

# ***How Does Emotion Impact the Market: Relevance Analysis Between News Sentiment and Industry Index***

**Yilinsaihan Bai**

*University of Illinois Urbana-Champaign, Champaign, USA  
yb16@illinois.edu*

**Abstract.** Whether news events have impacts on the equity market is a topic of widespread concern. The efficient market hypothesis holds that market prices can reflect public information in a timely manner. This article investigates the relationship between news events and the equity market, analyzing the role of sentiment across 11 industry sectors in the US stock market. This study uses GDELT to retrieve event-driven news through a large, weighted query bank and compute title-based sentiment scores, then map each article to sectors via word-embedding similarities. Sentiment is further quantified using a headline polarity measure. Fixed-effects regressions with HAC standard errors are used to test the predictive power of global and sector-specific sentiment on total and excess returns, while incorporating lagged sentiment to avoid look-ahead bias. The study further examines market regimes (down markets) as candidate settings in the regression framework. The results indicate that headline sentiment has limited explanatory power in the full sample once standard market controls are included, but sector-idiosyncratic sentiment becomes more informative in down-market regimes, consistent with sentiment-driven sector dispersion during periods of heightened risk aversion. These findings can help investors conducting fundamental analysis better interpret news and assess when sentiment is most likely to matter for sector-level return dynamics.

**Keywords:** Sentiment Analysis, News Event Impact, Industry Index, Relevance Analysis

## **1. Introduction**

The stock market is a fundamental component of the financial market. Timing and stock selection have always been the focus of investors. According to the Efficient Market Hypothesis, stock prices in a semi-strong efficient market reflect all public information. As the media becomes more instant and regulated, modern financial investors and researchers are supposed to find more correlations between stock performance and financial news. Considering that investors have their own judgment of each piece of news, a piece of news might add confidence to investors or discourage them from investing in a particular asset. Therefore, it is reasonable to assume that the news affects the stock market through its sentiment. Although the personal judgment of specific news is hard to predict due to complex factors, such as a person's optimism, trading experience, and risk tolerance, one can obtain a sentiment index for each piece of news using a data-mining technique on a large dataset.

A specific stock index depends on multiple factors, such as a firm's financial reports, government policies, and the overall performance of the financial market. Therefore, a specific model to analyze a single stock movement is neither achievable nor reliable. In such circumstances, market or industry indices, which summarize the performance of stocks grouped by specific sectors and help investors

reduce the volatility of single-stock movements, become more attractive. A practical and reliable study of industry index and current financial news would provide crucial information for investors.

There is extensive literature showing that the sentiment expressed in news articles can affect stock prices and predict returns. Along with the rapid development of Web 2.0 technologies, the volume of information has been growing geometrically. News, forums, and blog articles enable the connection, distribution, and sharing of information. Therefore, the internet's function is not limited to information publishing; it also integrates data and information from the interpersonal network, thereby enhancing the impact and transmission of online information. Given the massive development and broad adoption of big data, it is certain that one can excavate valuable information from unstructured text data. Since data mining technology enables the extraction of information from large-scale text data, more and more researchers have taken advantage of it. Gruhl and Daniel predicted sales of books according to blog and web contents [1]; Asur and Huberman found the correlation between box office results and Twitter posts [2]; Mishane predicted box office by studying viewer sentiments from blog contents [3]. On the other hand, big data is significantly applied in stock prediction. Schumaker and Chen found a correlation between news and stock prices by analyzing financial news [4]; Wysocki examined the relationship between the stock forum and stock prices, concluding that overnight forum content could be used to predict stock price movements [5].

