CIRJE-F-1248

Investment with New Sentiment Analysis in Japanese Stock Market: Expert Knowledge Can Still Outperform ChatGPT

Zhenwei Lin
The University of Tokyo

Masafumi Nakano
GCI Asset Management

Akihiko Takahashi
The University of Tokyo

March 2025

Investment with New Sentiment Analysis in Japanese Stock Market: Expert knowledge can still outperform ChatGPT

Zhenwei Lin,
Graduate School of Economics, University of Tokyo, 7-3-1 Hongo Bunkyo-ku, Tokyo, Japan, 113-0033

Masafumi Nakano [*],
GCI Asset Management, 9F Tokiwabashi Tower, 2-6-4 Otemachi, Chiyoda-ku, Tokyo, Japan, 100-0004

Akihiko Takahashi,
Graduate School of Economics, University of Tokyo, 7-3-1 Hongo Bunkyo-ku, Tokyo, Japan, 113-0033

First version: November 29, 2024, This version: March 27, 2025

Abstract

This paper presents a novel approach to sentiment analysis in the context of investments in the Japanese stock market. Specifically, we begin by creating an original set of keywords derived from news headlines sourced from a Japanese financial news platform. Subsequently, we develop new polarity scores for these keywords, based on market returns, to construct sentiment lexicons. These lexicons are then utilized to guide investment decisions regarding the stocks of companies included in either the TOPIX 500 or the Nikkei 225, which are Japan's representative stock indices. Furthermore, empirical studies validate the effectiveness of our proposed method, which significantly outperforms a ChatGPT-based sentiment analysis approach. This provides strong evidence for the advantage of integrating market data into textual sentiment evaluation to enhance financial investment strategies.

Keywords: sentiment analysis, text mining, large language models, natural language processing, ChatGPT, Japanese stock market, TOPIX 500, Nikkei 225, investment, alpha creation, risk-adjusted returns

[*]The findings and conclusions presented in this paper are based on the authors' analysis and interpretation of data. The authors do not guarantee the accuracy or completeness of the information provided. The content of this paper should not be considered as a recommendation or endorsement of any specific investment strategy or security. Investors should exercise their own judgment and seek professional advice before making any investment decisions.

1 Introduction

It is well recognized that stock price prediction is one of the central issues in financial investment. However, it is very challenging because financial markets are driven by the interaction between various factors such as economic, political, and psychological factors [1]. Therefore, it is important to collect and utilize appropriate data which affect the stock price movement as much as possible.

Initially, many researchers have focused on the market information as shown in [2–5] due to its accessibility. Specifically, market information data including return, volatility and dividend yield are mainly represented by numeric values, which are easier to handle than text data. However, numerical data alone often fail to capture psychological or contextual factors, leading to increased interest in text data. We note that text data analysis is inherently complex and has thus become a major research topic as natural language processing (NLP) (e.g., [6–11]).

Nonetheless, text data now attract more attention in financial investment due to the recent development of NLP especially boosted by artificial neural networks (ANNs). Before its appearance, the main stream of sentiment analysis is rule-based approach, where predefined linguistic rules by keywords, patterns, and linguistic heuristics are used to determine sentiment. See Tetlock [12] and Garcia [13] for analysis in the financial field.

Therefore, researchers have developed new approaches, such as Word2Vec [10], to address these limitations. Word2Vec, a shallow, two-layered neural network, generates vector representations of words known as word embeddings to capture contextual relationships and semantic meanings more effectively, thereby overcoming the rigidity of traditional rule-based methods. For applications in financial investment decisions, see Pagolu et al. [14] and Sun et al. [15], for instance.

Furthermore, ANNs have become increasingly important in text data analysis with the development of deep learning techniques [16], which enables fast and precise learning of multilayered ANNs. Since the deep-layered complex structure allows for the accurate approximation of non-linear functions as reported in Cybenko [17], a growing number of studies have applied ANNs to financial investment problems (e.g., [18–24]). In particular, the recent work by Lopez [24] exploits news headlines to predict stock price movements in the U.S. market using sentiment scores derived from ChatGPT, that is a large-scale language model (LLM) based on the Generative Pre-trained Transformer (GPT) architecture, to make a significant advancement in the field of NLP.

Although Lopez [24] demonstrates the excellence of ChatGPT in reading comprehension and question answering for predicting stock returns from text data, it remains uncertain whether these capabilities can be equally effective in languages fundamentally different from English. Additionally, the short test period of less than two years poses a limitation, as it hinders robust statistical analysis over extended timeframes.

In order to address these unresolved questions, Nakano and Yamaoka [31] thoroughly investigate the ChatGPT application to the sentiment analysis for stock return prediction, with a novel approach that extends beyond the ChatGPT framework. In particular, they explore the utilization of text data as a new α return resource for individual stock investment, and propose a new construction scheme of a polarity dictionary based on ChatGPT, which is also directly used for sentiment analysis. Specifically, their research focuses on analyzing non-English text data, expanding the application of sentiment analysis by ChatGPT to a broader range of markets and languages. Through extensive analysis of statistical test, they

identify the mean reversion feature following the release of negative news in Japanese large-cap companies.

However, we note that the biggest weak point of ChatGPT is its black-box nature of the response generation process, which motivates us to simultaneously explore interpretable methods such as a polarity approach to calculate sentiment scores. In particular, this paper proposes a new method using market return data to construct an original polarity dictionary, which is able to outperform ChatGPT approach for financial investment in the Japanese stock market. Departing from macro-level LLM approaches, our method emphasizes a micro-level perspective by developing an original polarity dictionary grounded in financial expert knowledge.

Specifically, the construction of a polarity dictionary is divided into two steps: creating a set consisting of target words extracted by morphological analysis and calculating the polarity values in the set. This paper presents novel schemes in both steps, which can be easily applied to stock markets in other countries with slight modifications.

We remark that there is a widely-used open-source Japanese morphological analysis tool called MeCab, which is often used to break down Japanese text into smaller components such as words, phrases, and grammatical elements. However, MeCab often fails to achieve the correct segmentation due to company-specific proper nouns and finance-related specialized terms.

Hence, we propose a simple and effective method that improves the MeCab segmentation. That is, the text is segmented using Japanese postpositional particles, symbols, and some other specific words as delimiters. In addition, we filter out noise words unlikely to influence investment outcomes (e.g., those with low frequency and whose average returns underperform transaction cost).

Also, let us remark that the polarities of words are generally based on the positive or negative labels manually attached to each sentence text by volunteers without financial expertise. Therefore, traditional methods do not take the expert knowledge and the real market reaction into account at all, which is true for large language models (LLMs) because market stock return data are not used directly in the training of LLMs.

Since our main purpose is to create an investment-focused polarity dictionary, the use of market return data seems more suitable to create a polarity dictionary which directly incorporates market features. Furthermore, we also test the case where market premium is used rather than raw return data. Here, market premium refers to the excess return of an individual stock over the market return, adjusted for systematic risk using its beta. This market premium isolates firm-specific factors from the whole market movements. These concepts provide a basis for our sentiment analysis approach, which integrates textual data with market-driven insights.

All of these procedures contrast sharply with LLMs such as ChatGPT, which rely on extensive, non-specialized text data interpreted by non-financial experts. In summary, the superiority of our proposed polarity dictionary lies in its ability to more accurately capture the domain-specific sentiment in financial news, which is directly learned from market return data.

As a result, the empirical studies demonstrate that our developed polarity dictionary scheme outperforms ChatGPT-based approach, especially for a period from April 2016 to Nov 2024, during which the benchmark index TOPIX records considerable positive returns. In particular, the close comparison analysis reveals the limitation of ChatGPT-based sentiment analysis which derives from its ignorance of market characteristics. In contrast, our proposed

method employing advanced segmentation and market excess return data successfully enables a more investment-focused sentiment analysis.

Lastly, we briefly review the other existing research about text analysis in Japanese stock market. Okimoto and Hirasawa [25] point out that text information has a strong impact on TOPIX return in the next day, where positive/negative polarity analysis is determined by news tag information defined in advance. Also, Goshima and Takahashi [26] use deep learning techniques to analyze the relationship between Reuters news and TOPIX. We remark that these works do not explore the connection between news and individual stocks. On the contrary, Katayama and Tsuda [27] and Akita [28] show that news information is useful to predict the individual stock price movement. Particularly for sentiment analysis, the former takes a polarity dictionary approach, while the latter adopts Paragraph Vector which extends the idea of Word2Vec to capture the semantic meaning of the entire paragraphs or documents.

Nishimura et al. [29] and Nakatani et al. [30] use text mining with morphological analysis based on MeCab, an open-source Japanese morphological analysis tool to develop and estimate three-factor term structure models with investor attitude factors, which are extracted by a text mining method of a large text base of daily financial news. Particularly, they estimate the entire yield curve and three factors using observed interest rates and frequencies of relevant word groups chosen from textual data. Moreover, they show that the estimated three factors, extracted only from the bond market data, are able to explain the movement in Nikkei 225 index.

The remainder of the paper is organized as follows: Section 2 summarizes text and market return data used in this paper. Section 3 explains our proposed methods for sentiment analysis. Analyzing the relationship between the news and stock returns, performances of the resulting investment strategies are shown in Section 4. Section 5 concludes. Appendix describes details of the performance measures appearing in this paper.

2 Financial Data

Our investment universe is composed of companies that are constituents of the TOPIX 500 and Nikkei 225 (NK225) indices, selected based on their liquidity in trading their stocks.

The TOPIX 500 is a capitalization-weighted stock price index, composed of 500 large-cap companies with high liquidity, all listed on the Tokyo Stock Exchange (TSE) and chosen by the Japan Exchange Group (JPX). In contrast, the Nikkei 225 (NK225) is a price-weighted index consisting of 225 large-cap companies selected by Nihon Keizai Shimbun, Inc.

It is worth noting that some companies are constituents of both indices. To define our universe, we take the union of the companies included in these two indices. We conduct an annual review and update of our investment universe since both TOPIX 500 and NK225 are periodically replaced. This approach ensures that our universe consists of the most liquid stocks in the Japanese market.

2.1 Text Data

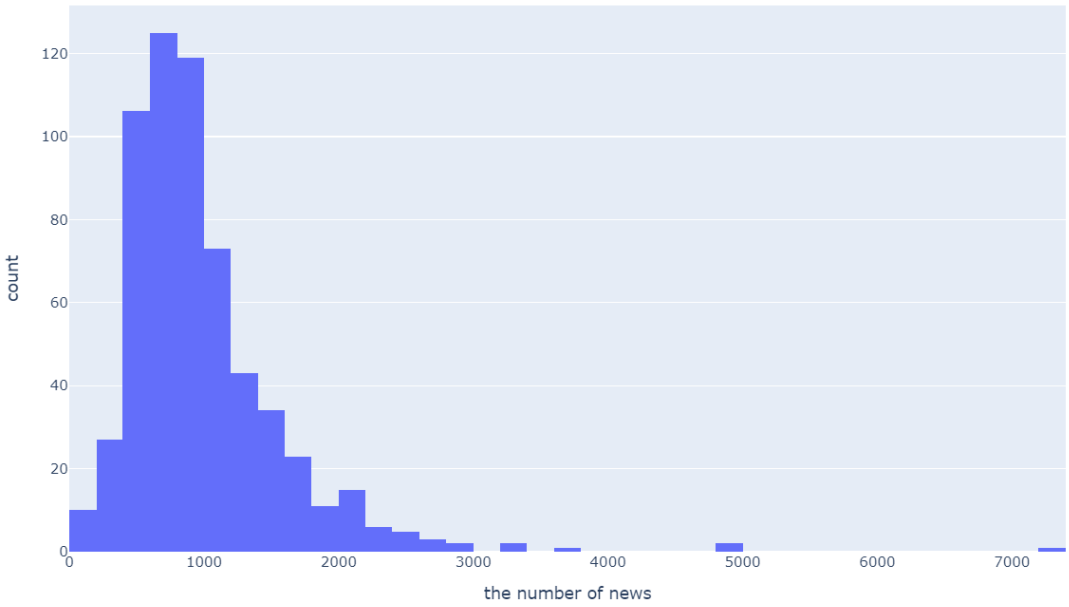
Now we acquire news headlines regarding the companies within our universe from October 1, 2013 to October 18, 2024. Particularly, for each company in our universe, we access the dedicated page of a financial news platform and retrieve all news articles listed under its news section. The collected news covers a broad spectrum of topics, including company-specific developments, industry-related events, and macroeconomic updates. It is important to note

that we exclude news directly associated with technical analysis, as the primary focus of this study is to explore the impact of news on stock price movements independent of technical analysis.

