CIRJE-F-1267

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 2026

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 4, 2026

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

Stock price prediction is a central but challenging problem in financial investment, because market dynamics reflect economic, political, and psychological forces (Yoo et al., 2005). While researchers have traditionally relied on numerical market variables (e.g., returns, volatility, dividend yields), such data often miss contextual and behavioral information, motivating the use of text data and NLP methods.

Initially, many researchers have focused on the market information as shown in prior work (Fama, 1965, 1970; Campbell and Thompson, 2007; Nakano et al., 2017) 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) (Bengio et al., 2000; Blei et al., 2003; Pennington et al., 2014; Lodhi et al., 2002; Mikolov et al., 2013; McCallum and Nigam, 1998).

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 their emergence, the mainstream of sentiment analysis is rule based approach, where predefined linguistic rules by keywords, patterns, and linguistic heuristics are used to determine sentiment. See Tetlock (2007) and Garcia (2013) for analysis in the financial field.

Therefore, researchers have developed new approaches, such as Word2Vec (Mikolov et al., 2013), 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. (2016) and Sun et al. (2017), for instance.

Furthermore, ANNs have become increasingly important in text data analysis with the development of deep learning techniques (Hinton et al., 2006), 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 (1989), a growing number of studies have applied ANNs to financial investment problems (Ballings et al., 2015; Chong et al., 2017; Huck, 2010; de Oliveira et al., 2013; Nakano et al., 2018; Nakano and Takahashi, 2020; Lopez-Lira and Tang, 2023).

In particular, the recent work by Lopez-Lira and Tang (2023) exploits news headlines to predict stock price movements in the U.S. market using sentiment scores derived from ChatGPT, which 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-Lira and Tang (2023) demonstrate 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.

To extend this line of inquiry, Nakano and Yamaoka (2023) examine the application of ChatGPT to Japanese financial news over a longer sample period. Their findings suggest that textual information also contains economically meaningful signals in the Japanese market. However, two important issues remain open: whether text-based asset pricing mecha-

nisms documented in English language and U.S. settings generalize to different linguistic and institutional environments, and whether LLM based sentiment measures are economically interpretable.

First, most existing evidence on text-based asset pricing relies on English language corpora and U.S. market institutions (Tetlock, 2007; Loughran and McDonald, 2011; Li, 2010; Loughran and McDonald, 2016). It is therefore unclear whether the documented pricing mechanisms possess external validity, that is, whether they generalize to markets characterized by different linguistic structures and institutional environments. In addition to language specific features, institutional features may also shape how information is conveyed and processed in financial markets.

For instance, prior evidence suggests that Japanese managers have incentives to avoid explicitly forecasting losses and may exhibit optimistic bias in management forecasts (Cho et al., 2011), which may be consistent with corporate narratives and related reporting practices conveying unfavorable information in relatively formal and indirect ways. This institutional and linguistic environment may affect how quickly and how accurately investors extract sentiment from headlines, making Japan a useful laboratory to test whether text-to-return mechanisms documented in English language settings generalize beyond U.S. markets.

From a financial economics perspective, this cross market test helps distinguish general information processing frictions from language or institution specific communication artifacts. In particular, if negative information is conveyed more implicitly, the mapping from textual tone to prices may be weaker and more sign-asymmetric around bad news, which has economically meaningful implications for price discovery and return predictability.

Since language specific features and domain specific vocabulary have been shown to materially affect sentiment classification (Li, 2010), examining non-English markets is essential for assessing the generalizability of prior findings. Related work in textual finance shows that linguistic conventions and disclosure style can materially affect measured tone and its economic interpretation (Li, 2010; Loughran and McDonald, 2016). This further suggests that sentiment measures calibrated in English language settings may transfer imperfectly to Japanese, potentially leading to information loss or systematic distortion.

To make this argument empirically testable, we examine whether the mapping from textual signals to stock returns differs systematically between positive and negative news in the Japanese information environment. In particular, if unfavorable information tends to be communicated more implicitly in Japanese financial news, price responses to negative textual signals may be more muted and more sign-asymmetric than responses to positive signals. We evaluate this prediction in our empirical analysis, including quartile-wise tests that stratify firms by news coverage to mitigate concerns that the results are driven only by high attention firms. Consistent with this hypothesis, our event based tests in Section 4.4.4 provide suggestive evidence that positive signals remain more statistically significant than negative signals overall.

Second, while LLM based approaches exhibit strong predictive performance, their black-box nature limits economic interpretability. This matters not only for transparency: interpretability allows us to test which linguistic components move prices and whether the resulting return predictability is consistent with economically meaningful mechanisms (e.g., delayed information processing versus generic correlations).

In empirical economics, interpretability is particularly important because researchers aim to relate statistical regularities to economically meaningful mechanisms rather than relying solely on predictive accuracy (Mullainathan and Spiess, 2017; Loughran and McDonald,

2016). Even when LLM based sentiment measures perform well, it remains difficult to identify which specific linguistic components drive return predictability. This motivates an explicitly constructed, word-level lexicon in which the mapping from linguistic components to return predictability can be inspected and tested.

These two considerations, namely external validity and interpretability, motivate the present study. We focus on the Japanese stock market, one of the largest and most liquid equity markets globally, making it a natural laboratory in which economically meaningful pricing effects can be evaluated. This enables us to speak to a core question in financial economics: whether textual return predictability reflects universal investor information processing frictions or market specific communication norms. At the same time, Japan offers a distinct information environment in linguistic and institutional terms, which provides a stringent setting to evaluate whether text-based asset pricing mechanisms generalize beyond the U.S. context.

While prior Japanese studies have explored textual information in financial markets, their focus has been different from ours. For example, Okimoto and Hirasawa (2014) and Goshima and Takahashi (2016) analyze the relationship between news content and aggregate market indices such as TOPIX. Nishimura et al. (2019) and Nakatani et al. (2020) incorporate textual factors into yield curve modeling using morphological analysis tools such as MeCab. Other studies, including Katayama and Tsuda (2018) and Akita et al. (2016), investigate sentiment effects at the firm level using conventional dictionary based or embedding approaches.

Taken together, prior Japanese studies show that text contains economically meaningful information, but an important gap remains for equity investment applications. Existing approaches either focus on aggregate outcomes or rely on sentiment measures whose mapping to investable performance is indirect and difficult to interpret at the lexical level. Moreover, for Japanese headlines, segmentation errors driven by firm specific proper nouns and finance specific terms can materially affect the construction of reliable word based signals.

Our study addresses these issues by developing an investment oriented and interpretable sentiment framework and by evaluating its incremental value against established NLP and LLM benchmarks. Importantly, in the Japanese context, prior studies often do not evaluate firm-level implementable strategies under a unified protocol that (i) benchmarks directly against LLM-based sentiment measures and (ii) tests whether any return predictability survives standard asset-pricing controls. Our design is intended to isolate the incremental contribution of an interpretable, return-disciplined lexicon relative to these established baselines.

Motivated by these gaps, our contribution lies in developing an interpretable, investment oriented sentiment framework that complements macro-level LLM approaches. Here, we note that a key limitation of ChatGPT is its black-box nature of the response generation process, which motivates us to explore interpretable methods such as a polarity approach to calculate sentiment scores.

In contrast to macro level LLM sentiment signals, our method emphasizes a microlevel, word level perspective by developing an original polarity dictionary grounded in financial expert knowledge. The procedure is explicitly designed to preserve finance specific terminology, retain only tokens that are plausibly relevant for investment decisions, and discipline the dictionary using realized market reactions. In particular, this paper proposes a new method using market return data to construct an original polarity dictionary, and our empirical studies show that it can outperform a ChatGPT based sentiment approach for financial investment in the Japanese stock market.

To clarify the incremental contribution relative to existing NLP based work, we benchmark

our framework against two baselines: (i) a ChatGPT based sentiment measure and (ii) a standard bag of words lexicon baseline that computes sentiment from token occurrences (i.e., a BoW representation based on token presence) using an externally constructed polarity dictionary (Ito et al., 2018). We then test whether the resulting strategy returns remain after controlling for standard risk factors, e.g., Fama–French three-factor model (Fama and French, 1993), and show that our approach delivers positive and statistically significant factor adjusted performance whereas baseline methods do not under the same evaluation protocol.

Specifically, the construction of a polarity dictionary is divided into two steps: (i) creating a set consisting of target words extracted by morphological analysis and (ii) 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. For Step (i), 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 correct and economically meaningful segmentation due to company specific proper nouns and finance related specialized terms. Hence, we propose a simple and effective method that improves MeCab segmentation: 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 those whose average returns underperform transaction costs).