Meanwhile, researchers apply sentiment analysis from social media data to stock market prediction. Since investors rely more on online information for investment decisions, they participate in discussions and exchanges in online forums. Based on Liu's psychological studies, the expression of information has a considerable impact on human beings' decision-making [6]. Therefore, it is reasonable to assume that the sentiment contained in big data can affect investors' decision-making. Das collected Yahoo stock forum data, applied several classification algorithms to determine the sentiment expressed in a forum post, and concluded that there was a correlation between sentiment and stock prices [7]. Nguyen and Shirai recognized the importance of the text's theme and applied a support vector machine to predict stock price movements [8]. Maragoudakis handpicked the sentiment-indicating word bank and assigned weights to these words (-2, -1, 0, 1, 2). This methodology demonstrated moderate accuracy [9]. Seghal and Song constructed a model based on Web data analysis, divided sentiment into five categories, and introduced Trust Value as the variable measuring trustworthiness of text sources [10]. Alsing and Banceci applied text analysis using a sentiment dictionary and explained why sentiment analysis based on machine learning required manual marking when building a training set [11].

Based on the review of available literature, current research agrees that big data, such as news, blogs, and forum data, contain a considerable amount of valuable information, especially in measuring the impact of sentiment on financial market movement. Therefore, a sentiment analysis based on text big data has a sufficient theoretical basis. On the other hand, previous studies focus on movements of individual stocks. In contrast, this study focuses on movements of industry indices to mitigate volatility and errors arising from movements of individual stocks.

This paper presents a framework that translates news into daily sentiment indicators for both global news sentiment and sector-specific sentiment, and examines whether these indicators can predict sector ETF returns and excess returns. News articles are gathered from the GDELT DOC news database using targeted keyword searches, covering the period from January 1, 2024, to September 30, 2025. Sentiment is assessed using TextBlob on article titles, while GloVe word embeddings assign each article to relevant sectors. To analyze the data, this study uses panel regressions with sector fixed effects and, to handle heteroscedasticity, estimates the model with HAC standard errors using statsmodels. The results indicate that lagged global sentiment generally lacks predictive power, whereas sector-specific idiosyncratic sentiment becomes significantly negative during market downturns. Key contributions of this study include: (i) employing an extensive weighted query bank

to extract event-driven news, (ii) utilizing embedding-based soft sector classification to convert unstructured text into daily sector panels, and (iii) applying market-regime-specific return regressions to pinpoint when the sentiment-return relationship becomes evident.

The rest of this paper is structured as follows. Section 2 details the measurement of sentiment and the development of a sentiment index utilizing financial news data. Section 3 presents the methodology used to analyze the relationship between sentiment and industry-sector index returns. In Section 4, the data and empirical findings across various sectors are examined. Finally, Section 5 concludes the work.

## 2. Methodology for sentiment analysis

This work aims to mine stock-related emotional factors from large amounts of unstructured text data, such as news and reviews, and to compute the correlation between text sentiment and the movement of industry indexes. Because news sentiment is conveyed in natural language, the purpose of text analysis in this work is to process it and obtain quantifiable emotional indicators that can represent relevant public opinion information.

This analysis utilizes news headlines to assess sentiment and determine a corresponding sentiment score. The sentiment score is computed from financial news article titles by employing the TextBlob library in Python. TextBlob is a rule-based natural language processing tool that performs sentiment analysis using a pre-trained lexicon called PatternAnalyzer. Each term or phrase within this lexicon is assigned both a sentiment polarity and a subjectivity score [12].

### 2.1. Sentiment score

For each news item, two indicators are computed related to the sentiment of the news headlines:

**Polarity:** Indicates whether sentiment is negative or positive, with values from  $-1$  (completely negative) to  $+1$  (completely positive).

**Subjectivity:** Shows how much of the content is based on opinion, ranging from  $0$  (fully objective) to  $1$  (entirely subjective).

