Foundation/FinGPT.
[2] Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063* (2019).
[3] Yalanati Ayyappa, B Vinay Kumar, Sudhabatthula Padma Priya, Siddiboena Akhila, Tamma Priya Vardhan Reddy, and Shaik Mohammad Goush. 2023. Forecasting Equity Prices using LSTM and BERT with Sentiment Analysis. In *2023 International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 643–648.
[4] Aysun Bozanta, Sabrina Angco, Mucahit Cevik, and Ayse Basar. 2021. Sentiment Analysis of StockTwits Using Transformer Models. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 1253–1258.
[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[6] Jeffrey A Busse, T Clifton Green, and Narasimhan Jegadeesh. 2012. Buy-side trades and sell-side recommendations: Interactions and information content. *Journal of Financial markets* 15, 2 (2012), 207–232.
[7] Hailiang Chen, Prabuddha De, Yu Jeffrey Hu, and Byoung-Hyoun Hwang. 2014. Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies* 27, 5 (2014), 1367–1403.
[8] Poongjin Cho, Ji Hwan Park, and Jae Wook Song. 2021. Equity research report-driven investment strategy in Korea using binary classification on stock price direction. *IEEE Access* 9 (2021), 46364–46373.
[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[10] Michael Dowling and Brian Lucey. 2023. ChatGPT for (finance) research: The Bananarama conjecture. *Finance Research Letters* 53 (2023), 103662.
[11] Jennifer Francis and Leonard Soffer. 1997. The relative informativeness of analysts' stock recommendations and earnings forecast revisions. *Journal of Accounting Research* 35, 2 (1997), 193–211.
[12] Alexander Glodd and Diana Hristova. 2023. Extraction of Forward-looking Financial Information for Stock Price Prediction from Annual Reports Using NLP Techniques. (2023).
[13] Anne Lundgaard Hansen and Sophia Kazinnik. 2023. Can ChatGPT Decipher Fedspeak? *Available at SSRN* (2023).
[14] Frank Hodge and Maarten Pronk. 2006. The impact of expertise and investment familiarity on investors' use of online financial report information. *Journal of Accounting, Auditing & Finance* 21, 3 (2006), 267–292.
[15] Allen H Huang, Amy Y Zang, and Rong Zheng. 2014. Evidence on the information content of text in analyst reports. *The Accounting Review* 89, 6 (2014), 2151–2180.
[16] Narasimhan Jegadeesh, Joonghyuk Kim, Susan D Krische, and Charles MC Lee. 2004. Analyzing the analysts: When do recommendations add value? *The journal of finance* 59, 3 (2004), 1083–1124.
[17] Narasimhan Jegadeesh and Woojin Kim. 2006. Value of analyst recommendations: International evidence. *Journal of Financial Markets* 9, 3 (2006), 274–309.
[18] Jang Ho Kim. 2023. What if ChatGPT were a quant asset manager. *Finance Research Letters* (2023), 104580.
[19] KPMG. 2014. *A New Vision of Value: Connecting Corporate and Societal Value Creation.* Technical Report. KPMG International. https://assets.kpmg.com/content/dam/kpmg/pdf/2014/06/kpmg-survey-business-reporting.pdf Accessed: 21-06-2023.
[20] Sangah Lee, Hansol Jang, Yunmee Baik, Suzi Park, and Hyopil Shin. 2020. Kr-bert: A small-scale korean-specific language model. *arXiv preprint arXiv:2008.03979* (2020).
[21] Yongjae Lee, John RJ Thompson, Jang Ho Kim, Woo Chang Kim, Francesco A Fabozzi, Frank J Fabozzi, Jim Musumeci, Bruce Feibel, Bradford Cornell, Zoltán Nagy, et al. 2023. An overview of machine learning for asset management. *The Journal of Portfolio Management* 49, 9 (2023), 31–63.