For Step (ii), let us remark that the polarities of words in traditional approaches are generally based on positive or negative labels manually attached to text by volunteers without financial expertise. Therefore, traditional methods do not take expert knowledge and the real market reaction into account at all, which is also relevant for large language models (LLMs) because market stock return data are not used directly in the training of LLMs to learn return labeled mappings from text. 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 whole market movements. These concepts provide a basis for our sentiment analysis approach, which integrates textual data with market driven insights and contrasts 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 domain specific sentiment in financial news, learned directly from realized market reactions, while providing an interpretable link between textual components and asset pricing outcomes.

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 Oct 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.

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 A defines the performance measures used throughout the paper, and Appendices B–D provide additional methodological details and supplementary empirical results.

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

We collect news headlines for the companies in our investment universe from October 1, 2013 to October 18, 2024 from Kabutan, a leading Japanese financial news platform widely utilized by domestic retail investors [1].

The collected headlines cover a broad spectrum of topics, including company-specific developments, industry-related events, and macroeconomic updates. At the same time, we exclude news directly associated with technical analysis, because the purpose of this study is to examine the impact of news on stock price movements independently of technical analysis.

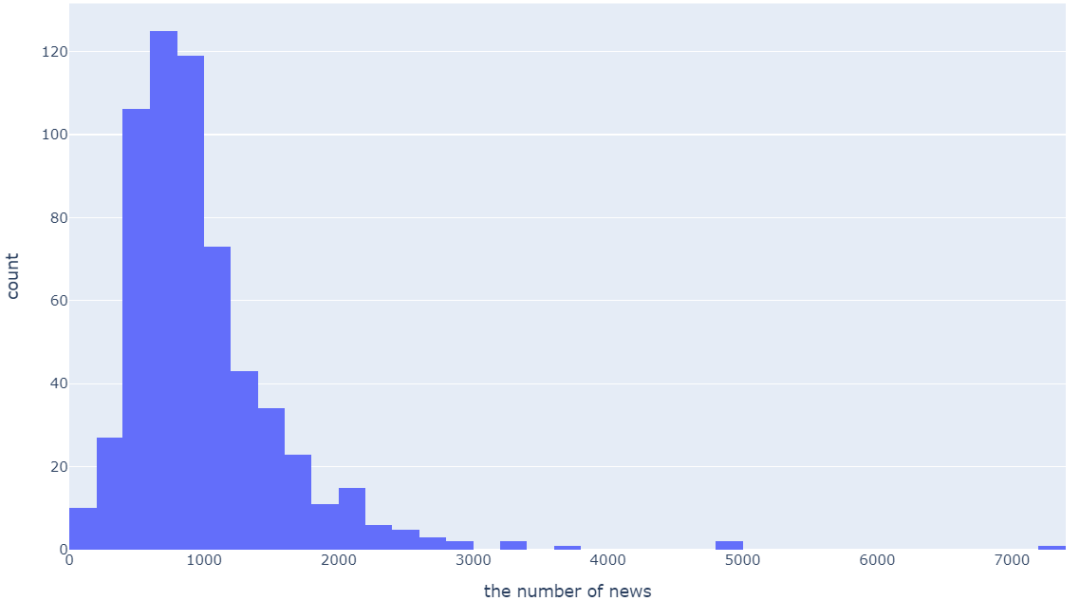
We also note that the headline categories provided by the original data source are pre-classified by the provider. Hence, when we refer to category-based distinctions in the subsequent analysis, we use the provider's classification labels and do not manually assign categories ourselves.

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.

[1] https://kabutan.jp/

Fig. 1: Number of News and Company



To further characterize this heterogeneity in information exposure, we examine how firm size varies with news coverage. Specifically, we sort firms in ascending order of M_i and construct four fixed quartiles $Q_1, \ldots, Q_4$ (from low to high news coverage). We summarize firm size by computing the average market capitalization for each quartile as of 2024/10/18.

Table 1: Average market capitalization by news-frequency quartile (2024/10/18)

quartile	Average market capitalization (JPY million)
Q_1	3.864874×10^5
Q_2	5.807455×10^5
Q_3	1.153306×10^6
Q_4	3.672440×10^6

Table 1 shows that firms in higher news-frequency quartiles tend to have larger market capitalizations. This relationship motivates our later robustness analysis in Section 4.4.3-4.4.4 that evaluates strategy performance within the same news-exposure strata. For clarity, this news-coverage quartile classification is used solely for ex-post stratification in our robustness analysis, and it is not used in our main analysis.

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 2 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 2: 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.

2.3 Firm Characteristics and Risk-Free Rate

In addition to daily prices, we obtain firm characteristics from Bloomberg. Specifically, we retrieve market capitalization and the price-to-book ratio (PBR). Market capitalization is used (i) to summarize firm size in Table 1 and (ii) to construct market-capitalization-weighted variants of our strategy. PBR is used as a value characteristic; when constructing the TOPIX500-based factor controls, we define book-to-market as the inverse of PBR (see Appendix B for details).

For the factor-model analysis, we proxy the daily risk-free rate RF_t using TONAR (Tokyo Overnight Average Rate), obtained from Bloomberg and converted to a daily rate.

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: $N \in \mathcal{F}_{\text{pattern}} \rightarrow$ (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 3: 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 [2]:

$$N_k^{NF} = s_0^k \; d_1^k \; s_1^k \; d_2^k \cdots d_{n_k}^k \; s_{n_k}^k, \quad d_i^k \in D,$$

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.

(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 4 shows all the elements in WL_{custom}.

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

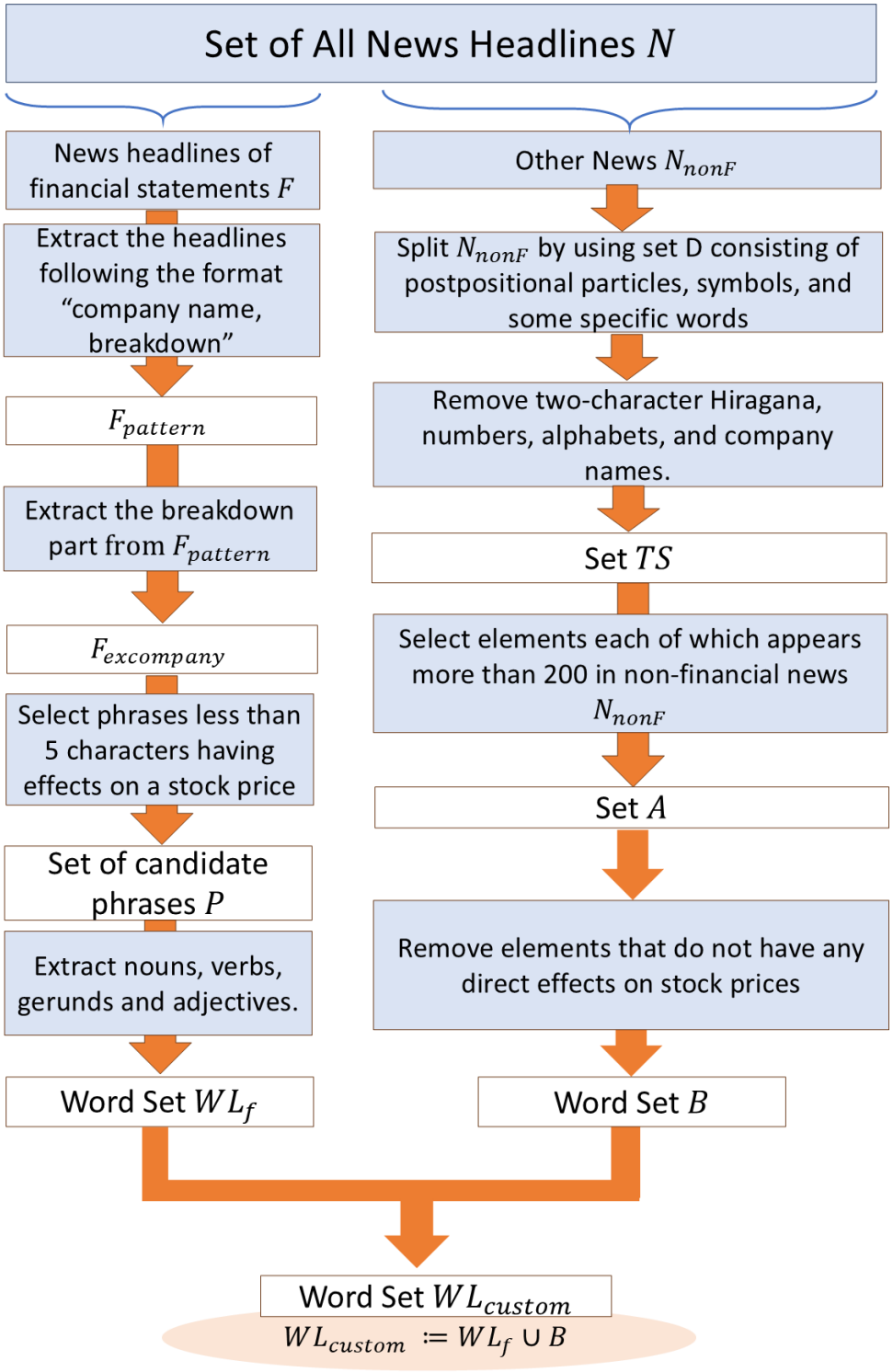
Table 4: Key Japanese Words (their English translations) in WL_{custom}

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 $\{\text{ts}^m_{i,j}\}^{M_i,m^j_i}_{j=1,m=1}$, 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 consisting 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 three 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 ChatGPT directly whether a news headline has a positive or negative effect on stock prices. Specifically, we apply ChatGPT[3] 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 Bag of Words Baseline with an External Polarity Dictionary

As an additional benchmark, we introduce a standard bag-of-words (BoW) sentiment method based on an external financial-domain polarity dictionary. In this study, we use the publicly available Japanese dictionary associated with Ito et al. (2018), and we refer to this benchmark as BoW in the following empirical analysis.