As a result, our dataset comprises 614,115 news headlines and we set $M = 614,115$. Also in the following section, let $E := \{i\}_{i=1}^{N_E}$ with $N_E = 631$ represents the set of companies. For each company i, M_i denotes the total number of associated news headlines. In other words, 631 companies were listed in either the TOPIX 500 or the NK225 over the 11-year period from 2013 to 2024.

While there are more than 900 news headlines for each company on average, the number of news articles varies significantly across the companies. The histogram in Figure 1, with the horizontal axis representing the number of news M_i, and the vertical axis indicating the number of companies, shows that the maximum number of news headlines reaches 7,249 while the minimum is 27. This figure suggests that certain companies receive significantly more attention than others, and that many companies cluster around the median of 831, that is, roughly 6 news headlines per month.

Fig. 1: Number of News and Company



Furthermore, we denote the news headlines and their corresponding timestamps for company i as $\{N_{i,j}\}_{j=1}^{M_i}$ and $\{ts_{i,j}^m\}_{j=1,m=1}^{M_i,m_i^j}$, respectively. Here, an $N_{i,j}$ may appear at multiple times with different timestamp $\{ts_{i,j}^m\}_{m=1}^{m_i^j}$. For example, a quite common news headline "Notice Concerning the Status of Repurchase of Shares of Common Stock" has been reported at 68 different timestamps in the past with regard to Toyota.

2.2 Market Return Data

For our numerical experiments, we obtain daily price data $\{P_{t,i}\}_{t,i}$ during the period from October 1, 2013 to October 18, 2024 from Bloomberg. The price data are adjusted to account for stock splits and dividend distributions. Here, t and i represent the time index and the company index, respectively. Using this data, we compute two types of returns. Namely, the close-to-close daily return $ON := \{r^{cc}_{t,i}\}_{t,i}$ and open-to-close intraday return $IN := \{r^{oc}_{t,i}\}_{t,i}$ are calculated as follows:

$$r^{cc}_{t,i} := P_{t^c,i}/P_{t-1^c,i} - 1, \qquad r^{oc}_{t,i} := P_{t^c,i}/P_{t^o,i} - 1, \tag{1}$$

where t^o and t^c indicate the market's opening and closing times on date t, respectively. The close-to-close return $r^{cc}_{t,i}$ reflects all market activity between the closing time of the previous day and current day t. The open-to-close return, on the other hand, isolates the market dynamics that occur exclusively during the trading hour, starting from the opening price and ending at the closing price.

Table 1 summarizes the descriptive statistics of those returns from October 1, 2013 to October 18, 2024, where the mean and standard deviation are annualized for comparison. Also, this table shows the close-to-close daily return of the TOPIX500 index r^{tp}_t as a benchmark.

Table 1: Descriptive statistics of the daily and intraday return

	intraday return $\{r^{oc}_{t,i}\}_{t,i}$	daily return $\{r^{cc}_{t,i}\}_{t,i}$	daily TOPIX500 return $\{r^{tp}_t\}_t$
Mean	-5.0%	12.4%	11.0%
Standard Deviation	26.1%	31.0%	18.4%
Skew	-0.32	0.68	-0.39
Kurtosis	20.08	46.76	10.25

Let us note that although the most well-known index in the Japanese stock market is TOPIX, we focused on TOPIX 500 as our benchmark. This is because TOPIX comprises over 1,500 constituent companies, which means that using TOPIX as a benchmark would entail substantial costs for collecting related news information. Notably, companies included in TOPIX 500 account for more than 90% of the total market capitalization of TOPIX. Furthermore, the correlation between TOPIX 500 and TOPIX exceeds 0.99, suggesting that employing TOPIX 500 as a proxy for the market index poses no significant issues.

3 Methodology: New Proposed Method

This section explains a novel method based on a Bag-of-Words (BoW) approach, which assesses sentiment by calculating the aggregate polarity of all the words contained in a news headline.

Particularly, we develop a new sentiment lexicon, which is a dictionary that assigns numerical polarity scores for sentiment to words or phrases in text data. These scores indicate whether a word or phrase has a positive or negative meaning, as well as its intensity. To create the sentiment lexicon, we compile a list of specific words or phrases (word keys) and assign each a corresponding sentiment score. Then, each news headline is classified as positive, neutral, or negative if its sentiment score is greater than, equal to, or less than 0, respectively.

Moreover, in the following empirical study section we will compare our proposed method with two existing methods for sentiment analysis: one with a MeCab-based Bag-of-Words (BoW) approach and the other with ChatGPT.

3.1 Selection of Keywords: Proposed Extraction Method

This subsection provides a detailed explanation of how to create a word set WL_{custom} consisting of keywords included in the news headlines. Since Japanese sentences are not separated by spaces, morphological analysis is particularly crucial at the outset. This study employs the following original method to extract words: Our custom extraction method is designed to effectively capture domain-specific expressions especially in finance, and ensures that contextually important words, which standard morphological analysis may overlook, are appropriately identified. In the following we outline this approach by introducing formal notations where appropriate.

First, we introduce two disjoint subsets of all news headlines denoted by $\mathcal{N}$:

$$\mathcal{F} \subset \mathcal{N} = \{N_{i,j}\}_{\forall i,j}, \quad \mathcal{N}_{\text{nonF}} \subset \mathcal{N}, \quad \text{such that} \quad \mathcal{F} \cap \mathcal{N}_{\text{nonF}} = \varnothing,$$

where the number of elements in set $\mathcal{N}$ is 541,702, and $\mathcal{F}$ represents a set of news headlines related to financial statements (e.g., corporate earnings), whose number of elements is 93,633.

We remark that these 541,702 news headlines comprise different contents each other, while the 614,115($= M$) news headlines mentioned in Section 2.1 include repeated instances of the same headlines appearing at different time points.

On the contrary, $\mathcal{N}_{\text{nonF}}$ consists of the news headlines classified as "zairyo" (in Japanese) which indicates a set of factors influencing stock prices except financial statements, and covers headlines regarding product launches, market updates, and so on. The number of elements in $\mathcal{N}_{\text{nonF}}$ is 166,790.

In addition, we denote the elements of $\mathcal{N}_{\text{nonF}}$ and the number of its elements as N_k^{NF} and K, respectively, that is, $\mathcal{N}_{\text{nonF}} = \{N_k^{NF}\}_{k=1}^{K}$.

Articles in $\mathcal{F}$ are distinguished by their use of industry-specific terminology and recurring linguistic patterns Typically, each company releases these articles about once per quarter, each providing a concise summary of its earnings report.

For each company $i \in E$, let

$$T_i = \{t_{i,k}\}_{k=1}^{L_i}$$

denote the set of earnings announcement timestamps obtained from Bloomberg, where L_i is the number of such announcements for company i.

For each company i, let us recall M_i denote the number of news headlines associated with company i, and let the available news headlines and their corresponding timestamps be given by

$$\{N_{i,j}\}_{j=1}^{M_i} \quad \text{and} \quad \{ts_{i,j}^m\}_{j=1,m=1}^{M_i,m_i^j},$$

respectively. Here, an $N_{i,j}$ may appear at multiple times, $\{ts_{i,j}^m\}_{m=1}^{m_i^j}$. We then extract, for each company i, the subset of headlines whose timestamps match any of the earnings announcement timestamps:

$$En_i = \{N_{i,j} \mid ts_{i,j}^m \in T_i\}.$$

Then, the overall set of pre-selected news articles related to financial statements is defined as

$$\mathcal{F} := \bigcup_{i \in E} En_i; \; E = \{i\}_{i=1}^{N_E}.$$

By this definition, $\mathcal{F}$ includes only those articles whose timestamps coincide with Bloomberg-reported earnings announcement times.

Within $\mathcal{F}$, we extract those headlines that follow the format "会社名、内訳"("company name, breakdown"). Denote the subset by:

$$\mathcal{F}_{\text{pattern}} = \{N \in \mathcal{F} \mid N \text{ is of the form "会社名、内訳"("company name, breakdown")}\},$$

where the number of elements in $\mathcal{F}_{\text{pattern}}$ is 25,758, around 27.5% in $\mathcal{F}$. Next, for each $N \in \mathcal{F}_{\text{pattern}}$ which can be decomposed into three parts, namely,
"会社名 (company name)", punctuation mark "、", and breakdown "s", we extract the breakdown s and define the set of these breakdowns as $F_{\text{excompany}} = \{s\}$. Then, for each $s \in F_{\text{excompany}}$, we extract candidate phrases as follows:

Given a breakdown s, we start by manually selecting a phrase with 4 characters, which seems to have an effect on a stock price, and then select different ones until no such phrases with 4 characters are found. We repeatedly apply the same procedure to selecting such phrases that consist of 3,2, and 1 characters in descending order for the breakdown s. We remark that phrases with four or fewer characters are sufficient to capture concise expressions frequently appearing in financial news, while those with five or more characters are likely to have multiple meanings and are therefore excluded.

As a result, we define a set of phrases which seem to have effects on the stock price within breakdown s as $P(s)$, and the set of all candidate phrases as

$$P = \bigcup_{s \in F_{\text{excompany}}} P(s).$$

Finally, we extract nouns, verbs, gerunds, and adjectives from the set P. Then, we denote the collection of those words as our keyword set W_f.

The summary of our procedure to obtain the key word set W_f is as follows:

1. In $\mathcal{F}$, extract N following the format "company name, breakdown".
   $\rightarrow \mathcal{F}_{\text{pattern}} = \{N \in \mathcal{F} | \text{"company name, breakdown"}\}$.
2. In $\mathcal{F}_{\text{pattern}}$, extract the breakdown part denoted by s.
   $\rightarrow F_{\text{excompany}} = \{s\}$.
3. In a breakdown $s \in F_{\text{excompany}}$, select phrases less than 5 characters having effects on a stock price.
   $\rightarrow P(s) \rightarrow P = \bigcup_{s \in F_{\text{excompany}}} P(s)$.
4. In P, extract nouns, verbs, gerunds and adjectives.
   $\rightarrow W_f = \{\text{nouns, verbs, gerunds, adjectives}\}$.

To help understanding our procedure, let us show some examples of the extraction:
$N \in \mathcal{F}_{\text{pattern}} \rightarrow (\text{nouns, verbs, gerunds, adjectives}) \in W_f$.