[22] Edmund Kwong Wei Leow, Binh P Nguyen, and Matthew Chin Heng Chua. 2021. Robo-advisor using genetic algorithm and BERT sentiments from tweets for hybrid portfolio optimisation. *Expert Systems with Applications* 179 (2021), 115060.
[23] Baruch Lev and Feng Gu. 2016. *The end of accounting and the path forward for investors and managers.* John Wiley & Sons.
[24] Mei hua Liao and Chia Yun Chang. 2014. Analysts' Forecasts and Institutional Investors' Behavior. In *Proceedings of the 2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. 575–579.
[25] Hsiou-wei Lin and Maureen F McNichols. 1998. Underwriting relationships, analysts' earnings forecasts and investment recommendations. *Journal of accounting and economics* 25, 1 (1998), 101–127.
[26] Yu-Wen Liu, Liang-Chih Liu, Chuan-Ju Wang, and Ming-Feng Tsai. 2018. Riskfinder: A sentence-level risk detector for financial reports. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. 81–85.
[27] Andrew W Lo, Manish Singh, Jim Musumeci, Zoltán Nagy, Guido Giese, Xinxin Wang, Andre Mirabelli, Nick Keywork, Avi Turetsky, Barry Griffiths, et al. 2023. From ELIZA to ChatGPT: The Evolution of Natural Language Processing and Financial Applications. *The Journal of Portfolio Management* (2023).
[28] Alejandro Lopez-Lira and Yuehua Tang. 2023. Can chatgpt forecast stock price movements. *Return Predictability and Large Language Models* (2023).
[29] Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance* 66, 1 (2011), 35–65.
[30] Tim Loughran and Bill McDonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research* 54, 4 (2016), 1187–1230.
[31] Thien Hai Nguyen, Kiyoaki Shirai, and Julien Velcin. 2015. Sentiment analysis on social media for stock movement prediction. In *International Conference on Advanced Data Mining and Applications*. Springer, 227–240.
[32] Derya Othan, Zeynep Hilal Kilimci, and Mitat Uysal. 2019. Financial sentiment analysis for predicting direction of stocks using bidirectional encoder representations from transformers (BERT) and deep learning models. In *Proc. Int. Conf. Innov. Intell. Technol.* 30–35.
[33] Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020. An empirical study of tokenization strategies for various Korean NLP tasks. *arXiv preprint arXiv:2010.02534* (2020).
[34] Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. 2003. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)*. 616–623.
[35] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).
[36] Emre Şaşmaz and F Boray Tek. 2021. Tweet sentiment analysis for cryptocurrencies. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*. IEEE, 613–618.
[37] Matheus Gomes Sousa, Kenzo Sakiyama, Lucas de Souza Rodrigues, Pedro Henrique Moraes, Eraldo Rezende Fernandes, and Edson Takashi Matsubara. 2019. BERT for stock market sentiment analysis. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 1597–1601.
[38] Ming-Feng Tsai and Chuan-Ju Wang. 2017. On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research* 257, 1 (2017), 243–250.
[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[40] Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076* (2019).
[41] Sida I Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 90–94.
[42] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).
[43] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-Source Financial Large Language Models. *arXiv preprint arXiv:2306.06031* (2023).
[44] Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097* (2020).