Let the external sentiment lexicon be denoted by

$$SL_{\text{BoW}} := \{(\text{word}_k, P_k^{\text{BoW}})\}_k,$$

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

where P_k^{BoW} is the polarity score of word$_k$ given in the external dictionary.

Using SL_{BoW}, we compute the sentiment score of the j-th news headline of company i, denoted by $S_{i,j}^{\text{BoW}}$:

$$S_{i,j}^{\text{BoW}} = \sum_{\text{word}_k \in SL_{\text{BoW}}} \mathbf{1}_{\text{word}_k \in N_{i,j}} P_k^{\text{BoW}},$$

where $\mathbf{1}_{\text{word}_k \in N_{i,j}}$ is a binary indicator that equals 1 if word$_k$ appears in the news headline $N_{i,j}$, and 0 otherwise.

Thus, unlike our proposed methods, this benchmark uses externally provided polarity information and does not estimate word polarity from stock-price reactions or market premiums in our sample. The BoW benchmark is evaluated under the common backtesting framework described in Section 4.2.

4.1.3 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 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 $(\{P^{\ell}_{y-1,k}\}_k)$ and SL^{ℓ}_{y-1}, respectively. Formally, the sentiment lexicons are represented as follows:

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

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

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

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

$$S^{\ell,m,d,y}_{i,j} = \sum_{\text{word}_k \in SL^{\ell}_{y-1}} 1_{\text{word}_k \in N^{m,d,y}_{i,j}} P^{\ell}_{y-1,k}. \tag{19}$$

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

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

For each company i, all the sentiment scores $\{S^{\ell,m,d,y}_{i,j}\}_{j,m}$ corresponding to the news headlines $\{N^{m,d,y}_{i,j}\}_{j,m}$ 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

$$S^{\ell,d,y}_i := \sum_{j=1}^{J^{d,y}_i} \sum_{m=1}^{M^{d,y}_{i,j}} S^{\ell,m,d,y}_{i,j}. \tag{21}$$

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^{ℓ}_{MeCab} and $SL^{\ell}_{\text{custom}}$ 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.
- **BoW (external polarity dictionary)**;
  BoW is a bag-of-words baseline using an external polarity dictionary; see Section 4.1.2.
- **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)**;
  MMPM 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.

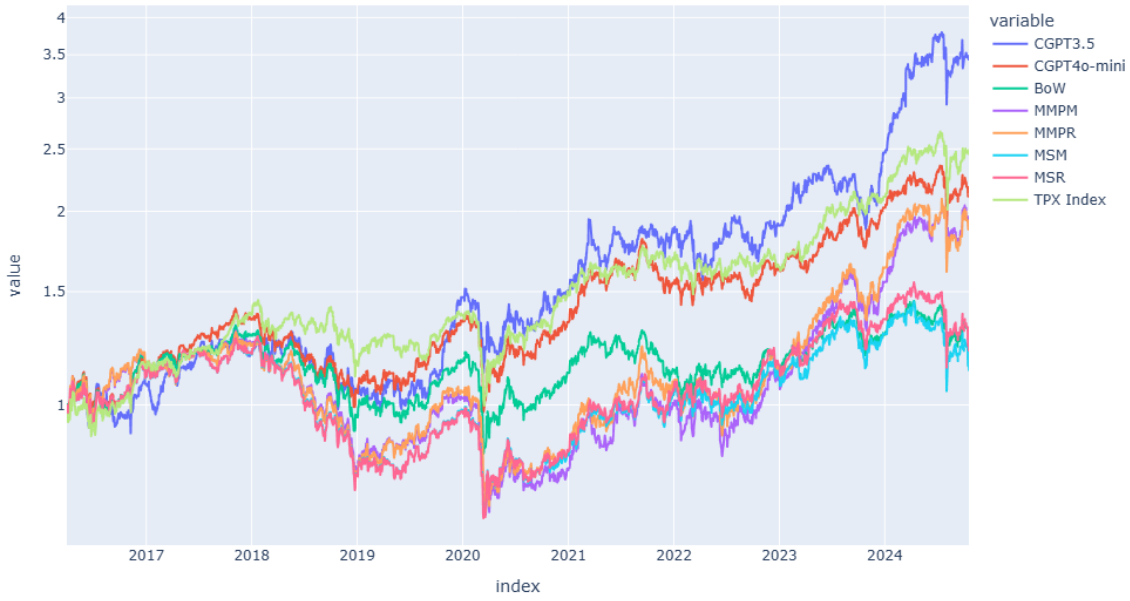
Table 5 and Figure 3 show the performance of each method. Here and hereafter, CR, SD, DD, MDD, ShR, SoR, and StR stand for compound return (CR), standard deviation

(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 A

Table 5: 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 %
BoW	2.40 %	18.18 %	11.56 %	36.80 %	13.18 %	20.73 %	6.51 %
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 5 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. The external BoW baseline also fails to outperform TPX Index, although it performs better than the weakest MeCab variants.

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

4.4.1 Baseline results

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.
- **BoW (external polarity dictionary)**;
  BoW is a bag-of-words baseline using an external polarity dictionary; see Section 4.1.2.
- **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 seven strategies are summarized in the following Table 6 and Figure 4.

Table 6: 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 %
BoW	2.40 %	18.18 %	11.56 %	36.80 %	13.18 %	20.73 %	6.51 %

Fig. 4: Cumulative Returns

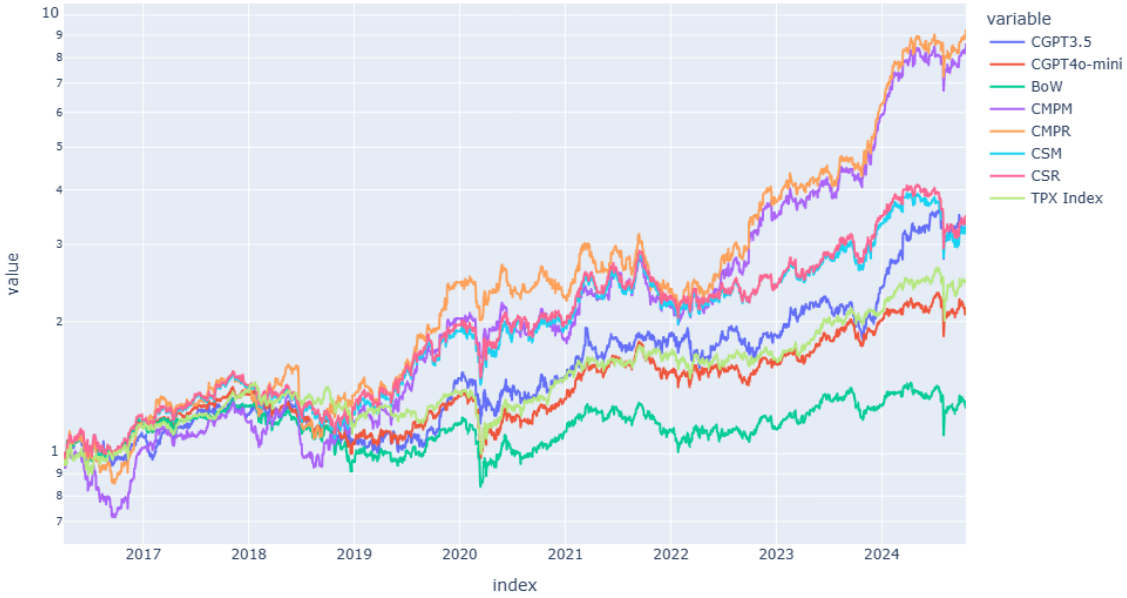


Table 6 summarizes the performance metrics of each strategy. The CMPR-based strategy achieves the highest compound return (CR) and the strongest risk-adjusted performance in terms of the Sharpe and Sortino ratios (ShR and SoR). The CMPM-based strategy performs similarly and attains the highest Sterling ratio (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, compared with the MeCab-based strategies in Section 4.3, the performance based on our original dictionaries is 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 Section 4.5.