(Examples in Japanese)

1. トヨタ、今期最終は 5 %増益へ→（増益）
2. トヨタ、今期最終を 27 %上方修正→（上方修正）

3. トヨタ、上期最終は 26 %減益で着地、未定だった今期配当は 15 円増配→（減益、増配）

The English translations are in the following:
(Examples)

1. Toyota's net profit for the current fiscal year is increased by 5%.
   → (Profit Increase)
2. Toyota raises its net profit forecast for the current fiscal year by 27%.
   → (Upward Revision)
3. Toyota's net profit for the first half landed at a 26% decrease. The previously undecided dividend for the current fiscal year is increased by 15 yen.
   → (Profit Decline, Dividend Increase)

Consequently, we obtain set W_f consisting of 20 Japanese keywords in financial statements as follows:

Table 2: Key Japanese Words (their English translations) in financial statements

減益 (Profit Decline)	増益 (Profit Increase)	上方修正 (Upward Revision)
下方修正 (Downward Revision)	黒字浮上 (Return to Profit)	赤字転落 (Turn to Deficit)
赤字拡大 (Expanding Deficit)	赤字縮小 (Deficit Reduction)	連続 (Consecutive)
下振れ (Downward Deviation)	上乗せ (Additional)	下回る (Fall Below)
赤字 (Deficit)	黒字 (Profit)	超過 (Excess)
増額 (Increase)	増配 (Dividend Increase)	減配 (Dividend Cut)
最高益 (Record Profit)	上振れ (Upward Swing)	

Next, for news headlines that do not focus on financial statements, we use a multi-step approach. Specifically, we proceed as follows:

1. *Initial Segmentation:*
   (a) Each news headline $N_k^{NF} \in \mathcal{N}_{\text{nonF}}$ is split into smaller segments using postpositional particles (e.g., “は”, “が”, “を”), symbols, and some specific words as delimiters. That is, the news article $N_k^{NF} \in \mathcal{N}_{\text{nonF}}$ can be expressed by using the set D consisting of postpositional particles, symbols, and some specific words as follows [1]:
   $$N_k^{NF} = s_0^k \; d_1^k \; s_1^k \; d_2^k \cdots d_{n_k}^k \; s_{n_k}^k, \quad d_i^k \in D,$$
   where n_k represents the number of occurrences of the elements $d_i^k \in D$ in N_k^{NF}.
   (b) Then for each k, we define a list of segments $S_k = (s_1^k, \cdots, s_{n_k}^k)$, where it is possible that $s_l^k = s_m^k$ for $l \neq m$.
   (c) Next, we define a set S_k^{set} from the list S_k by eliminating duplication of the elements, i.e., $s_l^k \neq s_m^k$ for $l \neq m$, $s_l^k, s_m^k \in S_k^{set}$.
   (d) Further, we define set $S_k^{filtered}$ from set S_k^{set} by removing elements (segments) consisting of exactly two Hiragana characters (Japanese specific characters) and those containing numbers, alphabets, or company names.

[1] Concrete elements in the set D will be given upon request.

(e) Finally, we define the set of total segment $TS := \cup_{k=1}^{K} S_k^{filtered}$, which consists of unique elements.

(f) Hereafter, tss_ℓ and N_{TS} stand for the ℓ-th (distinct) element of TS and the number of its elements N_{TS}, respectively.
That is, $TS = \{tss_\ell\}_{\ell=1}^{N_{TS}}$ with $tss_m \neq tss_n$ for $m \neq n$.

2. *Frequency Analysis and Selection:*
For each element $tss_\ell \in TS$, let us assign the number FR_ℓ which means the frequency of each element tss_ℓ as follows:

$$FR_\ell = \sum_{k=1}^{K} \#\{s_m^k = tss_\ell : \forall s_m^k \in S_k\}, \tag{2}$$

where $\#S$ is its counting measure, i.e. $\#S := \sum_{w \in S} 1$.

Then, we define a set of words or phrases, each of which appears more than 200 times in non-financial news:

$$A := \{tss_\ell \in TS \mid FR_\ell > 200\}. \tag{3}$$

3. We then manually remove elements that do not seem to have any direct effects on stock prices to define a set B consisting of 71 elements.

Finally, the union set of W_f and B consists of our custom keyword list:

$$WL_{\text{custom}} = W_f \cup B = \{\text{word}_k\}_{k=1}^{W}, \tag{4}$$

where $W = 88$ denotes the total number of words in WL_{custom}. We note that there exist common elements in W_f and B. The following Table 3 shows all the elements in WL_{custom}.

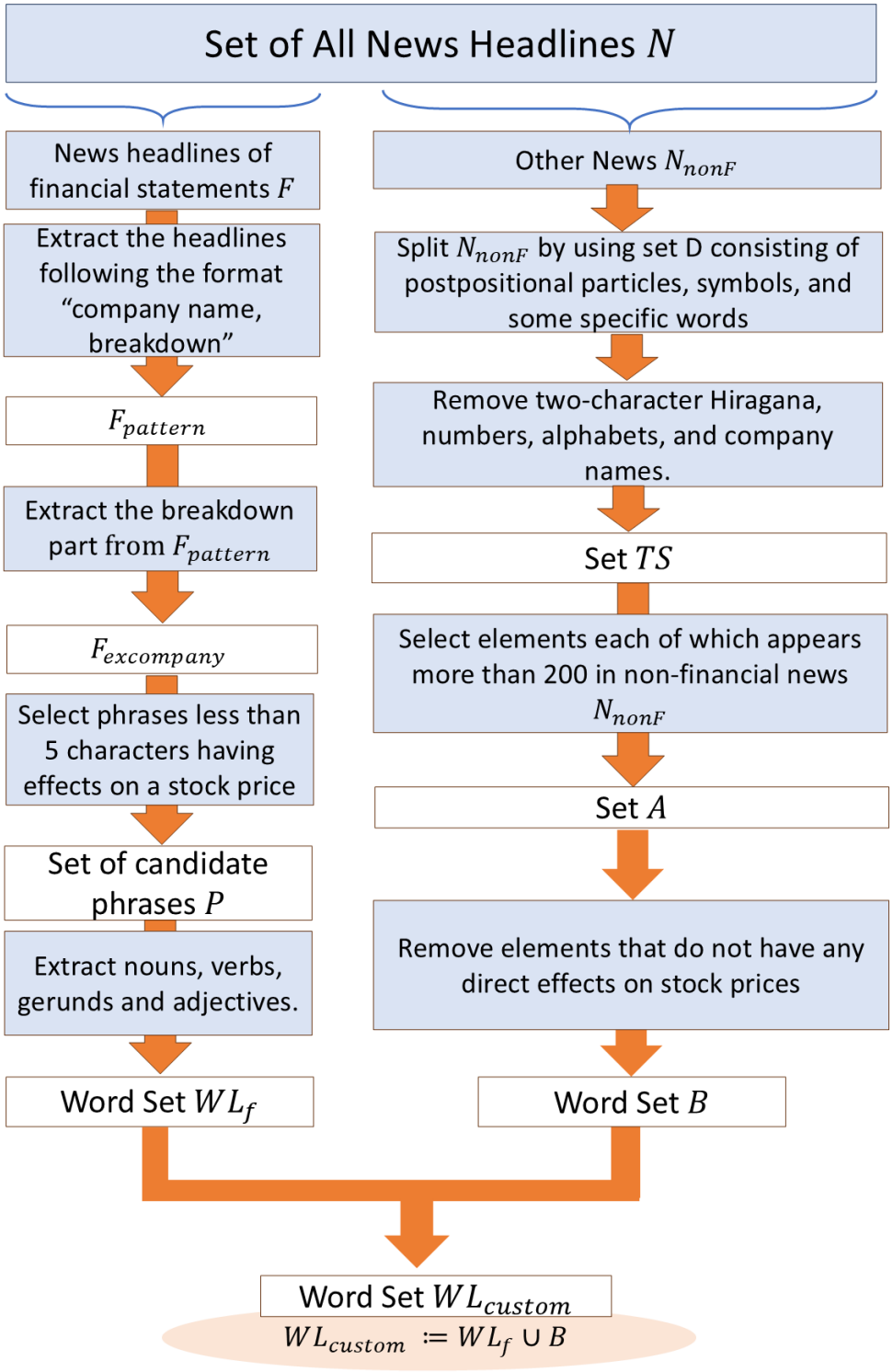
Table 3: Key Japanese Words (their English translations) in WL_{custom}

赤字拡大 (Expanding Deficit)	追い風 (Tailwind)	前年下回る (Below Last Year)
増益 (Profit Increase)	引き下げ (Lowering)	続急伸 (Continued Surge)
期待 (Expectation)	前年上回る (Above Last Year)	軟調 (Weak)
買い気配 (Strong Buying Interest)	赤字転落 (Turn to Deficit)	ネガティブ視 (Viewed Negatively)
軒並み高 (Broad Gains)	買い優勢 (Buying Dominance)	カバレッジ開始 (Coverage Initiation)
黒字浮上 (Return to Profit)	下振れ (Downward Deviation)	大幅続落 (Sharp Continued Decline)
減益 (Profit Decline)	増額 (Increase)	上方修正 (Upward Revision)
もみ合い (Range-bound)	売り優勢 (Selling Dominance)	格上げ (Upgrade)
高値更新 (New High)	カイ気配スタート (Strong Buy Start)	ストップ高 (Limit-up)
注目 (Attention)	上乗せ (Additional)	強気評価 (Bullish Rating)
急騰 (Rapid Rise)	強気 (Bullish)	減配 (Dividend Cut)
大幅高 (Significant Rise)	急反落 (Sharp Rebound)	買い推奨 (Buy Recommendation)
カイ気配 (Strong Buying Interest)	上振れ (Upward Swing)	懸念 (Concern)
下回る (Fall Below)	下方修正 (Downward Revision)	上昇 (Rise)
急伸 (Rapid Surge)	上場来高値 (All-time High)	大幅続伸 (Significant Continued Gain)
最高益 (Record Profit)	引き上げ (Increase)	警戒 (Caution)
黒字 (Profit)	上限 (Upper Limit)	売られる (Sold Off)
増配 (Dividend Increase)	反発 (Rebound)	大幅 (Significant)
続伸 (Continued Gain)	目標株価引き上げ (Target Price Increase)	大幅安 (Sharp Decline)
急落 (Plunge)	超過 (Excess)	嫌気 (Aversion)
新高値 (New High)	堅調 (Steady)	下落 (Decline)
続落 (Continued Decline)	連続 (Consecutive)	買い (Buying)
赤字縮小 (Deficit Reduction)	安い (Low)	急反発 (Rapid Rebound)
買い先行 (Buying Lead)	買収 (Acquisition)	高い (High)
好感 (Positive Reception)	格下げ (Downgrade)	大幅反発 (Strong Rebound)
想定以上 (Above Expectations)	計画上振れ (Plan Overshoot)	ストップ高買い気配 (Limit-up with Strong Buy)
買われる (Bought)	マイナス視 (Viewed Negatively)	材料視 (Viewed as a Factor)
自社株買い (Share Buyback)	年初来高値 (Year-to-date High)	赤字 (Deficit)
年初来高値更新 (New YTD High)	好調 (Strong Performance)	反落 (Pullback)
大幅反落 (Sharp Pullback)		

We note that methods such as MeCab and particle-based segmentation may result in inappropriate splits, making it difficult to accurately measure the frequency of meaningful words. On the contrary, our method can select keywords directly affecting stock prices buried in long segments, which frequently appear in the set of news headlines.

In summary, we show the flowchart for the construction of WL_{custom} below.

Fig. 2: Flowchart for the construction of WL_{custom}



3.2 Word Polarity Scores, Sentiment Lexicons, and Sentiment Scores

This subsection explains how to calculate the polarity scores of words and construct *sentiment lexicon* SL^ℓ ($\ell = 1, 2, 3, 4$), the set of a pair of a certain word and the corresponding polarity score, based on the word set WL_{custom}. In this subsection, we present how to obtain sentiment scores $S^\ell_{i,j}$ with type ℓ(=1,2,3,4) and news headline j of company i for WL_{custom}.

We remark that for simplicity, all the calculations in this subsection are shown by using the entire dataset during the period between October 1, 2013 and October 18, 2024. In the next section, we will adequately change the period of calculations for out-of-sample simulations to evaluate our proposed method in investment.

First, let us recall our notations: The set of all companies is represented as $E := \{i\}_{i=1}^{N_E}$, where i refers to company i in the set. For each company i, there is a certain number of associated news headlines, and this total number is denoted as M_i.

The actual news headlines for company i are labeled $\{N_{i,j}\}_{j=1}^{M_i}$, where each $N_{i,j}$ denotes news headline j of company i. Each headline also has its timestamps, denoted as $\{\text{ts}^m_{i,j}\}_{j=1,m=1}^{M_i,m^j_i}$, where each $\text{ts}^m_{i,j}$ consists of (month/date/year, time), and we use "time" and "month/date/year" extracted from the timestamp denoted by $\text{ts}^{\text{m,time}}_{i,j}$ and $t^m_{i,j}$, respectively. This means that for every news article $N_{i,j}$, there is a corresponding timestamp $ts^m_{i,j}$ to show when it is published. We remark that the same $N_{i,j}$ may appear at different timestamps $ts^m_{i,j}$, $m = 1, \cdots, m^j_i$.

There are two types of returns related to news headlines denoted hereafter by raw return (NR) and market premium (MP).