[45] Ming Zhang, Jiahao Yang, Meilin Wan, Xuejun Zhang, and Jun Zhou. 2022. Predicting long-term stock movements with fused textual features of Chinese research reports. *Expert Systems with Applications* 210 (2022), 118312. https://doi.org/10.1016/j.eswa.2022.118312

A LONG-SHORT PORTFOLIO PERFORMANCE

In the long-short scenario, the portfolio constructed using KR-FinBERT scores exhibited the best performance, followed by GPT 3.5 as shown in Figure 5 and 6. Conversely, in the long-only portfolios scenario, as discussed in section 5.2 and 5.3, the portfolio derived from scores generated by GPT 3.5 emerged as the top performer, with KR-FinBERT coming in second. This performance difference can be attributed to KR-FinBERT's enhanced ability to discern negative underlying sentiments, thereby generating profits from shorting those stocks. Nonetheless, KoBERT's performance remains weak in comparison to both KR-FinBERT and GPT 3.5.

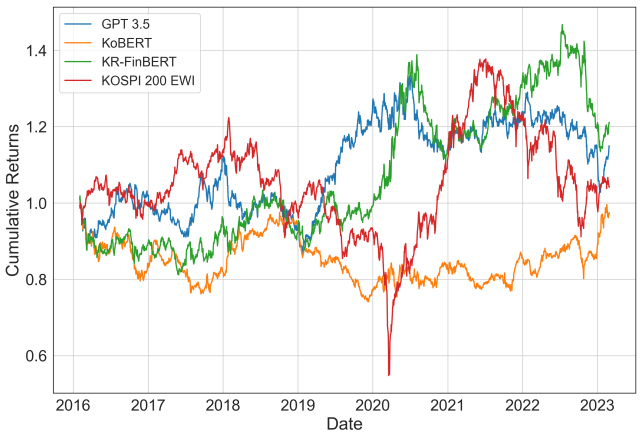


Figure 5: Cumulative Returns of Score-based Portfolios with Long-short Strategies

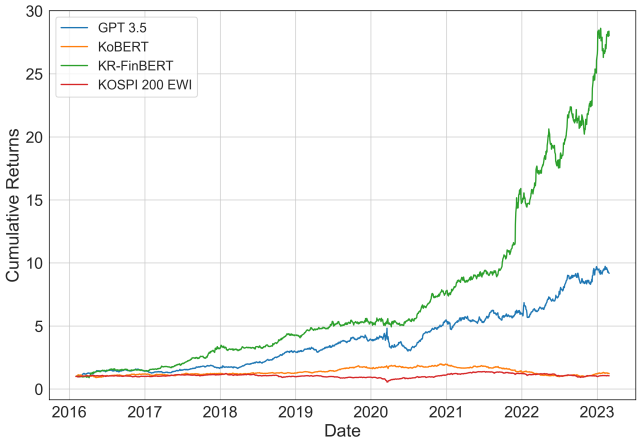


Figure 6: Cumulative Returns of Incremental Score-based Portfolios with Long-short Strategies

B ANALYSIS OF SENTIMENT SCORES BY LLMS

In further probing the resilience of sentiment score-based portfolios during the early stages of COVID-19, we examine the ability of models to identify negatively perceived stocks. We focused on industries significantly impacted during the initial phase of the pandemic from January to June 2020: Airlines, Hotels and Resorts, and Travel Agencies.

Figure 7 depicts the distribution of negative scores assigned by LLMs to stocks in the highlighted industries. KR-FinBERT consistently showed higher negative scores across all sectors, emphasizing its proficiency in understanding reports with negative undertones, particularly concerning the pandemic's adverse effect on travel and leisure sectors. In Airlines, KoBERT and GPT 3.5 clustered, with KoBERT slightly leading. For Hotels and Resorts, KoBERT had consistent scores, while GPT 3.5 had varied scores with occasional spikes indicating its ability to pinpoint underperforming stocks at certain times. Travel agencies had similar patterns with GPT 3.5 showing a higher average.

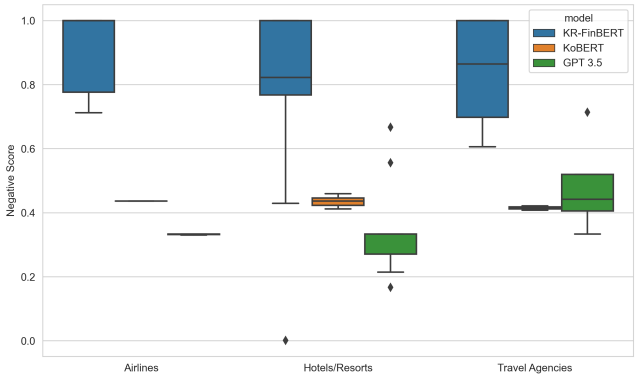


Figure 7: Negative Score Distribution by Industry and LLMs During the Early Stage of COVID-19 Pandemic

Industry	KR-FinBERT	KoBERT	GPT 3.5
Airlines	0.000779	0.000021	**0.003338**
Hotels/Resorts	**0.005128**	-0.002105	-0.000088
Travel Agencies	**0.009087**	0.003335	0.006501

Table 4: Long-Short Return by Industry and LLMs

Table 4 presents the average long-short returns for each LLM across various industry sectors during the aforementioned period. KR-FinBERT led in returns for the Hotels/Resorts and Travel agencies sectors, with GPT 3.5 and KoBERT following, mirroring the overall ranking of the long-short portfolio.