4.4.2 Factor-Model Analysis with TOPIX500-Based Three-Factor Controls

To examine whether the excess returns of the sentiment-based strategies can be explained by conventional risk factors, we implement daily time-series regressions using Fama–French 3 factor controls constructed on the TOPIX500 universe following Chen et al. (2025). See Appendix B for the detailed construction procedure of these factor returns.

Following the Fama–French 3 factor model (Fama and French, 1993), we estimate the following regression for each portfolio p:

$$R_{p,t} - \mathrm{RF}_t = \alpha_p + \beta_{p,\mathrm{MKT}} \left(\mathrm{MKT}_t - \mathrm{RF}_t\right) + \beta_{p,\mathrm{SMB}} \, \mathrm{SMB}_t + \beta_{p,\mathrm{HML}} \, \mathrm{HML}_t + \varepsilon_{p,t}, \tag{22}$$

where $R_{p,t}$ is the daily return of strategy portfolio p, RF_t is the daily risk-free return, and SMB_t and HML_t denote the size and value factors, respectively. We define MKT_t as the daily return of the TOPIX500 total return index, and use $\mathrm{MKT}_t - \mathrm{RF}_t$ as the market excess return in the regression. The intercept α_p is interpreted as the factor-adjusted alpha (i.e., abnormal return unexplained by the three factors).

Table 7 reports the regression results. The estimated alphas for CMPR and CMPM remain positive and statistically significant after controlling for market, size, and value exposures, indicating that their performance cannot be fully explained by these three conventional factors alone. In contrast, the benchmark methods generally show weaker factor-adjusted alphas, with some estimates being statistically insignificant and BoW exhibiting a significantly negative alpha.

Table 7: Fama–French-Type Three-Factor Regression Results (TOPIX500-Based Factors)

	α(%)	MKT-RF	SMB	HML	Annualized α(%)
CMPR***	0.08(2.67)	0.75(25.96)	0.94(3.90)	-0.19(-1.99)	21.31
CMPM**	0.08(2.58)	0.74(25.87)	0.85(3.54)	-0.20(-2.11)	20.51
CSR	0.03(1.17)	0.90(43.48)	0.99(5.73)	-0.22(-3.32)	6.68
CGPT3.5	0.03(1.13)	0.82(34.22)	1.02(5.09)	-0.29(-3.80)	7.50
CSM	0.02(1.07)	0.90(43.70)	0.92(5.37)	-0.22(-3.38)	6.08
CGPT4o-mini	-0.00(-0.17)	0.91(74.66)	0.78(7.70)	-0.15(-3.87)	-0.58
BoW**	-0.03(-2.50)	0.95(93.29)	0.90(10.61)	-0.19(-5.95)	-7.01

Notes: This table reports coefficients from daily OLS time-series regressions of strategy excess returns on TOPIX500-based Fama–French-type three-factor controls (MKT-RF, SMB, HML). Values in parentheses are t-statistics. α denotes the factor-adjusted abnormal return (intercept). Annualized α is computed as $252 \times \alpha$ using the (unrounded) daily intercept estimates. Significance stars (*, **, ***) are attached to the strategy name to indicate that the intercept α is statistically significant at the 10%, 5%, and 1% two-sided levels, respectively.

4.4.3 Robustness to Heterogeneous News Coverage

As shown in Figure 1, the number of news headlines differs substantially across firms, raising the concern that the strong performance of the CMPR-based strategy may be disproportionately driven by a small set of high-attention firms. To examine whether the CMPR signal remains effective within groups of firms with similar news exposure, we conduct a quartile-wise robustness check by stratifying firms into quartiles based on their firm-level news exposure.

Specifically, for each quartile Q_k, year y, and trading date d, we define the set of positive-signal firms within quartile Q_k as follows:

$$I^{+}_{k,d,y} := \left\{ i \in Q_k \,\middle|\, S_i^{4,d,y} > 0 \right\}, \tag{23}$$

where $S_i^{4,d,y}$ denotes the CMPR-based daily sentiment score for firm i, constructed using the same daily score construction method as in Section 4.2.

The quartile-wise CMPR_Q_k strategy is implemented by applying the same backtesting rule as in Section 4.2 with the investable universe restricted to Q_k: at each daily rebalancing step, we take long positions only in firms in $I^{+}_{k,d,y}$, allocate positions equally across the selected firms, and close positions when the signal becomes non-positive. The performance is evaluated by the same performance measures under the same transaction-cost condition as in Section 4.4.1.

As a quartile-specific benchmark, we also compute the equally weighted (EW) daily return of all firms in the same quartile:

$$R^{\mathrm{EW}}_{k,d} := \frac{1}{\#Q_k} \sum_{i \in Q_k} r^{cc}_{d,i}, \tag{24}$$

To ensure a consistent comparison with this EW benchmark, which is computed without trading cost, we also report the quartile-wise CMPR_Q_k strategy performance gross of transaction costs. We then compare the quartile-wise CMPR_Q_k strategy against this within-quartile EW benchmark.

Table 8 reports the quartile-wise backtest results. Across all four news-frequency quartiles, the CMPR-based strategy delivers higher compound returns than the corresponding within-quartile equal-weighted (EW) benchmark. The outperformance is most pronounced in the high-news-frequency group Q_4: the stratum where attention-intensive "star" firms are most prevalent, with a compound return of 36.9% versus 15.5% for the EW benchmark in Q_4. Notably, the within-quartile EW benchmark also attains its highest compound return in Q_4, suggesting that baseline performance may itself vary depending on news exposure. Importantly, however, even within Q_4 (holding news exposure and firm composition relatively fixed), the CMPR signal still substantially outperforms the within-quartile EW benchmark, indicating that the results are not driven solely by a star firm composition effect.

Economically large gains are also observed in the lower-coverage strata (Q_1: 26.1% vs. 8.14%; Q_2: 24.7% vs. 10.2%), implying that the signal's value is not confined to the most attention-intensive firms. In relative terms, the improvement is largest in Q_1, where CMPR delivers more than three times the within-quartile EW benchmark. Overall, because the comparison is conducted within each news-frequency stratum, systematic differences in firm composition across quartiles, including the higher prevalence of attention-intensive firms in Q_4, are largely held fixed. The remaining within-quartile performance gaps therefore suggest

that the sentiment signal contains incremental information even among firms with similar news exposure, rather than being driven solely by a high-attention (or star-firm composition) effect.

However, full-period portfolio backtests aggregate over both (i) the conditional return response to sentiment-relevant news and (ii) how often the strategy is actually invested (i.e., the share of days with active signals). This aggregation can mute full-sample performance measures in strata with more inactive (no-position) periods, or when eligible signals are spread across many names so that equal-weight positions become mechanically smaller. To isolate the conditional return response and to examine sign asymmetry more directly, we next conduct an event-based conditional return analysis within each quartile.

Table 8: Performance comparison between quartile-wise CMPR_Q_k strategies and within-quartile EW benchmarks

	CR	SD	DD	MDD	ShR	SoR	StR
CMPR_Q_4	36.91 %	27.83 %	16.89 %	32.35 %	132.63 %	218.53 %	114.07 %
EW Q4	15.50 %	18.26 %	11.62 %	32.61 %	84.88 %	133.38 %	47.54 %
CMPR_Q_3	12.31 %	26.31 %	16.48 %	55.50 %	46.79 %	74.72 %	22.19 %
EW Q3	12.18 %	17.59 %	11.25 %	35.30 %	69.20 %	108.25 %	34.49 %
CMPR_Q_2	24.72 %	25.02 %	13.93 %	44.93 %	98.80 %	177.47 %	55.03 %
EW Q2	10.21 %	16.64 %	10.60 %	33.37 %	61.39 %	96.32 %	30.61 %
CMPR_Q_1	26.09 %	20.57 %	10.73 %	25.04 %	126.85 %	243.30 %	104.22 %
EW Q1	8.14 %	15.42 %	9.80 %	32.13 %	52.78 %	83.06 %	25.33 %
TPX Index	10.68 %	17.54 %	11.23 %	31.42 %	60.89 %	95.10 %	33.99 %

4.4.4 Event-Based Conditional Return Test within News-Frequency Quantiles

Since the quartile-wise full-period backtest aggregates returns over the entire sample, it mixes two components: (i) the conditional return effect of sentiment when sentiment-relevant news arrives and (ii) the arrival frequency of such events. This mixture can be particularly important in low-news-frequency strata, where sparse event arrivals and inactive periods may dilute full-sample performance measures even when the conditional effect exists. We therefore complement the backtest with an event-based conditional return test that evaluates conditional returns only on sentiment-event days within each stratum.