The raw return set NR can be defined as follows:

$$NR = \left\{ r_{t^m_{i,j}+1,i} \;\middle|\; r_{t^m_{i,j}+1,i} = \begin{cases} r^{oc}_{t^m_{i,j}+1,i}, & \text{if } ts^{m,\text{time}}_{i,j} \geq 14:57, \\ r^{cc}_{t^m_{i,j}+1,i}, & \text{if } ts^{m,\text{time}}_{i,j} < 14:57, \end{cases} \right. \\ \left. i = 1, \ldots, N_E, \quad j = 1, \ldots, M_i, \quad m = 1, \ldots, m^j_i \right\}. \tag{5}$$

with

$$r^{oc}_{t,i} = P_{t^c,i}/P_{t^o,i} - 1, \qquad r^{cc}_{t,i} = P_{t^c,i}/P_{t-1^c,i} - 1, \tag{6}$$

as defined in Eq.(1), where t^o and t^c indicate the market's opening and closing times on date t, respectively, $t + l$ ($t - l$) denotes the l-th business day after (before) t.

The return calculation differs in the timestamp of a news event. If news published at or after 14:57, the open-to-close intraday return of the following trading day is used. If it publishes before 14:57, the close-to-close daily return is applied.

To calculate the market premium, the 260-day beta is necessary in our study. We note that the beta measures the sensitivity of a company's stock return to the market return, and company i's beta at time-t denoted by $\beta_{t,i}$ is calculated in the subsequent analysis as follows:

$$\beta_{t,i} = \frac{\sum_{l=1}^{260}(r^{cc}_{t-l,i} - \bar{r}_{t,i})(r_{t-l,mkt} - \bar{r}_{t,mkt})}{\sum_{l=1}^{260}(r_{t-l,mkt} - \bar{r}_{t,mkt})^2}, \bar{r}_{t,i} = \frac{1}{260}\sum_{l=1}^{260} r^{cc}_{t-l,i}, \bar{r}_{t,mkt} = \frac{1}{260}\sum_{l=1}^{260} r_{t-l,mkt}, \tag{7}$$

where $r_{t,mkt}$ stands for the market index (TOPIX 500). (See Eq.(1), again.)

Using the beta $\beta_{t,i}$, the market premium MP constituting of $\tilde{r}_{t^m_{i,j}+1,i}$ can be expressed as follows:

$$MP = \left\{ \tilde{r}_{t^m_{i,j}+1,i} \;\middle|\; \tilde{r}_{t^m_{i,j}+1,i} = \begin{cases} r^{oc}_{t^m_{i,j}+1} - \beta_{t+1,i} * r^{oc}_{t+1,mkt}, & if \quad ts^{m,\text{time}}_{i,j} \geq 14:57 \\ r^{cc}_{t^m_{i,j}+1} - \beta_{t+1,i} * r^{cc}_{t+1,mkt}, & if \quad ts^{m,\text{time}}_{i,j} < 14:57, \end{cases} \right. \\ \left. i = 1, \ldots, N_E, \quad j = 1, \ldots, M_i, \quad m = 1, \ldots, m^j_i \right\}. \tag{8}$$

by subtracting the market index beta component from individual stock returns, we aim to isolate the return factor derived from news events more clearly.

Next, we explain how to calculate the polarity of a word in the set WL_{custom} by the following four methods.

1. Simple Average Method:
   The simple average method is a basic method for estimating word polarity based on the associated stock returns. If a word denoted by word_k appears in the news headline set $\mathcal{N} = \{N_{i,j}\}_{i,j}$, all stock returns $r_{t^m_{i,j},i}$ corresponding to the news headlines containing word_k are collected. Then, P^1_k, i.e, the polarity score for "word_k" is defined by the arithmetic mean of these returns as follows:

   $$P^1_k := \frac{1}{\#M^k} \sum_{(i,j,t^m_{i,j}) \in M^k} r_{t^m_{i,j}+1,i}, \tag{9}$$

   where M^k represents the set of triplets of firm index, news headline index, release date. Let us remind that each word_k belongs to WL_{custom} in Section 3.1.

   Since it would be nonsense to consider words that do not contribute positively to investment returns, we adopt only those words whose average returns exceed the transaction cost into our polarity dictionary. Specifically, we assume a round-trip transaction cost equivalent to two ticks as the minimum threshold for acceptable returns. Among the TOPIX 500 constituents, the security with the largest tick size had a tick size of 6.7 bps as of 2024. Thus, we set our threshold at 15 bps, slightly more than twice this value.

   Moreover, if the total number of the selected words exceeds 100, we select only the top 100 words with the highest absolute polarity scores such that $|P^1_k| > 0.0015$. Then, the set that is the *sentiment lexicon* consisting of the retained words is denoted as SL^1 with $\ell = 1$ corresponding to "Simple Average Method".

   In summary, using P^1_{101} which is the 101-th largest absolute value in $\{P^1_k\}_k$, *sentiment lexicon* SL^1, which is a pair of word and polarity score, is defined by

   $$SL^1 := \{(\text{word}^1_k, P^1_k); \text{word}^1_k \in WL_{custom} \wedge abs(P^1_k) > \max(0.0015, P^1_{101})\} \tag{10}$$

2. Simple Average of Market Premium:
   To exclude the effects of total market directions, raw returns $r_{t^m_{i,j}+1,i}$ are replaced by the news market premium $\tilde{r}_{t^m_{i,j}+1,i}$ in P^1_k as follows:

   $$P^2_k := \frac{1}{\#M^k} \sum_{(i,j,t^m_{i,j}) \in M^k} \tilde{r}_{t^m_{i,j}+1,i}. \tag{11}$$

In analogy with SL^1, *sentiment lexicon* SL^2 corresponding to "Simple Average of Market Premium" is defined as follows:

Namely, given a set of the retained words denoted as $\{\text{word}_k^2\}_k$, *sentiment lexicon* SL^2 is defined by

$$SL^2 := \{(\text{word}_k^2, P_k^2); \text{word}_k^2 \in WL_{custom} \wedge abs(P_k^2) > \max(0.0015, P_{101}^2)\} \tag{12}$$

This method extracts the unique effects of specific words on stock returns after removing the influence of the entire market movement by using market premiums.

3. Multiple Regression with Simple Returns:
In this method, a regression model is defined to estimate the polarity of each word based on its contribution to stock returns. The simple linear regression equation is given as follows:

$$r_{t_{i,j}^m,i} = \sum_{\text{word}_k \in SL^1} a_k 1_{\text{word}_k \in N_{i,j}} + \epsilon_{t_{i,j}^m,i}, \quad \forall r_{t_{i,j}^m,i} \in NR, \tag{13}$$

where $\epsilon_{t_{i,j}^m,i}$ is a noise term.

Then, we define the estimated coefficient a_k for each word_k as its polarity score, denoted as $P_k^3 := a_k$. This method provides polarity scores considering simultaneous effects of words $\text{word}_k \in SL^1$ on $r_{t_{i,j}^m,i}$ within the same news headline.

The resulting set, i.e., the *sentiment lexicon* is denoted as SL^3 with $\ell = 3$ corresponding to "Multiple Regression with Simple Returns".

Namely, given the same set of the words as in SL^1, $\{\text{word}_k^1\}_k$, *sentiment lexicon* SL^3 is defined by

$$SL^3 := \{(\text{word}_k^3, P_k^3);\ \text{word}_k^3 \in SL^1\} \tag{14}$$

4. Multiple Regression with Market Premium:
To further refine the regression analysis, the dependent variable is replaced with the news market premium, $\tilde{r}_{t_{i,j}^m,i}$ excluding market-wide influences. The regression model is defined as follows: For each $\tilde{r}_{t_{i,j}^m,i} \in MP$,

$$\tilde{r}_{t_{i,j}^m,i} = \sum_{\text{word}_k \in SL^2} \tilde{a}_k 1_{\text{word}_k \in N_{i,j}} + \tilde{\epsilon}_{t_{i,j}^m,i}, \tag{15}$$

where the variables are same as those previously defined except that $r_{t_{i,j}^m,i}$ is replaced by $\tilde{r}_{t_{i,j}^m,i}$, and that only the words word_k included in SL^2 are used in the regression. The coefficient $\tilde{a}_k$ for each word word_k represents its polarity score denoted as $P_k^4 := \tilde{a}_k$, which particularly excludes systematic market effects.

We expect that this method offers a more robust measure of the words' intrinsic impact on firm-specific performance, as it filters out the influence of broader market movements.

The resulting set, i.e., the *sentiment lexicon* is denoted as SL^4 with $\ell = 4$ corresponding to "Multiple Regression with Market Premium".

Namely, given the same set of the words as in SL^2, $\{\text{word}_k^2\}_k$, *sentiment lexicon* SL^4 is defined by

$$SL^4 := \{(\text{word}_k^4, P_k^4);\ \text{word}_k^4 \in SL^2\} \tag{16}$$

Finally, we calculate the sentiment score of type ℓ for the j-th news headline of company i denoted by $S_{i,j}^{\ell}$ as follows:

$$S_{i,j}^{\ell} = \sum_{\text{word}_k \in SL^{\ell}} 1_{\text{word}_k \in N_{i,j}} P_k^{\ell}, \tag{17}$$

where $1_{\text{word}_k \in N_{i,j}}$ is a binary indicator that equals 1 if the word word_k appears in the news headline $N_{i,j}$, and 0 otherwise; P_k^{ℓ} denotes the polarity score of word_k in the lexicon SL^{ℓ}, $\ell = 1, \cdots, 4$.

4 Empirical Study

4.1 Other Methods for Comparison

This subsection briefly explains two existing methods for comparison with our proposed method in the empirical analysis.

4.1.1 Application of ChatGPT

Recently, it is well-known that ChatGPT is a large-scale language model developed by OpenAI based on the GPT architecture. We ask to ChatGPT directly whether a news headline has a positive or negative effect on stock prices. Specifically, we apply ChatGPT[2] as a non-linear function $cgpt : N_{i,j} \to \{-1, 0, 1\}$ $(i = 1, \cdots, N_E;\ j = 1, \cdots, M_i)$ with the following prompt:

> Forget all your previous instructions. Pretend you are a financial expert with stock recommendation experience. Is this headline good or bad for the stock price of Company i?
> Headline: "$N_{i,j}$"
> Answer 1 if it is good news, -1 if it is bad news, or 0 if it is uncertain. Provide only the number as your response.

As a result, we obtain the sentiment score $S_{i,j}^{0}$ with $\ell = 0$ corresponding to "ChatGPT" by $S_{i,j}^{0} := cgpt(N_{i,j})$, $i = 1, \cdots, N_E;\ j = 1, \cdots, M_i$.

In fact, ChatGPT processes the raw text data $N_{i,j}$, meaning that it takes into account the order and context of sentences, which is different from the Bag-of-Words (BoW) approach that is incapable of analyzing the sentiment of text data lacking explicit sentiment words. Thus, the sentiment obtained by ChatGPT are expected to differ from those by the Bag-of-Words approach such as our proposed methods.

4.1.2 Selection of Keywords by MeCab

MeCab is a widely used tool for Japanese morphological analysis, which segments text into morphemes (the smallest meaningful units) and assigns grammatical information such as parts of speech.

Particularly, in this analysis we apply MeCab to all news headlines in $\mathcal{N} = \{N_{i,j}\}_{\forall i,j}$ by dividing those into small word units and then removing unnecessary elements such as numbers and symbols (e.g., "$\cdot$" and "$*$"). In addition, we obtain the set for only the words

[2]We use gpt-3.5-turbo-0125 and gpt-4o-mini-2024-07-18.

that appear more than 1,000 times in the headlines to create our refined word set, named WL_{MeCab} consisting of 1,887 keywords. Then, we apply the same method as in Section 3.2 with replacing WL_{custom} by WL_{MeCab} to construct the sentiment lexicon and calculate sentiment scores based on WL_{MeCab}.

While MeCab is effective for general purposes, it occasionally fails to recognize compound nouns or domain-specific terms, leading to incorrect splitting and decreasing segmentation accuracy. This issue is particularly important in handling financial news, as it often includes numerous specialized terms. To overcome the problems, we have developed an original method in the previous subsection to enhance the quality of the word list.