Moreover, while our implementable strategy is long-only to maintain a realistic and transparent cost treatment (short selling would require additional assumptions about stock-borrowing fees and short-sale constraints), examining negative-sentiment events remains central from a financial-economics perspective. Market reactions to unfavorable information can differ in magnitude and timing from reactions to favorable information when disclosure may be indirect and limits to arbitrage are present. Accordingly, we analyze negative events alongside positive ones in the event-based framework to evaluate sign asymmetry in price responses.

We use the same fixed news-frequency quartiles $Q_1, \ldots, Q_4$ defined in Section 2.1. For each news headline $N_{i,j}^{m,d,y}$, we compute the headline-level CMPR sentiment score (for $\ell = 4$)

by

$$S_{i,j}^{4,m,d,y} := \sum_{\text{word}_k \in SL_{y-1}^4} \mathbf{1}_{\{\text{word}_k \in N_{i,j}^{m,d,y}\}} P_{y-1,k}^4, \tag{25}$$

where SL_{y-1}^4 and $P_{y-1,k}^4$ are the CMPR-based lexicon and polarity scores introduced in Section 4.4.1.

To align this test with the backtesting information set in Section 4.2.2, we restrict attention to pre-cutoff news-release observations, i.e., timestamped headline occurrences (i, j, m) satisfying $ts_{i,j}^{m,\text{time}} < 14{:}57$. For each such observation (i, j, m), we define the backtest-aligned post-signal return as

$$r_{i,j,m} := r_{t_{i,j}^m+1,i}^{cc}, \tag{26}$$

namely, the close-to-close return used in the backtesting convention for signals formed from news observed before the daily execution cutoff.

For each quantile Q_k, we then define the positive- and negative-sentiment event sets as

$$E_{k,\text{pos}} := \{(i, j, m) \mid i \in Q_k,\ ts_{i,j}^{m,\text{time}} < 14{:}57,\ S_{i,j}^{4,m,d,y} > 0\}, \tag{27}$$

$$E_{k,\text{neg}} := \{(i, j, m) \mid i \in Q_k,\ ts_{i,j}^{m,\text{time}} < 14{:}57,\ S_{i,j}^{4,m,d,y} < 0\}. \tag{28}$$

Using these sets, we compute the conditional mean returns

$$\mu_{k,\text{pos}} := \frac{1}{\#E_{k,\text{pos}}} \sum_{(i,j,m)\in E_{k,\text{pos}}} r_{i,j,m}, \tag{29}$$

$$\mu_{k,\text{neg}} := \frac{1}{\#E_{k,\text{neg}}} \sum_{(i,j,m)\in E_{k,\text{neg}}} r_{i,j,m}. \tag{30}$$

As a within-quantile baseline, we use the average daily return of all firms in the same quantile:

$$\mu_k^{\text{base}} := \frac{1}{T} \sum_d R_{k,d}^{\text{EW}}, \tag{31}$$

where $R_{k,d}^{\text{EW}}$ is the quartile-specific equal-weight daily return defined in Eq. (24), and T is the number of market days in the backtest period.

We conduct one-sided one-sample t-tests within each quantile, using μ_k^{base} as the null mean. Specifically, for positive-sentiment events, we test

$$H_0^{(\text{pos})} : \mu_{k,\text{pos}} \le \mu_k^{\text{base}} \quad \text{vs.} \quad H_1^{(\text{pos})} : \mu_{k,\text{pos}} > \mu_k^{\text{base}}, \tag{32}$$

and for negative-sentiment events, we test

$$H_0^{(\text{neg})} : \mu_{k,\text{neg}} \ge \mu_k^{\text{base}} \quad \text{vs.} \quad H_1^{(\text{neg})} : \mu_{k,\text{neg}} < \mu_k^{\text{base}}, \tag{33}$$

where H_0 denotes the null hypothesis and H_1 denotes the alternative hypothesis.

In addition to the within-quartile tests (Q_1–Q_4), we also report a pooled version of the event-based test using the full universe, denoted by $Q_0 := \bigcup_{k=1}^4 Q_k$. The pooled test evaluates whether the sign effect of CMPR sentiment is present on average across all firms, providing a summary benchmark that is not conditioned on news-frequency strata.

Table 9: Event-based conditional mean returns and one-sided one-sample t-tests within news-frequency quartiles

Quantile	pos ($S_4^{i,j} > 0$) Mean	t-stat	p-value	neg ($S_4^{i,j} < 0$) Mean	t-stat	p-value	Baseline μ_k^{base}
Q_1	0.4363%	3.7610	8.9687×10^{-5}	-0.2544%	-2.4381	7.5541×10^{-3}	0.0375%
Q_2	0.3327%	3.9025	4.9570×10^{-5}	-0.0794%	-1.4061	8.0013×10^{-2}	0.0456%
Q_3	0.2284%	3.0764	1.0595×10^{-3}	-0.2263%	-3.4340	3.0508×10^{-4}	0.0532%
Q_4	0.2585%	5.2400	8.3523×10^{-8}	-0.1141%	-4.1887	1.4308×10^{-5}	0.0655%
Q_0	0.2804%	8.3042	5.6726×10^{-17}	-0.1435%	-5.7142	5.7275×10^{-9}	0.0494%

Table 9 reports conditional mean returns, t-statistics, and one-sided p-values for positive- and negative-sentiment events within each quartile (Q_1–Q_4). Positive-sentiment events are statistically significant at the 1% level in all four quartiles. Negative-sentiment events are also statistically significant at the 1% level in three of the four quartiles (Q_1, Q_3, and Q_4), but weaker and marginally significant at the 10% level in Q_2. For reference, we also report a pooled benchmark using the full firm universe (Q_0), which shows that both positive and negative sentiment events move in the predicted directions on average, with stronger statistical significance for positive-sentiment events.

These event-based results complement the quartile-wise backtest in Section 4.4.3 by isolating conditional return effects on sentiment-event days within each stratum. Notably, even in the upper-middle news-coverage group Q_3, Table 9 shows statistically significant event-day effects, despite muted portfolio-level outperformance in the full-period backtest (Table 8).

Taken together, the within-quartile event-based evidence mitigates the concern that our findings are driven solely by a small number of high-attention ("star") firms with exceptionally high news counts, since statistically significant sentiment-event effects are observed across multiple strata, including low-news-coverage strata (Q_1 and Q_2), rather than being confined to a narrow subset of firms.

An additional implication of Table 9 is a clear positive–negative asymmetry in the event-based effects. In the pooled benchmark (Q_0), the positive-sentiment effect is statistically far more pronounced than the negative-sentiment effect, and this ordering also holds in three of the four news-coverage quartiles (except Q_3). These results are consistent with the hypothesis that favorable information is more explicitly encoded in Japanese financial news headlines, whereas unfavorable information can be harder to capture systematically with word-based measures. We interpret this evidence as suggestive, rather than causal, of how language environment and disclosure style may affect the transmission of textual information into stock prices.

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

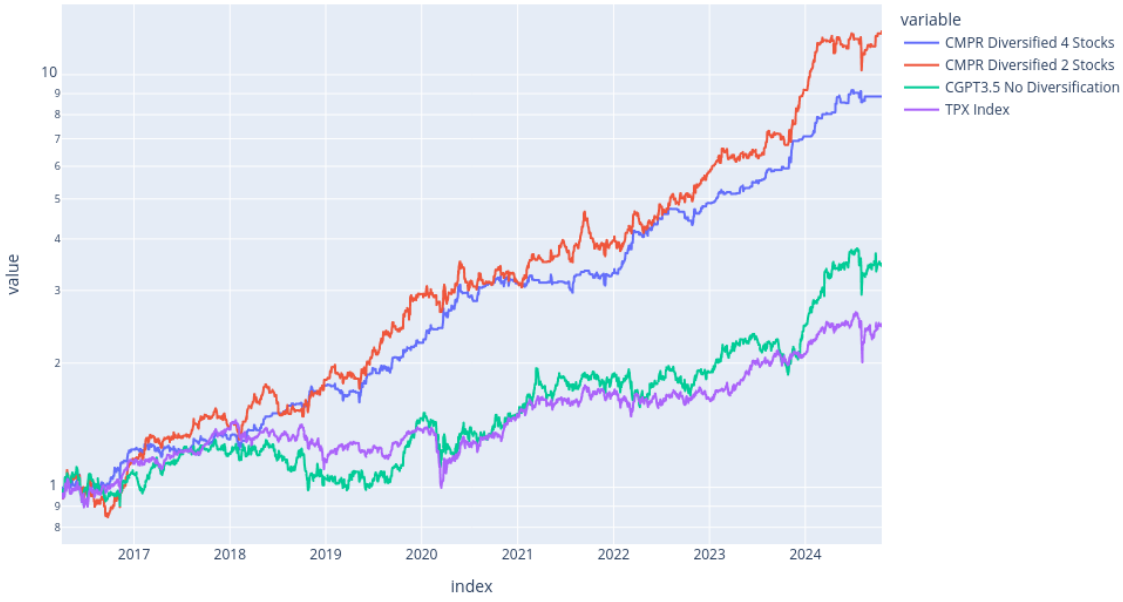
	CR	SD	DD	MDD	ShR	SoR	StR
CMPR Diversified 4 Stocks	27.53 %	14.52 %	8.29 %	15.55 %	189.60 %	332.20 %	177.08 %
CMPR Diversified 2 Stocks	32.70 %	21.06 %	12.26 %	23.03 %	155.31 %	266.64 %	141.97 %
CMPR Diversified 6 Stocks	13.70 %	9.82 %	5.51 %	8.19 %	139.49 %	248.58 %	167.37 %
CMPR No Diversification	27.35 %	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 be realized in practice, this subsection investigates the performance of strategies by using TOPIX futures as a trading instrument combined with our original lexicons (dictionaries).

We extend the backtesting period for the futures strategy to April 14, 2025, to include the severe market crash in early April 2025, testing the robustness of the strategy under extremely volatile market conditions.

Let us note that since trading hours in JPX are extended on November 5, 2024, we adjust our return calculations accordingly. Specifically, equity markets and TOPIX futures close at 15:00 and 15:15 until November 4, 2024, while it is extended to 15:30 and 15:45 after November 5, 2024, respectively. In the following, whenever two times are shown in the format hh:mm (hh':mm'), the time outside the parentheses applies to dates before 5 November 2024, while the time in parentheses applies to 5 November 2024 and afterwards.

To accommodate the extended trading hours, we adjust the return calculation to construct our lexicons(dictionaries): for news released at or after 14:57 (15:27), we apply the following trading day's open-to-close intraday return; for news published before 14:57 (15:27), we use the standard close-to-close daily return.

4.6.1 Daily Calculation of the market sentiment score

Since the closing time for TOPIX futures is 15:15(15:45), we collect news released until 15:12(15:42) from 0:00 for the TOPIX futures strategy.

Namely, at 15:12(15:42) 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(15:42) 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(15:42) until 15:15(15:45) 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(15:45) 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:

$$\begin{aligned} S_{i,j}^{\ell,m,d,y} &:= \sum_{\text{word}_k \in SL_{y-1}^{\ell}} 1_{\text{word}_k \in N_{i,j}^{m,d,y}} P_{y-1,k}^{\ell}, \text{ for } \ell = 1,2,3,4, \\ S_{i,j}^{0,m,d,y} &= cgpt(N_{i,j}^{m,d,y}), \quad \text{for } \ell = 0, \end{aligned} \tag{34}$$

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

$$S_i^{\ell,d,y} := \sum_{j=1}^{J_i^{d,y}} \sum_{m=1}^{M_{i,j}^{d,y}} S_{i,j}^{\ell,m,d,y}. \tag{35}$$

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:

$$\begin{aligned} S_{\text{topx500}}^{\ell,d,y,1} &:= \sum_{i \in \text{TOPIX500}^{d,y}} w_i^{d,y} S_i^{\ell,d,y}, \\ S_{\text{topx500}}^{\ell,d,y,2} &:= \sum_{i \in \text{TOPIX500}^{d,y}} S_i^{\ell,d,y}, \end{aligned} \tag{36}$$

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 $S^{\ell,d,y,k}_{\text{topx500}} > 0$ $(k = 1, 2)$, we take a long position of TOPIX futures at 15:15. Also, if $S^{\ell,d,y,k}_{\text{topx500}} \leq 0$, 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 (36).

- **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_SR_mcap 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 11 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 11: Performance Metrics (Sorted by Sharpe Ratio)

	CR	SD	DD	MDD	ShR	SoR	StR
topix100_MPR_mcap	17.81 %	12.61 %	7.37 %	15.84 %	141.18 %	241.49 %	112.44 %
topix30_MPR_mcap	13.05 %	10.63 %	6.25 %	11.31 %	122.73 %	208.82 %	115.39 %
topix100_MPR	14.18 %	12.04 %	7.10 %	10.32 %	117.72 %	199.61 %	137.42 %
topix100_MPM_mcap	14.95 %	12.90 %	7.77 %	14.02 %	115.92 %	192.40 %	106.63 %
topix30_MPR	11.47 %	10.73 %	6.43 %	13.83 %	106.92 %	178.37 %	82.95 %
topix100_MPM	13.08 %	12.34 %	7.45 %	12.88 %	106.01 %	175.61 %	101.58 %
topix30_MPM	10.54 %	11.01 %	6.83 %	13.28 %	95.75 %	154.25 %	79.34 %
topix30_MPM_mcap	10.31 %	10.91 %	6.77 %	10.30 %	94.52 %	152.33 %	100.11 %
topix500_MPR_mcap	12.49 %	13.65 %	8.23 %	19.92 %	91.49 %	151.79 %	62.71 %
topix30_SR_mcap	13.56 %	15.13 %	9.45 %	16.94 %	89.62 %	143.45 %	80.06 %
topix500_MPM_mcap	11.92 %	13.81 %	8.36 %	18.22 %	86.32 %	142.54 %	65.42 %
topix30_SM_mcap	12.72 %	15.56 %	9.94 %	18.09 %	81.74 %	127.99 %	70.30 %
topix30_SR	11.90 %	15.06 %	9.46 %	17.36 %	79.05 %	125.89 %	68.57 %
topix100_SR_mcap	12.47 %	16.31 %	10.21 %	24.24 %	76.47 %	122.20 %	51.46 %
topix30_SM	10.85 %	15.25 %	9.65 %	22.50 %	71.16 %	112.50 %	48.23 %
topix100_SM_mcap	11.26 %	16.86 %	10.72 %	24.71 %	66.81 %	105.06 %	45.58 %
topix100_SR	11.20 %	17.19 %	11.29 %	26.34 %	65.16 %	99.24 %	42.53 %
topix100_SM	11.16 %	17.32 %	10.91 %	24.28 %	64.42 %	102.28 %	45.95 %
topix500_CGPT_3.5	2.48 %	3.94 %	2.63 %	8.63 %	62.87 %	94.05 %	28.72 %
topix500_SR_mcap	10.72 %	17.59 %	11.59 %	24.94 %	60.92 %	92.53 %	42.98 %
topix500_CGPT_3.5_mcap	3.65 %	6.10 %	3.85 %	13.35 %	59.83 %	94.83 %	27.35 %
topix100_CGPT_3.5_mcap	3.47 %	6.28 %	3.92 %	12.51 %	55.33 %	88.58 %	27.75 %
topix500_MPM	7.30 %	13.57 %	8.62 %	15.66 %	53.78 %	84.63 %	46.61 %
TPX Index(future)	10.11 %	19.55 %	12.57 %	32.64 %	51.73 %	80.44 %	30.98 %
topix500_SM_mcap	9.09 %	18.09 %	11.80 %	26.32 %	50.28 %	77.06 %	34.55 %
topix500_SM	8.96 %	18.77 %	12.31 %	30.96 %	47.72 %	72.75 %	28.93 %
topix500_SR	7.97 %	18.30 %	12.13 %	30.79 %	43.57 %	65.72 %	25.89 %
topix500_MPR	5.49 %	13.12 %	8.61 %	19.01 %	41.88 %	63.81 %	28.89 %
topix30_CGPT_3.5	2.42 %	5.95 %	3.85 %	14.80 %	40.61 %	62.81 %	16.33 %
topix500_CGPT_4o_mini	7.39 %	18.41 %	12.05 %	25.47 %	40.16 %	61.36 %	29.03 %
topix30_CGPT_4o_mini	6.19 %	15.57 %	9.86 %	28.96 %	39.76 %	62.78 %	21.37 %
topix30_CGPT_3.5_mcap	2.65 %	6.72 %	4.49 %	16.69 %	39.49 %	59.06 %	15.90 %
topix30_CGPT_4o_mini_mcap	5.54 %	15.22 %	9.87 %	29.91 %	36.39 %	56.13 %	18.52 %
topix100_CGPT_4o_mini	5.78 %	16.43 %	10.64 %	26.66 %	35.19 %	54.33 %	21.69 %
topix100_CGPT_3.5	1.97 %	5.71 %	3.74 %	17.42 %	34.51 %	52.69 %	11.32 %
topix500_CGPT_4o_mini_mcap	5.58 %	16.20 %	10.51 %	27.53 %	34.47 %	53.14 %	20.28 %
topix100_CGPT_4o_mini_mcap	4.87 %	15.87 %	10.38 %	31.00 %	30.69 %	46.94 %	15.71 %