4.2 Backtesting Framework

This subsection outlines the backtesting framework utilized to evaluate sentiment-driven trading strategies in the Japanese stock market. Specifically, the polarity scores and sentiment lexicons are adaptively updated on an annual basis, using historical return data available until the end of the preceding year. This approach acknowledges yearly variations in company listings, returns, and associated news headlines. Sentiment scores are calculated daily at 14:57 JST, exclusively considering news headlines released up to that time from 0:00 JST. Subsequently, investment portfolios are adjusted based on these sentiment scores: a stock is purchased (long position) if its sentiment score is positive and closed out when the sentiment score becomes non-positive. Portfolios are equally weighted, and transaction costs are accounted for in the evaluation of investment returns

4.2.1 Yearly update of the polarity score

In conducting out-of-sample simulations, we adaptively update the polarity scores $\{P_k^\ell\}_k$ of each word and sentiment lexicons SL^ℓ ($\ell = 1, 2, 3, 4$) based on historical return data at the end of the year. This adaptive updating is necessary because, when initiating simulation analysis from a specific date (e.g., 2016/01/01), the available data including stock returns of companies and their news headlines are restricted to information up to the end of the previous year (in this example, 2015/12/31), dating back to the original data acquisition date (2013/10/01). Consequently, there are year-to-year variations in the sets of existing companies $\{i\}_i$, their stock returns $\{r_i\}_i$, and associated news headlines $\{N_{i,j}\}_{i,j}$. Therefore, we denote the polarity scores and sentiment lexicons employed in year y by ($\{P_{y-1,k}^\ell\}_k$) and SL_{y-1}^ℓ, respectively. Formally, the sentiment lexicons are represented as follows:

$$\begin{aligned} &SL_{y-1}^1 = \{(\text{word}_{y-1,k}^1; P_{y-1,k}^1)\}_k;\ SL_{y-1}^2 = \{(\text{word}_{y-1,k}^2, P_{y-1,k}^2)\}_k, \\ &SL_{y-1}^3 = \{(\text{word}_{y-1,k}^1, P_{y-1,k}^3)\}_k;\ SL_{y-1}^4 = \{(\text{word}_{y-1,k}^2, P_{y-1,k}^4)\}_k. \end{aligned} \tag{18}$$

4.2.2 Daily Calculation of the sentiment score

Since the Tokyo Stock Exchange closes at 15:00 JST, we focus exclusively on news headlines published up to 14:57 JST for our daily portfolio rebalancing. In other words, we take no action regarding any news released after 14:57 JST on a given trading day. Concretely, at 14:57 on each trading day d in year y, we collect all news headlines related to companies that are part of the TOPIX 500 or Nikkei 225 as of that day. We only include headlines published between 0:00 and 14:57 on day d.

We remark that news announced after 14:57 until 15:00 are not considered in our strategy, because we do not have enough time to incorporate information during the last 3 minutes into our positions. In addition, our separate research reveals that news released from 15:00 to 0:00 do not have meaningful impacts on the stock prices at the closing time of the following day, which is consistent with our intuition.

Next, the set of news headlines of company i appearing on a fixed day d in year y is represented as $\{N_{i,j}^{m,d,y}\}_{j,m}$ $(j = 1, 2, \ldots, J_i^{d,y})$, $(m = 1, 2, \ldots, M_{i,j}^{d,y})$, meaning that the company i's unique headline labeled as j shows up at different time points $m = 1, 2, \ldots, M_{i,j}^{d,y}$ on day d in year y. Here, for a given day d in year y, $J_i^{d,y}$ denote the total number of news headlines of company i, and $M_{i,j}^{d,y}$ stands for the total number of appearances of company i's headline j.

Using the sentiment lexicon SL_{y-1}^{ℓ} constructed at the end of the previous year, we compute the type ℓ sentiment score $\{S_{i,j}^{\ell,m,d,y}\}_{i,j,m}$ for each news headlines $N_{i,j}^{m,d,y}$ as follows:

For $\ell = 1, 2, 3, 4$,

$$S_{i,j}^{\ell,m,d,y} = \sum_{\text{word}_k \in SL_{y-1}^{\ell}} 1_{\text{word}_k \in N_{i,j}^{m,d,y}} P_{y-1,k}^{\ell}. \tag{19}$$

For $\ell = 0$, i.e., when using ChatGPT, we follow Section 4.1.1 to set

$$S_{i,j}^{0,d,y} = cgpt(N_{i,j}^{m,d,y}) \text{ with } cgpt : N_{i,j}^{m,d,y} \to \{-1, 0, 1\}. \tag{20}$$

For each company i, all the sentiment scores $\{S_{i,j}^{\ell,m,d,y}\}_{j,m}$ corresponding to the news headlines $\{N_{i,j}^{m,d,y}\}_{j,m}$ are aggregated to derive the daily sentiment score of each company.

Thus, the overall sentiment score for company i on day d in year y is defined as

$$S_i^{\ell,d,y} := \sum_{j=1}^{J_i^{d,y}} \sum_{m=1}^{M_{i,j}^{d,y}} S_{i,j}^{\ell,m,d,y}. \tag{21}$$

In this equation, the summation is taken across all distinct headlines $j = 1, 2, \ldots, J_i^{d,y}$ and, for each j, across all its occurrences $m = 1, 2, \ldots, M_{i,j}^{d,y}$.

4.2.3 Portfolio Construction

If the sentiment score for company i is positive, namely $S_i^{\ell,d,y} > 0$, we decide to take a long position of company i at the market close in our portfolio. Also, if we take a long position of company j on day $d-1$ and $S_j^{\ell,d,y} \leq 0$, we close out the long position of stock j at the market close. Furthermore, the portfolio on the date d is constructed with equal-weight allocation across all selected stocks (i.e., all stocks i such that $S_i^{\ell,d,y} > 0$).

Let us remark that we consider 2.5 bps trading cost into each trading for making/closing a position because the average 1 tick size within the TOPIX 500 constituents is 2.49 bps. Any performance metrics or returns are reported net of these transaction costs.

Hereafter for readability, we use abbreviated notations, SL_{MeCab}^{ℓ} and $SL_{\text{custom}}^{\ell}$ with the year index y omitted.

4.3 Performance by MeCab-Based Approach

This subsection shows performances of trading strategies based on the sentiment lexicon SL^{ℓ}_{MeCab}. Particularly, we compare performances using the following methods:

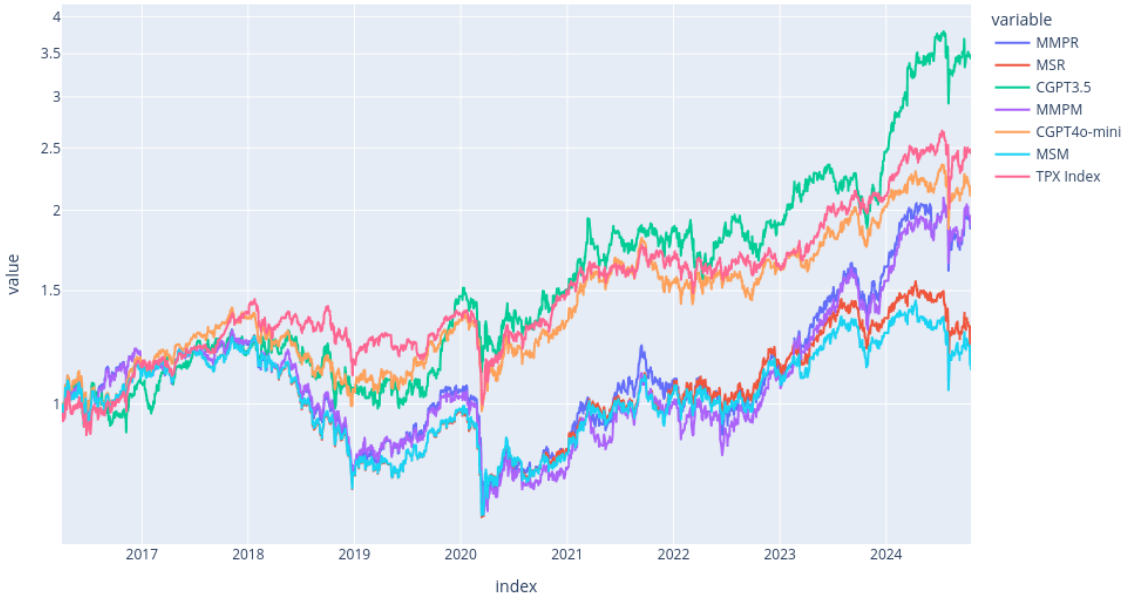
- **CGPT3.5**;
  CGPT3.5 applies ChatGPT3.5 to determine the sentiment of news headlines.
- **CGPT4o-mini**;
  CGPT4o-mini applies ChatGPT4o-mini to determine the sentiment of news headlines.
- **MeCab Simple Mean (MSM)**;
  MSM applies a financial dictionary (SL^{1}_{MeCab}), where polarity of a word is computed as the simple average of stock returns associated with that word in the news.
- **MeCab Market Premium Mean(MMPM)**;
  MPMM applies a financial dictionary (SL^{2}_{MeCab}), where polarity of a word is computed as the simple average of market premiums linked to that word in the news.
- **MeCab Stock Return Regression (MSR)**;
  MSR uses a financial dictionary (SL^{3}_{MeCab}) created by a regression model, and employs this model to determine polarity of a word based on its regression coefficient when regressed on stock returns.
- **MeCab Market Premium Regression (MMPR)**;
  MMPR uses a financial dictionary (SL^{4}_{MeCab}) created by a regression model, and applies this model to determine polarity of a word based on its regression coefficient when regressed on market premiums.

Table 4 and Figure 3 show the performance of each method. Here and hereafter, CR, SD, DD, MDD, ShR, SoR, and StR stand for compound return (CR), standard deviation (SD), downside deviation (DD), maximum drawdown (MDD), Sharpe ratio (ShR), Sortino ratio (SoR), and Sterling ratio (StR), respectively, each of which definition and explanation is given in Appendix: Performance measure.

Table 4: Performance Metrics (Sorted by Sharpe Ratio)

	CR	SD	DD	MDD	ShR	SoR	StR
CGPT3.5	14.57 %	23.38 %	13.61 %	26.01 %	62.31 %	107.01 %	56.00 %
TPX Index	10.68 %	17.54 %	11.23 %	31.42 %	60.89 %	95.10 %	33.99 %
CGPT4o-mini	8.40 %	18.40 %	11.44 %	30.91 %	45.64 %	73.36 %	27.16 %
MMPM	7.29 %	21.20 %	13.28 %	48.32 %	34.40 %	54.91 %	15.09 %
MMPR	7.11 %	21.62 %	13.70 %	47.62 %	32.90 %	51.93 %	14.94 %
MSR	1.99 %	19.89 %	12.63 %	47.84 %	9.99 %	15.73 %	4.15 %
MSM	1.18 %	19.85 %	12.65 %	47.37 %	5.93 %	9.30 %	2.49 %

Fig. 3: Cumulative Returns



It is observed in Table 4 that only the strategies with CGPT3.5 outperform a benchmark index, TOPIX denoted as TPX Index in terms of the compound return (CR) and all risk-adjusted returns such as Sharpe, Sortino and Sterling Ratios (ShR, SoR, StR). On the contrary, the performances of all of the MeCab-based strategies are worse than TOPIX.

To clarify the reason why MeCab-based methods do not work effectively, a closer look at the words contained in the created financial dictionary reveals that its segmentation accuracy is insufficient. For example, "SoftBank" is incorrectly divided into "Soft" + "Bank". Also, "四半期 (quarter period)" divided into "四半 (quarter) + 期 (period)" or "四 (four) + 半期 (half period)", fails to achieve the intended segmentation.

Therefore, to outperform the CGPT-based method by using a financial dictionary-based approach, it seems necessary to develop an alternative segmentation method which is not based on the existing MeCab approach.