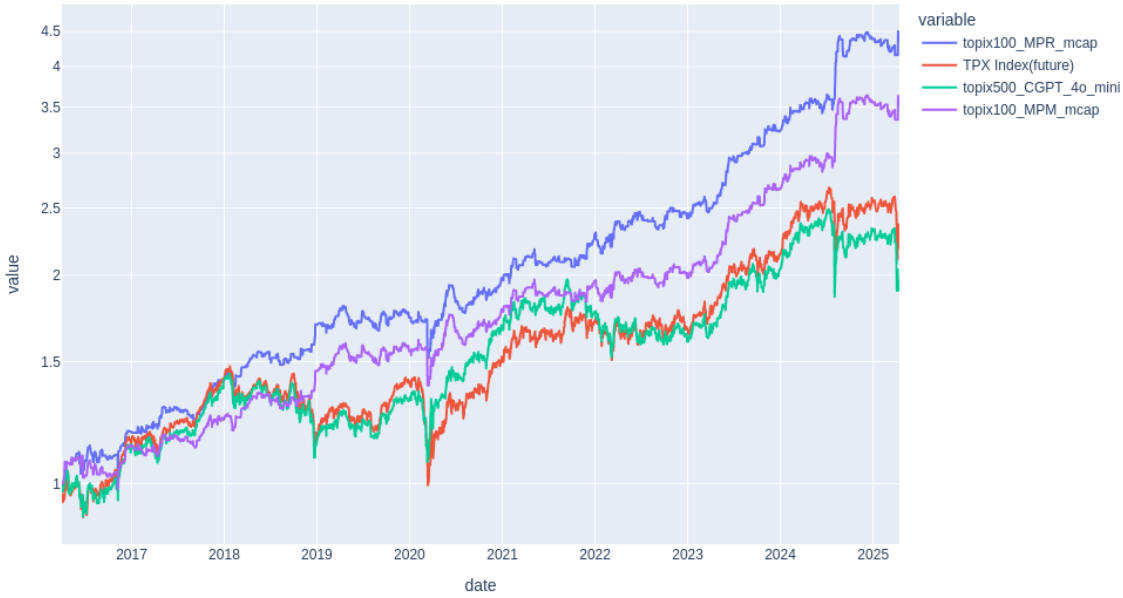
Firstly, it shows that in terms of Sharpe Ratio, most of our strategies (20 out of 24) outperform the benchmark TOPIX futures whose Sharpe Ratio=51.73%, while CGPT-based strategies do only 3 out of 12. Moreover, the best Sharpe Ratio created by TOPIX100_MPR_mcap

among our strategies is more than 140%, while that by TOPIX500_CGPT_3.5 among CGPT-based strategies is just around 63%.

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, 17.81%) and second best (TOPIX100_MPM_mcap, 14.95%) compound returns, as well as those of a CGPT-based strategy with the best compound returns (TOPIX500_CGPT_4o_mini, 7.39%) and the benchmark TOPIX futures(10.11%).

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 proposes a novel approach to sentiment analysis to enhance investment decision-making in the Japanese stock market. We first extract a corpus-derived set of finance-specific keywords from Japanese news headlines and construct sentiment lexicons by assigning polarity scores that are directly linked to market returns. In particular, by incorporating market excess

returns and regression-based estimation into the lexicon construction, our approach provides a practical way to translate textual information into interpretable sentiment signals.

Extensive empirical analyses show that the proposed strategies achieve much higher risk adjusted returns than existing NLP baselines (including MeCab-based approaches and a bag-of-words benchmark) as well as ChatGPT-based sentiment signals. Importantly, time-series regressions on the Fama–French three-factor model indicate that the returns of our proposed strategies are not fully explained by standard risk-factor exposures, because the intercepts, i.e., the factor-adjusted alphas remain positive and statistically significant. In contrast, none of the benchmark methods deliver comparably significant alphas under the same evaluation protocol.

Moreover, our event-based conditional tests provide evidence of sign asymmetry in market reactions: conditional return effects tend to be more robust for positive-sentiment news than for negative-sentiment news, suggesting that the incorporation of textual information into prices differs between favorable and unfavorable disclosures. Overall, these findings highlight that domain-specific expertise combined with market-driven sentiment analysis remains valuable for generating economically meaningful and factor-robust performance in the Japanese stock market.

Acknowledgement

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

Appendix A Performance measure

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

- Compound return (CR):

$$CR \equiv \left\{ \prod_{t=1}^{T} (1 + R_t) \right\}^{1/T} - 1. \tag{37}$$

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

- Standard deviation (SD):

$$SD \equiv \left\{ \frac{1}{T} \sum_{t=1}^{T} (R_t - \bar{R})^2 \right\}^{1/2}, \ \bar{R} \equiv \frac{1}{T} \sum_{t=1}^{T} R_t. \tag{38}$$

  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):

$$DD \equiv \left\{ \frac{1}{T} \sum_{t=1}^{T} \min(0, R_t)^2 \right\}^{1/2}. \tag{39}$$

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

- Maximum drawdown (MDD):

$$MDD \equiv \max_{1 \leq t \leq T} \frac{M_t - V_t}{M_t}, \quad M_t \equiv \max_{0 \leq s \leq t} V_s. \tag{40}$$

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.

In short, 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):

$$ShR \equiv (\bar{R} - r_f)/SD, \tag{41}$$

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):

$$SoR \equiv (\bar{R} - r_f)/DD. \tag{42}$$

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

- Sterling Ratio (StR):

$$StR \equiv (\bar{R} - r_f)/MDD. \tag{43}$$

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

Appendix B Construction of TOPIX500-Based Fama–French-Type Three-Factor Controls

This appendix describes the construction of the Fama–French three-factor control returns used in the factor-model analysis in Section 4.4.2. We construct factor portfolios on the TOPIX500 universe with quarterly rebalancing and point-in-time sorting variables (market capitalization and PBR), and then compute daily MKT, SMB, and HML returns.

Universe and rebalancing schedule. The factor portfolios are constructed from the TOPIX500 constituent universe. Portfolio membership is rebalanced on the last trading day of each quarter (i.e., the last trading day of March, June, September, and December), and the resulting factor portfolio composition is held until the next quarterly rebalancing date.

Sorting variables. At each rebalancing date, we obtain firm-level market capitalization and price-to-book ratio (PBR) for all eligible TOPIX500 constituents. To align the value factor with the standard Fama–French definition, we define the book-to-market measure as the inverse of PBR:

$$\mathrm{BM}_{i,t} = \frac{1}{\mathrm{PBR}_{i,t}}. \tag{44}$$

Thus, firms with lower PBR correspond to higher book-to-market firms (value firms), and firms with higher PBR correspond to lower book-to-market firms (growth firms).

Portfolio formation. At each quarterly rebalancing date, firms are first split into two groups by market capitalization (Small and Big) using the cross-sectional median market capitalization within the TOPIX500 universe. Independently, firms are sorted into three groups by book-to-market using the 30th and 70th percentiles: Low (L), Medium (M), and High (H). Combining these two sorts yields six portfolios:

$$S/L,\ S/M,\ S/H,\ B/L,\ B/M,\ B/H.$$

Within each portfolio, daily returns are computed as equal-weighted returns of constituent stocks and held fixed until the next rebalancing date.

Factor return definitions. Using the six portfolio returns, the size factor (*Small Minus Big*, SMB) and the value factor (*High Minus Low*, HML) are constructed as:

$$\mathrm{SMB}_t = \frac{1}{3}\left(R_{S/L,t} + R_{S/M,t} + R_{S/H,t}\right) - \frac{1}{3}\left(R_{B/L,t} + R_{B/M,t} + R_{B/H,t}\right), \tag{45}$$

$$\mathrm{HML}_t = \frac{1}{2}\left(R_{S/H,t} + R_{B/H,t}\right) - \frac{1}{2}\left(R_{S/L,t} + R_{B/L,t}\right), \tag{46}$$

where $R_{\cdot,t}$ denotes the daily portfolio return at date t.

Timing and data availability. All sorting variables (market capitalization and PBR) are measured at each quarterly rebalancing date using point-in-time information available as of that date. Factor portfolios are then held over the subsequent trading days until the next rebalancing.

Regression. We estimate the following daily time-series regression for each strategy portfolio p:

$$R_{p,t} - \mathrm{RF}_t = \alpha_p + \beta_{p,\mathrm{MKT}}\left(\mathrm{MKT}_t - \mathrm{RF}_t\right) + \beta_{p,\mathrm{SMB}}\,\mathrm{SMB}_t + \beta_{p,\mathrm{HML}}\,\mathrm{HML}_t + \varepsilon_{p,t}, \tag{47}$$

where MKT_t denotes the daily TOPIX500 index return and RF_t is the daily risk-free rate. We proxy RF_t by the overnight call rate (TONAR), converted to a daily rate. We estimate the coefficients by ordinary least squares using daily observations, and interpret α_p as the factor-adjusted abnormal return.

Appendix C Rationale for the Dictionary Frequency Threshold

In this appendix, we provide an additional justification for the frequency threshold used in Eq. (3), where we define the candidate word/phrase set by requiring that each element appears more than 200 times in non-financial news headlines.

To generalize Eq. (3), for a threshold $t > 0$, let us define

$$A_{>t} := \{tss_\ell \in TS \mid FR_\ell > t\}, \tag{48}$$

where TS is the set of tokenized strings obtained from the headline corpus and FR_ℓ is the empirical frequency of token tss_ℓ in the full sample (see Section 3.1). Note that the set A in Eq. (3) corresponds to $A_{>200}$.

We first evaluate how well the high-frequency vocabulary set $A_{>t}$ covers economically relevant headlines in $\mathcal{N}_{\text{nonF}}$. For this purpose, we define the coverage rate

$$\text{Coverage}(t) := \frac{\#\{h \in \mathcal{N}_{\text{nonF}} \mid \exists w \in A_{>t} \text{ s.t. } w \preceq h\}}{\#\{h \in \mathcal{N}_{\text{nonF}}\}}, \tag{49}$$

where $w \preceq h$ means that the word/phrase w appears as a substring of headline h.

Table 12 reports Coverage(t), the corresponding vocabulary size $\#A_{>t}$, and the marginal increase in coverage when the threshold is lowered. As expected, the coverage rate increases as the threshold decreases. However, the incremental gain becomes small once the threshold is lowered below $t = 200$, while the vocabulary size continues to expand substantially. In particular, Coverage(200) = 96.2%, indicating that the threshold $t = 200$ already captures the vast majority of non-financial (material) headlines.

Table 12: Coverage of $\mathcal{N}_{\text{nonF}}$ headlines by frequency threshold

Frequency threshold t	$\#A_{>t}$	Coverage(t)	Marginal increase in coverage
> 500	72	87.9%	–
> 400	93	91.5%	3.63%
> 300	110	93.6%	2.09%
> 200	161	96.2%	2.63%
> 100	319	98.7%	2.50%
> 50	580	99.4%	0.69%

Next, we examine whether the threshold provides sufficient statistical precision for estimating word-level polarity. In our framework, word polarity is constructed from the empirical mean of post-news market excess returns. Hence, for a word with N associated observations and variance σ^2, the standard error of the sample-mean estimator is approximated by

$$\text{SE} \approx \frac{\sigma}{\sqrt{N}}. \tag{50}$$

Therefore, higher-frequency words yield more precise polarity estimates.

To align this approximation with the data used in polarity estimation, we estimate σ using the pooled standard deviation of post-news market excess returns across all news events. The

resulting daily standard deviation is approximately 1.99%. Thus, for a word appearing 200 times, the standard error is approximately

$$\frac{1.99\%}{\sqrt{200}} \approx 0.141\% = 14.1 \text{ bps}. \tag{51}$$

In Section 3.2, we adopt 15 bps as the threshold for economically meaningful polarity so that the expected return signal exceeds typical round-trip transaction costs in practical implementation. From this viewpoint, the threshold $t = 200$ is also justified because the estimation error (approximately 14.1 bps) is below the economically meaningful decision threshold (15 bps).

Overall, the threshold $t = 200$ satisfies two conditions simultaneously:

(i) it achieves high headline coverage (Coverage(200) = 96.2%), and

(ii) it provides sufficient statistical precision for word-level polarity estimation relative to the 15 bps economic threshold.

For these reasons, we adopt the frequency threshold $t = 200$ in Eq. (3).

Appendix D MeCab-based Benchmark: Threshold Choice and Sensitivity Analysis

As for frequency threshold of the MeCab-based benchmark, we adopt a more conservative threshold than in our proposed dictionary construction. This choice is deliberate and is motivated by a different design objective.

The proposed method extracts recurring expressions directly from headline strings, whereas the MeCab-based benchmark relies on morphological segmentation. In Japanese financial headlines, morphological segmentation can generate fragmented tokens (e.g., partial proper nouns, split compound expressions, or semantically weak fragments), especially when company-specific or finance-specific expressions are segmented imperfectly. When the frequency threshold is set too low, such low-frequency fragments increase rapidly and make polarity estimation unstable.

To mitigate this issue, we use a stricter threshold (1,000 occurrences) in the MeCab-based benchmark, which suppresses unstable low-frequency fragments and retains only frequently observed tokens. At the same time, the resulting vocabulary remains sufficiently large to cover major recurring expressions in the headline corpus.

To confirm that the empirical conclusion is not driven by this particular threshold choice, we additionally conduct a sensitivity analysis over a wide range of thresholds for the MeCab-based MMPR strategy:

$$t \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000\}.$$

Table 13 reports the corresponding performance metrics.

The main conclusion is robust. While certain thresholds improve some measures relative to the TPX index, the MeCab-based results vary substantially across thresholds and do not exhibit stable performance. Moreover, even under comparatively favorable threshold choices, the MeCab-based benchmark generally remains less reliable and inferior to the proposed methods. Overall, these findings suggest that the primary limitation of the MeCab-based

benchmark is structural—stemming from segmentation quality and fragmentation-driven noise—rather than a consequence of an arbitrary single-threshold selection.

Table 13: Sensitivity of the MeCab-based MMPR strategy to the frequency threshold

Method	CR	SD	DD	MDD	ShR	SoR	StR
MMPR2500	20.38%	23.60%	13.16%	39.48%	86.33%	154.85%	51.61%
MMPR3500	17.71%	25.94%	15.71%	62.34%	68.28%	112.75%	28.41%
MMPR2000	15.05%	22.56%	13.51%	41.83%	66.70%	111.37%	35.98%
MMPR4000	15.78%	24.17%	14.92%	52.13%	65.26%	105.75%	30.26%
MMPR3000	15.02%	24.02%	13.72%	52.33%	62.54%	109.48%	28.71%
TPX Index	10.68 %	17.54 %	11.23 %	31.42 %	60.89 %	95.10 %	33.99 %
MMPR200	12.79%	26.43%	15.39%	43.93%	48.39%	83.11%	29.11%
MMPR1500	9.69%	21.76%	13.12%	46.50%	44.54%	73.87%	20.84%
MMPR1000	7.11 %	21.62 %	13.70 %	47.62 %	32.90 %	51.93 %	14.94 %
MMPR500	0.86%	21.99%	14.04%	47.30%	3.89%	6.10%	1.81%

References

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.

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

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

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.

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.

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

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

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).

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.

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.

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).

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.

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

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.

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.

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

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

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.

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.

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

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.

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

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

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

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

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

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

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.

Nishimura, K. G., Sato, S., & Takahashi, A. (2019). Term structure models during the global financial crisis: A parsimonious text mining approach. Asia-Pacific Financial Markets, 26(3), 297-337.

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.

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

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10 - Ks. The Journal of finance, 66(1), 35-65.

Li, F. (2010). The information content of forward - looking statements in corporate filings—A naïve Bayesian machine learning approach. Journal of accounting research, 48(5), 1049-1102.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. Journal of accounting research, 54(4), 1187-1230.

Ito, T., Sakaji, H., Tsubouchi, K., Izumi, K., & Yamashita, T. (2018, June). Text-visualizing neural network model: understanding online financial textual data. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 247-259). Cham: Springer International Publishing.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56.

Chen, Y., Nakagawa, K., Kimura, Y., & Inoue, K. (2025). Are managerial cognitive biases priced in Japan? Evidence from cross-sectional portfolio returns. Finance Research Letters, 107940.

Cho, M., Hah, Y. D., & Kim, O. (2011). Optimistic bias in management forecasts by Japanese firms to avoid forecasting losses. The International Journal of Accounting, 46(1), 79-101.