4.4 Performance by Our Custom Dictionary Approach

This subsection shows performances of trading strategies based on our original dictionary developed in Section 3.1. Particularly, we compare performances using the following methods:

- **CGPT3.5**;
  CGPT3.5 applies ChatGPT3.5 to determine the sentiment of news headlines.
- **CGPT4o-mini**;
  CGPT4o-mini applies ChatGPT4o-mini to determine the sentiment of news headlines.
- **Custom Simple Mean (CSM)**;

CSM uses the sentiment lexicon (SL^1_{custom}), where polarity of a word is computed as the simple average of stock returns associated with that word in the news.

- **Custom Market Premium Mean (CMPM)**;
  CMPM uses the sentiment lexicon (SL^2_{custom}), where polarity of a word is computed as the simple average of market premiums linked to that word in the news.

- **Custom Stock Return Regression (CSR)**;
  CSR uses the sentiment lexicon (SL^3_{custom}) created with a regression model, and employs this model to determine polarity of a word based on its regression coefficient when regressed on stock returns.

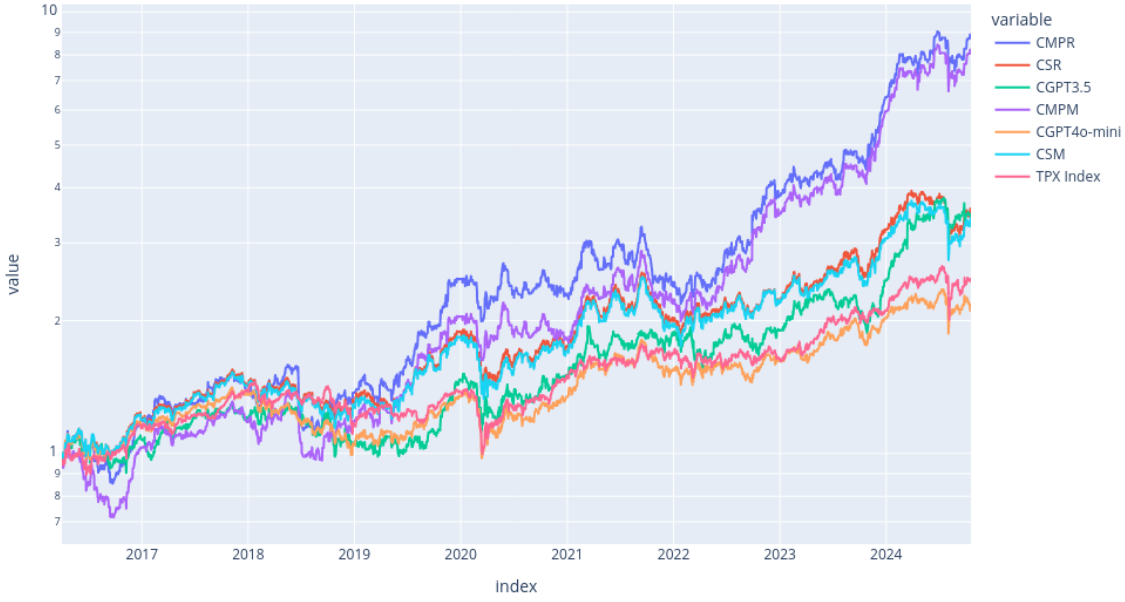
- **Custom Market Premium Regression (CMPR)**;
  CMPR uses the sentiment lexicon (SL^4_{custom}) created with a regression model, and applies this model to determine polarity of a word based on its regression coefficient when regressed on market premiums.

The performances of the six strategies are summarized in the following Table 5 and Figure 4.

Table 5: Performance Metrics (Sorted by Sharpe Ratio)

	CR	SD	DD	MDD	ShR	SoR	StR
CMPR	27.35 %	25.79 %	15.53 %	33.09 %	106.04 %	176.09 %	82.65 %
CMPM	26.22 %	25.65 %	15.46 %	31.45 %	102.21 %	169.53 %	83.35 %
CSR	14.98 %	21.85 %	13.07 %	29.62 %	68.59 %	114.62 %	50.59 %
CSM	14.34 %	21.78 %	13.04 %	30.40 %	65.82 %	109.96 %	47.16 %
CGPT3.5	14.57 %	23.38 %	13.61 %	26.01 %	62.31 %	107.01 %	56.00 %
TPX Index	10.68 %	17.54 %	11.23 %	31.42 %	60.89 %	95.10 %	33.99 %
CGPT4o-mini	8.40 %	18.40 %	11.44 %	30.91 %	45.64 %	73.36 %	27.16 %

Fig. 4: Cumulative Returns



It is observed in Table 5 that the CMPR-based strategy shows the best performance in terms of the compound return (CR) and all risk-adjusted returns such as Sharpe, Sortino and Sterling Ratios (ShR, SoR, StR). In addition, CMPR and CMPM-based strategies achieve higher performance than CGPT-based strategies in terms of the compound return (CR) and all risk-adjusted returns (ShR, SoR, StR). Namely, the strategies based on our original dictionary together with our regression model using market premium as a dependent variable, as well as simple average of market premium outperform the CGPT-based strategies.

Moreover, comparing with the MeCab-based strategies in Section 4.3, the performance based on our original dictionaries are much better, which implies that our original dictionaries can capture the specific context and terminology of financial news more accurately Moreover, by applying the 'market premium' to eliminate the impact of overall market trends, we have been able to more clearly extract firm-specific positive news factors, which has resulted in an improvement of the risk-adjusted returns.

However, it should be noted that there are certain days on which only a limited number of stocks are traded. Consequently, the investment performance on those days relies heavily on a small subset of stocks, potentially leading to unstable results and increasing investment risk. This issue will be addressed explicitly in the subsequent subsection.

4.5 Diversification

This subsection examines the effect of introducing diversification constraints to the CMPR and CGPT strategies, which ensure that trading is executed only if at least n (such as $n = 2, 4, 6, 8$) stocks have a positive sentiment score on a given day. This additional constraint aims to mitigate dependence on sentiment derived from only a small number of stocks, thereby

achieving more stable investment performance. The performance metrics of CMPR and CGPT with/without diversification strategies sorted by Sharpe Ratio are summarized in the table below.

Table 6: Performance Metrics for Diversified Portfolios (Sorted by Sharpe Ratio)

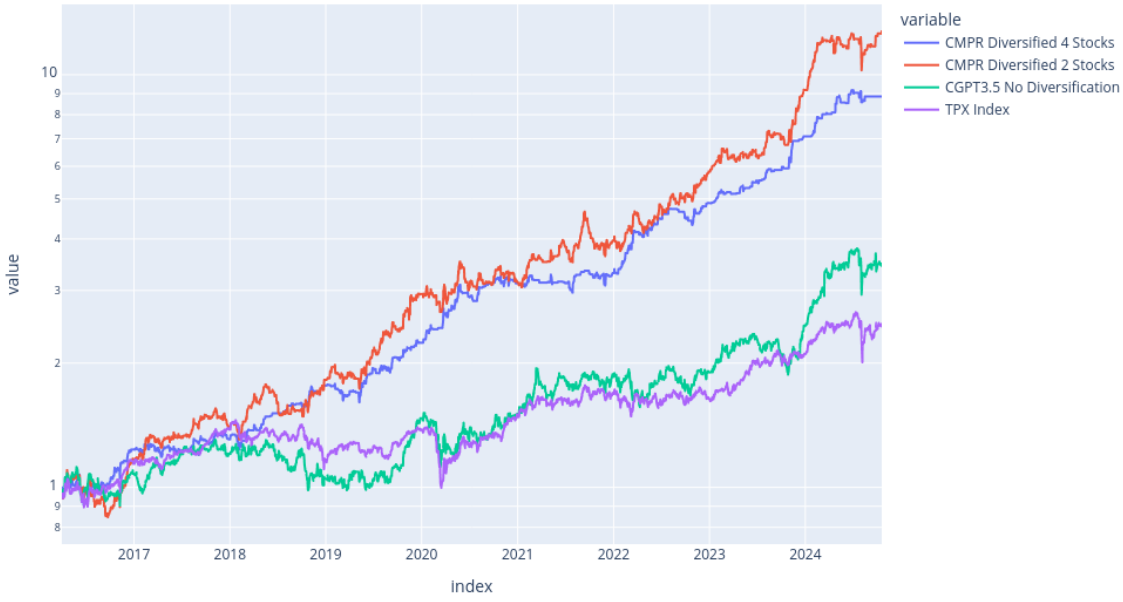
	CR	SD	DD	MDD	ShR	SoR	StR
CMPR Diversified 4 Stocks	27.53 %	14.52 %	8.29 %	15.55 %	189.60 %	332.20 %	177.08 %
CMPR Diversified 2 Stocks	32.70 %	21.06 %	12.26 %	23.03 %	155.31 %	266.64 %	141.97 %
CMPR Diversified 6 Stocks	13.70 %	9.82 %	5.51 %	8.19 %	139.49 %	248.58 %	167.37 %
CMPR No Diversification	27.35 %	25.79 %	15.53 %	33.09 %	106.04 %	176.09 %	82.65 %
CMPR Diversified 8 Stocks	6.03 %	6.81 %	4.01 %	6.69 %	88.62 %	150.43 %	90.09 %
CGPT3.5 Diversified 6 Stocks	9.56 %	12.51 %	7.94 %	20.43 %	76.40 %	120.40 %	46.76 %
CGPT3.5 Diversified 8 Stocks	6.80 %	9.27 %	6.32 %	18.36 %	73.40 %	107.69 %	37.04 %
CGPT3.5 Diversified 4 Stocks	12.11 %	16.91 %	10.39 %	21.12 %	71.59 %	116.52 %	57.31 %
CGPT3.5 Diversified 2 Stocks	13.99 %	21.38 %	12.69 %	24.54 %	65.46 %	110.28 %	57.03 %
CGPT3.5 No Diversification	14.57 %	23.38 %	13.61 %	26.01 %	62.31 %	107.01 %	56.00 %
TPX Index	10.68 %	17.54 %	11.23 %	31.42 %	60.89 %	95.10 %	33.99 %
CGPT4o-mini Diversified 4 Stocks	9.01 %	18.31 %	11.33 %	29.67 %	49.23 %	79.52 %	30.37 %
CGPT4o-mini Diversified 6 Stocks	8.86 %	18.19 %	11.27 %	29.43 %	48.69 %	78.62 %	30.10 %
CGPT4o-mini Diversified 2 Stocks	8.40 %	18.40 %	11.44 %	30.91 %	45.64 %	73.36 %	27.16 %
CGPT4o-mini No Diversification	8.40 %	18.40 %	11.44 %	30.91 %	45.64 %	73.36 %	27.16 %
CGPT4o-mini Diversified 8 Stocks	7.46 %	17.96 %	11.19 %	29.25 %	41.54 %	66.64 %	25.50 %

Compared to the results with no diversification, we observe that CMPR-based strategies with diversification except for 8 stocks improve Sharpe ratio. Especially, diversification with 4 stocks is most effective, which shows improvement in terms of the compound return (CR) and all risk-adjusted returns (ShR, SoR, StR). On the contrary, there are very few cases for simultaneous buying signals for 8 stocks, which makes its performance considerably worse.

Also, strategies based on CGPT3.5 with all diversification cases and CGPT4o-mini with diversification except 8 stocks outperform the corresponding no diversification strategies. However, all CGPT4o-mini-related strategies are still inferior to the benchmark index.

Next, Figure 5 below shows the time series of cumulative returns for our strategies generating the best (CMPR Diversified 2 stocks return = 32.70%) and second best (CMPR Diversified 4 stocks return = 27.53%) compound returns, as well as those of a CGPT-based strategy with the best compound returns (CGPT 3.5 No Diversification return = 14.57%) and the benchmark TOPIX (TPX Index return = 10.68%).

Fig. 5: Cumulative Returns: Top 2 with CGPT3.5 No Diversification



4.6 Trading Strategy with Futures

First, we note that futures contracts on TOPIX are the most liquid trading instruments with the lowest transaction costs in the Japanese equity market. Hence, to show reliable results through simulations, which should be most likely to realize in practice, this subsection investigates the performance of strategies by using TOPIX futures as a trading instrument combined with our original lexicons (dictionaries).

4.6.1 Daily Calculation of the market sentiment score

Since the closing time for TOPIX futures is 15:15, we collect news released until 15:12 from 0:00 for the TOPIX futures strategy.

Namely, at 15:12 on each trading day d in year y, we collect all news headlines concerning companies i within the TOPIX 500 that were announced between 0:00 and 15:12 on day d. These headlines are denoted by $\{N_{i,j}^{m,d,y}\}_{j,m}$. In this notation, the index j identifies each distinct news headlines, while the index m distinguishes multiple occurrences of the same headline if it appears more than once on day d.

News announced after 15:12 until 15:15 are not considered in this strategy, because we do not have sufficient time to incorporate information during the last 3 minutes into our positions. Moreover, our separate research reveals that news released from 15:15 to 0:00 do not have meaningful impacts on the stock prices at the closing time of the following day, which confirms our intuition.

Here, let us remark that we use the financial dictionary created with the information available up to the end of the last year, following the method described in Section 4.2. Consequently, the overall market sentiment score is obtained by the following procedure.

First, let us remind that the type ℓ sentiment score $S_{i,j}^{\ell,m,d,y}$ for the j-th news headline of company i on date d is given by:

$$\begin{aligned} S_{i,j}^{\ell,m,d,y} &:= \sum_{\text{word}_k \in SL_{y-1}^{\ell}} 1_{\text{word}_k \in N_{i,j}^{m,d,y}} P_{y-1,k}^{\ell}, \text{ for } \ell = 1,2,3,4, \\ S_{i,j}^{0,m,d,y} &= cgpt(N_{i,j}^{m,d,y}), \quad \text{for } \ell = 0, \end{aligned} \tag{22}$$

where $cgpt : N_{i,j}^{m,d,y} \to \{-1, 0, 1\}$. Then, we aggregate the sentiment scores $\{S_{i,j}^{\ell,m,d,y}\}_{j,m}$ on the day d to derive the daily sentiment score of company i. More precisely, let $J_i^{d,y}$ denote the total number of unique news headlines related to company i on day d in year y. For each unique news headline j (where $j = 1, 2, \ldots, J_i^{d,y}$), let $M_{i,j}^{d,y}$ represent the number of times the news headline j appears on the day d; i.e., $m = 1, 2, \ldots, M_{i,j}^{d,y}$

Thus, the overall sentiment score for company i on day d in year y is defined as

$$S_i^{\ell,d,y} := \sum_{j=1}^{J_i^{d,y}} \sum_{m=1}^{M_{i,j}^{d,y}} S_{i,j}^{\ell,m,d,y}. \tag{23}$$

Although the process of calculating sentiment scores for an individual company is identical to that described in Section 4.2, it is necessary in this strategy to additionally calculate a market sentiment score in order to determine positions for market index, i.e. TOPIX futures. Let us note that while TOPIX 500 is not exactly equivalent to TOPIX itself, companies included in TOPIX 500 account for more than 90% of the market capitalization among TOPIX. Therefore, considering factors such as news collection costs, we have decided to approximate TOPIX by TOPIX 500.

Concretely, we calculate the market sentiment score by aggregating sentiment scores for individual stocks in the following two way: the one market capitalization-weighted average method $S_{\text{TOPIX500}}^{\ell,d,y,1}$, and the other simple aggregation method $S_{\text{TOPIX500}}^{\ell,d,y,2}$. That is, when we denote $w_i^{d,y}$ as the market capitalization weight of company i in TOPIX on date d in year y, a sentiment score $S_{\text{TOPIX500}}^{\ell,d,y,k}$ $(k = 1, 2)$ is defined as:

$$\begin{aligned} S_{\text{topx500}}^{\ell,d,y,1} &:= \sum_{i \in \text{TOPIX500}^{d,y}} w_i^{d,y} S_i^{\ell,d,y}, \\ S_{\text{topx500}}^{\ell,d,y,2} &:= \sum_{i \in \text{TOPIX500}^{d,y}} S_i^{\ell,d,y}, \end{aligned} \tag{24}$$

where "TOPIX500^{d,y}" stands for the universe of TOPIX 500 in day d of year y.

4.6.2 Futures Position Construction

Next, we briefly explain how to construct a position for TOPIX futures based on the market sentiment score. Especially, if the overall market sentiment is positive, that is $S_{\text{topx500}}^{\ell,d,y,k} > 0$ $(k = 1, 2)$, we take a long position of TOPIX futures at 15:15. Also, if $S_{\text{topx500}}^{\ell,d,y,k} \leq 0$, we close all the position at 15:15. In this strategy, transaction costs for futures positions are set at 1 basis point (0.01%) in each way, applied separately to both making and closing futures positions.

4.6.3 Investment Performance

Following the procedure explained in the previous sections, we implement eight types of trading strategies based on our original financial dictionaries with ChatGPT3.5 and ChatGPT4o-mini. below.

- **TOPIX500_SM** and **TOPIX500_SM_mcap** ;
  Both TOPIX500_SM and TOPIX500_SM_mcap use a financial dictionary (SL^1_{custom}) for polarity calculation of sentiment score of each company. As for the calculation of market sentiment score, the former uses the simple aggregation method ($S^{\ell,d,y,2}_{\text{TOPIX500}}$) while the latter employs the market capitalization-weighted average method ($S^{\ell,d,y,1}_{\text{TOPIX500}}$) in (24).

- **TOPIX500_MPM** and **TOPIX500_MPM_mcap** ;
  Both TOPIX500_MPM and TOPIX500_MPM_mcap use a financial dictionary (SL^2_{custom}). The remainder is identical to the above and is thus omitted here.

- **TOPIX500_SR** and **TOPIX500_SR_mcap** ;
  Both TOPIX500_SR and TOPIX500_MPM_SR use a financial dictionary (SL^3_{custom}). The remainder is identical to the above and is thus omitted here.

- **TOPIX500_MPR** and **TOPIX500_MPR_mcap** ;
  Both TOPIX500_MPR and TOPIX500_MPR_mcap use a financial dictionary (SL^4_{custom}). The remainder is identical to the above and is thus omitted here.

- **TOPIX500_CGPT_3.5** and **TOPIX500_CGPT_3.5_mcap** ;
  Both TOPIX500_CGPT_3.5 and TOPIX500_CGPT_3.5_mcap use ChatGPT3.5 to determine the sentiment of news articles. The remainder is identical to the above and is thus omitted here.

- **TOPIX500_CGPT_4o_mini** and **TOPIX500_CGPT_4o_mini_mcap**;
  Both TOPIX500_CGPT_4o_mini and TOPIX500_CGPT_4o_mini_mcap use ChatGPT 4o-mini to determine the sentiment of news articles. The remainder is identical to the above and is thus omitted here.

In addition, to incorporate our sentiment scores into TOPIX futures efficiently, we conduct the same analysis only for the constituent companies in the TOPIX Large 100 and TOPIX Core 30, which consist of top 100 and 30 largest market capitalization companies with high liquidity, respectively.

Table 7 below shows the resulting investment performances sorted by Sharpe Ratio for the cases of TOPIX 500 (e.g., denoted as TOPIX500_MPR), TOPIX Large 100 (e.g., denoted as TOPIX100_MPR) and TOPIX Core 30 (e.g., denoted as TOPIX30_MPR), as well as the benchmark TOPIX futures denoted as TOPIX Index (futures).

Table 7: Performance Metrics (Sorted by Sharpe Ratio)

	CR	SD	DD	MDD	ShR	SoR	StR
TOPIX100_MPR_mcap	18.59 %	12.35 %	7.23 %	15.84 %	150.46 %	257.11 %	117.37 %
TOPIX30_MPR_mcap	13.55 %	10.48 %	6.25 %	9.37 %	129.23 %	216.79 %	144.67 %
TOPIX100_MPR	15.09 %	11.76 %	6.95 %	10.60 %	128.28 %	217.11 %	142.29 %
TOPIX100_MPM_mcap	15.85 %	12.74 %	7.78 %	14.02 %	124.38 %	203.64 %	113.04 %
TOPIX100_MPM	14.69 %	12.04 %	7.26 %	12.88 %	121.96 %	202.21 %	114.01 %
TOPIX500_MPR_mcap	15.18 %	13.54 %	8.10 %	19.92 %	112.10 %	187.45 %	76.19 %
TOPIX30_MPR	11.71 %	10.53 %	6.40 %	12.66 %	111.21 %	182.96 %	92.54 %
TOPIX500_MPM_mcap	14.20 %	13.73 %	8.33 %	18.22 %	103.41 %	170.41 %	77.93 %
TOPIX30_MPM	11.13 %	10.80 %	6.78 %	12.73 %	103.02 %	164.04 %	87.38 %
TOPIX30_MPM_mcap	10.72 %	10.72 %	6.74 %	9.20 %	99.97 %	158.92 %	116.48 %
TOPIX100_SR_mcap	16.33 %	16.34 %	10.03 %	21.72 %	99.92 %	162.83 %	75.16 %
TOPIX30_SR_mcap	15.19 %	15.33 %	9.48 %	19.78 %	99.06 %	160.17 %	76.79 %
TOPIX30_SM_mcap	15.03 %	15.49 %	9.59 %	20.58 %	97.03 %	156.72 %	73.02 %
TOPIX100_SM_mcap	15.21 %	16.74 %	10.33 %	24.71 %	90.84 %	147.27 %	61.56 %
TOPIX30_SR	13.34 %	15.26 %	9.49 %	19.11 %	87.46 %	140.68 %	69.83 %
TOPIX100_SR	14.34 %	16.55 %	10.38 %	26.23 %	86.62 %	138.20 %	54.67 %
TOPIX100_SM	13.85 %	16.78 %	10.51 %	23.57 %	82.53 %	131.79 %	58.77 %
TOPIX500_SR_mcap	13.82 %	17.48 %	11.28 %	27.62 %	79.09 %	122.55 %	50.05 %
TOPIX30_SM	12.13 %	15.42 %	9.65 %	22.56 %	78.66 %	125.64 %	53.76 %
TOPIX500_CGPT_3.5_mcap	4.38 %	6.20 %	3.85 %	13.18 %	70.74 %	113.84 %	33.25 %
TOPIX500_CGPT_3.5	2.61 %	4.05 %	2.71 %	8.63 %	64.37 %	96.30 %	30.20 %
TOPIX500_MPR	8.11 %	12.83 %	8.41 %	18.50 %	63.19 %	96.33 %	43.81 %
TOPIX100_CGPT_3.5_mcap	4.04 %	6.39 %	3.94 %	13.89 %	63.18 %	102.51 %	29.07 %
TOPIX500_SR	11.30 %	18.04 %	11.84 %	29.76 %	62.61 %	95.39 %	37.96 %
TOPIX500_SM	11.43 %	18.48 %	11.99 %	30.69 %	61.88 %	95.35 %	37.25 %
TPX Index (future)	11.57 %	19.09 %	12.27 %	32.64 %	60.61 %	94.29 %	35.45 %
TOPIX500_CGPT_4o_mini_mcap	9.14 %	15.82 %	10.01 %	27.18 %	57.76 %	91.23 %	33.61 %
TOPIX500_SM_mcap	10.05 %	17.80 %	11.53 %	31.66 %	56.46 %	87.15 %	31.74 %
TOPIX500_CGPT_4o_mini	9.76 %	18.08 %	11.70 %	25.47 %	53.97 %	83.42 %	38.32 %
TOPIX30_CGPT_3.5_mcap	3.59 %	6.80 %	4.47 %	18.29 %	52.85 %	80.34 %	19.65 %
TOPIX30_CGPT_4o_mini_mcap	7.76 %	15.03 %	9.73 %	29.69 %	51.60 %	79.76 %	26.13 %
TOPIX100_CGPT_4o_mini	8.22 %	15.98 %	10.15 %	26.66 %	51.42 %	80.92 %	30.82 %
TOPIX100_CGPT_4o_mini_mcap	7.85 %	15.46 %	9.88 %	30.89 %	50.75 %	79.40 %	25.40 %
TOPIX500_MPM	6.15 %	13.04 %	8.53 %	17.70 %	47.19 %	72.14 %	34.75 %
TOPIX30_CGPT_4o_mini	6.92 %	15.11 %	9.71 %	28.96 %	45.81 %	71.31 %	23.90 %
TOPIX30_CGPT_3.5	2.68 %	6.00 %	3.84 %	14.80 %	44.58 %	69.69 %	18.08 %
TOPIX100_CGPT_3.5	2.56 %	5.80 %	3.74 %	15.17 %	44.10 %	68.40 %	16.86 %

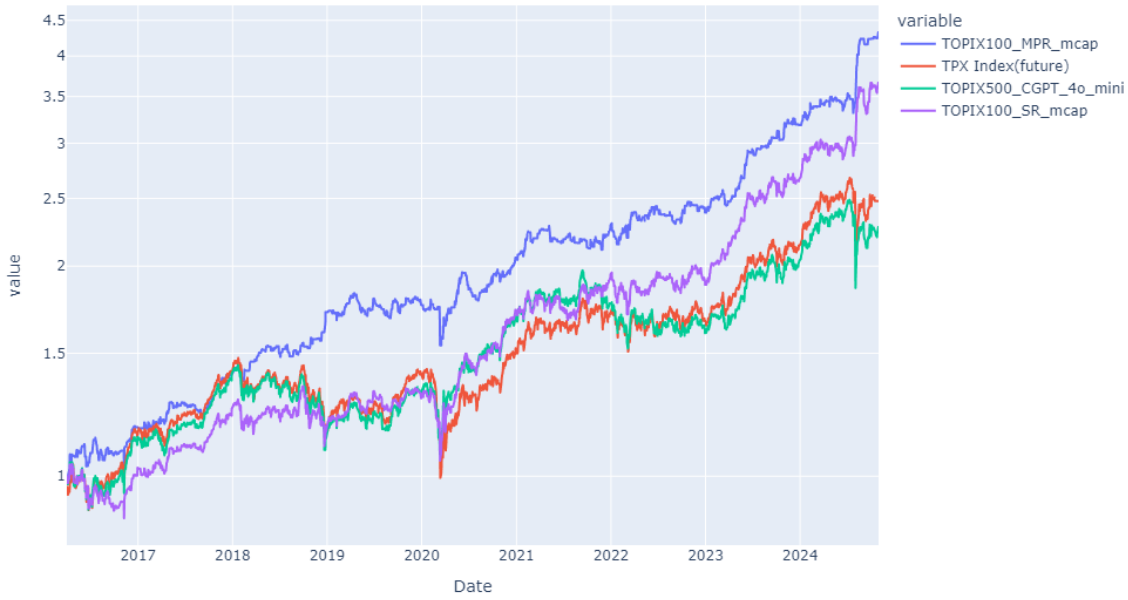
Firstly, it shows that in terms of Sharpe Ratio, most of our strategies (22 out of 24) outperform the benchmark TOPIX futures whose Sharpe Ratio=60.61%, while CGPT-based strategies do only 3 out of 12. Moreover, the best Sharpe Ratio created by TOPIX100_MPR_mcap among our strategies is more than 150%, while that by TOPIX500_CGPT_3.5_mcap among CGPT-based strategies is just around 71%.

In addition, we observed that under the fixed universe set, i.e. labeled as TOPIX30,

TOPIX100, or TOPIX500, our proposed market premium and regression based-strategy with market capitalization weighting, namely, labeled as 30, 100, or 500_MPR_mcap outperforms our other strategies, and substantially does ChatGPT-based methods as well as the benchmark TOPIX futures.

Next, Figure 6 below shows the time series of cumulative returns for our strategies generating the best (TOPIX100_MPR_mcap, 18.59%) and second best (TOPIX100_SR_mcap, 16.33%) compound returns, as well as those of a CGPT-based strategy with the best compound returns (TOPIX500_CGPT_4o_mini_mcap, 9.14%) and the benchmark TOPIX futures(11.57%).

Fig. 6: Cumulative Returns: Top2 with TOPIX500_CGPT_4o_mini



These results indicate that market return driven sentiment analysis, particularly with the use of market capitalization weight and custom dictionaries based on expert knowledge, significantly enhances investment performance beyond that achievable by generic language models such as ChatGPT.

5 Conclusion

This paper introduces a novel approach to sentiment analysis aimed at enhancing investment decision-making in the Japanese stock market. The proposed methodology begins by developing an original set of finance-specific keywords, extracted from news headlines published on a Japanese financial news platform. Using these keywords, sentiment lexicons are constructed by assigning polarity scores that are directly linked to corresponding market returns. In particular, by incorporating market premiums and regression models into the construction

of lexicons, the paper offers an innovative perspective on the integration of sentiment analysis into financial investment strategies.

Extensive empirical testing demonstrates that this approach significantly enhances investment performance compared to traditional natural language processing (NLP) methods, such as MeCab-based strategies, and advanced large language model-based approaches, specifically ChatGPT. The results highlight that integrating financial expert knowledge into sentiment analysis—via the construction of specialized dictionaries and the incorporation of market-premium, particularly in conjunction with regression techniques—leads to superior investment outcomes. Strategies utilizing the customized lexicons consistently achieve higher compound returns and improved risk-adjusted performance metrics, including Sharpe, Sortino, and Sterling ratios, outperforming those based on generic methods. These findings suggest that, despite the remarkable advancements in NLP models such as ChatGPT, domain-specific expertise combined with market-driven sentiment analysis remains critically valuable in generating alpha—risk-adjusted excess returns over the market index—in financial markets.

Acknowledgement

We appreciate Kyo Yamamoto and Soichiro Takahashi at GCI Asset Management Inc. for their valuable comments.

Appendix: Performance measure

Here, we summarize the definition of performance measures used in the current paper.

- Compound return (CR):

$$CR \equiv \left\{ \prod_{t=1}^{T} (1 + R_t) \right\}^{1/T} - 1. \tag{25}$$

  This is one of the most fundamental performance measures, which corresponds with a geometric average of the portfolio returns $\{R_t\}$.

- Standard deviation (SD):

$$SD \equiv \left\{ \frac{1}{T} \sum_{t=1}^{T} (R_t - \bar{R})^2 \right\}^{1/2}, \ \bar{R} \equiv \frac{1}{T} \sum_{t=1}^{T} R_t. \tag{26}$$

  This is one of the most basic variables both in theory and practice for portfolio risk management or derivatives pricing and hedging, which is also known as volatility.

- Downside deviation (DD):

$$DD \equiv \left\{ \frac{1}{T} \sum_{t=1}^{T} \min(0, R_t)^2 \right\}^{1/2}. \tag{27}$$

  Differently from SD, this risk measure regards only negative return as risk, which seems reasonable for investment performance evaluation.

- Maximum drawdown (MDD):

$$MDD \equiv \max_{1 \leq t \leq T} \frac{M_t - V_t}{M_t}, \quad M_t \equiv \max_{0 \leq s \leq t} V_s. \tag{28}$$

MDD is a famous concept in hedge fund risk management, where drawdown denotes a decline from the past peak value M_t to the present value V_t.

Shortly, this measure tells us the worst scenario for a given investment horizon. That is, it represents how much loss an investor suffers from if he/she enter and exit an investment at the worst timing.

As it is widely recognized in practice that investment performance largely depends on its starting and exiting timing, MDD is thought to be an important measure. Namely, small MDD implies that an investor has not suffered from a large loss, whenever he/she starts the investment, at least on the past data.

- Sharpe ratio (ShR):

$$ShR \equiv (\bar{R} - r_f)/SD, \tag{29}$$

where r_f denotes a risk-free rate. In investment performance evaluation, risk-adjusted returns are often regarded as the most important measures. Among them, ShR is the most famous one, which is also a basic quantity in the field of financial economics. In this paper, we assume $r_f = 0$ because Bank of Japan guides short-term rates at -0.1% and the 10-year bond yield around 0% during most of the test period.

- Sortino ratio (SoR):

$$SoR \equiv (\bar{R} - r_f)/DD. \tag{30}$$

SoR is also useful because it adjusts risk by using DD, which makes it possible to focus on only downside risk.

- Sterling Ratio (StR):

$$StR \equiv (\bar{R} - r_f)/MDD. \tag{31}$$

StR is a measure of risk-adjusted return that uses drawdown measures as denominator.

References

[1] Yoo, P. D., Kim, M. H., & Jan, T. (2005, November). Machine learning techniques and use of event information for stock market prediction: A survey and evaluation. In International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06) (Vol. 2, pp. 835-841). IEEE.

[2] Fama, E. F. (1965). The behavior of stock-market prices. The journal of Business, 38(1), 34-105.

[3] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The journal of Finance, 25(2), 383-417.

[4] Campbell, J. Y., & Thompson, S. B. (2007). Predicting excess stock returns out of sample: Can anything beat the historical average?. The Review of Financial Studies, 21(4), 1509-1531.

[5] Nakano, M., Takahashi, A., & Takahashi, S. (2017). Generalized exponential moving average (EMA) model with particle filtering and anomaly detection. Expert Systems with Applications, 73, 187-200.

[6] Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. Advances in neural information processing systems, 13.

[7] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[8] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[9] Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. Journal of machine learning research, 2(Feb), 419-444.

[10] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.

[11] McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, No. 1, pp. 41-48).

[12] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of finance, 62(3), 1139-1168.

[13] Garcia, D. (2013). Sentiment during recessions. The journal of finance, 68(3), 1267-1300.

[14] Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016, October). Sentiment analysis of Twitter data for predicting stock market movements. In 2016 international conference on signal processing, communication, power and embedded system (SCOPES) (pp. 1345-1350). IEEE.

[15] Sun, T., Wang, J., Zhang, P., Cao, Y., Liu, B., & Wang, D. (2017, August). Predicting stock price returns using microblog sentiment for chinese stock market. In 2017 3rd International Conference on Big Data Computing and Communications (BIGCOM) (pp. 87-96). IEEE.

[16] Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554.

[17] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems, 2(4), 303-314.

[18] Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. Expert Systems with Applications, 42(20), 7046-7056.

[19] Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. Expert Systems with Applications, 83, 187-205.

[20] Huck, N. (2010). Pairs trading and outranking: The multi-step-ahead forecasting case. European Journal of Operational Research, 207(3), 1702-1716.

[21] de Oliveira, F. A., Nobre, C. N., & Zarate, L. E. (2013). Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index-Case study of PETR4, Petrobras, Brazil. Expert Systems with Applications, 40(18), 7596-7606.

[22] Nakano, M., Takahashi, A., & Takahashi, S. (2018). Bitcoin technical trading with artificial neural network. Physica A: Statistical Mechanics and its Applications, 510, 587-609.

[23] Nakano, M., & Takahashi, A. (2020). A new investment method with AutoEncoder: Applications to crypto currencies. Expert Systems with Applications, 162, 113730.

[24] Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? return predictability and large language models. arXiv preprint arXiv:2304.07619.

[25] Okimoto, T., & Hirasawa, E. (2014). Stock market predictability using news indexes. Security Analysis Journal, 52(4), 67-75.

[26] Goshima, K., & Takahashi, H. (2016). Quantifying news tone to analyze the Tokyo Stock Exchange with deep learning. Security Analysis Journal, 54(3), 76-86.

[27] Katayama, D., & Tsuda, K. (2018). A method of measurement of the impact of Japanese news on stock market. Procedia computer science, 126, 1336-1343.

[28] Akita, R., Yoshihara, A., Matsubara, T., & Uehara, K. (2016, June). Deep learning for stock prediction using numerical and textual information. In 2016 IEEE : ACIS 15th International Conference on Computer and Information Science (ICIS) (pp. 1-6). IEEE.

[29] Nishimura, K. G., Sato, S., Akihiko Takahashi, A. Term Structure Models During the Global Financial Crisis: A Parsimonious Text Mining Approach. Asia-Pacific Financial Markets, Volume 26, pages 297–337.

[30] Nakatani, S., Nishimura, K. G., Saito, T., Akihiko Takahashi, A. (2020). Interest Rate Model with Investor Attitude and Text Mining. IEEE Access, Volume 8, pages 86,870 - 86,885.

[31] Nakano, M., & Yamaoka, T. (2023). Enhancing Sentiment Analysis based Investment by Large Language Models in Japanese Stock Market. SSRN id=4511658.