Similarly, the polarity and subjectivity of a sentence can be calculated by the following equations:

$$\text{Polarity}(\text{sen}) = (-1)^{N_{\text{neg}}} \sum_{\text{word}_j} \text{Freq}(\text{word}_j) \times \text{Polarity}(\text{word}_j) \times \text{Intensity}(\text{word}_j)$$

$$\text{Subjectivity}(\text{sen}) = \sum_{\text{word}_j} \text{Freq}(\text{word}_j) \times \text{Subjectivity}(\text{word}_j) \times \text{Intensity}(\text{word}_j)$$

where  $\text{Freq}(\text{word}_j)$  represents the frequency of  $\text{word}_j$  and  $N_{\text{neg}}$  represents the number of negations in the sentence. This study will focus on the polarity of the news title and denote  $\text{pol}_i \in [-1, 1]$  as the sentiment polarity of the title of the news  $i$ .

### 2.2. Text representation: turning words into vectors

This section considers how to represent sector-related information in mathematical form, enabling automatic assignment of news to different sectors. A tokenization function  $\text{Tok}(\cdot)$  is defined for the news  $i$ ,

$$\text{Tok}(\text{Title}_i) = [w_{i,1}, w_{i,2}, \dots, w_{i,m_i}],$$

where  $m_i$  is the number of word tokens in the title. Then the title vector is defined as the mean of the embeddings of in-vocabulary tokens:

$$v_i = \frac{1}{m_i} \sum_{j=1}^{m_i} e(w_{i,j}),$$

where the function  $e(\cdot)$  is a word embedding vector function that maps the word to a high-dimensional vector. In the empirical analysis, the glove-wiki-gigaword-100 embedding vector function from the Python package genism is used.

For each industry sector, a sector prototype vector is created using specific anchor phrases. This involves selecting 20 phrases that are most strongly associated with each sector, for instance, “semiconductor” and “GPU” for the Technology sector, or “crude oil” and “OPEC” for the Energy sector. These selected phrases form the anchor phrase list, and one can denote the anchor phrase list for the sector  $s$  as

$$\{(a_{s,1}, \lambda_{s,1}), \dots, (a_{s,J_s}, \lambda_{s,J_s})\},$$

where  $a_{s,j}$  is a short phrase string, and  $\lambda_{s,j} > 0$  is its anchor weights, representing the importance of the phrase, and  $J_s$  is the number of anchor phrases for this sector. Here, each anchor phrase  $a_{s,j}$  is firstly mapped to a vector by the same embedding rule described above, and denote the sector prototype vector as

$$V_s = \frac{\sum_{j=1}^{J_s} \lambda_{s,j} e(a_{s,j})}{\sum_{j=1}^{J_s} \lambda_{s,j}},$$

and  $V_s$  represents the sector’s importance-weighted average meaning within the embedding space.

### 2.3. Mapping articles to sectors: soft assignment

Next, this section illustrates how to map each article to the vectors using soft assignment. Define the cosine similarity for any two non-zero vectors  $x, y$  as

$$\cos(x, y) = \frac{x^T y}{|x| \cdot |y|},$$

and for each article  $i$  and sector  $s$ , compute:

$$\text{sim}_{i,s} = \cos(v_i, V_s).$$

This will yield a value in  $[-1, 1]$ . Larger values mean the article title is more semantically similar to that sector. Then the softmax probabilities with temperature are applied, which converts similarities into probabilities using a softmax transformation:

$$p_{i,s} = \frac{\exp(\tau \cdot \text{sim}_{i,s})}{\sum_{s' \in S} \exp(\tau \cdot \text{sim}_{i,s'})},$$

where  $p_{i,s}$  is the probability weight to assign article  $i$  to sector  $s$ ,  $\tau > 0$  is the temperature parameter controlling the sharpness, and  $S$  is the set for all sectors. Large  $\tau$  represents the preference for the probabilities to become one sector dominates, and small  $\tau$  represents a preference for probabilities to be more distributed across sectors.

Then the article-level polarity is converted into a daily time series, and one can generate both the market-wide and sector sentiment indices. Define the global daily sentiment  $G_t$  as

$$G_t = \frac{\sum_{i \in I_t} w_k(i) \times pol_i}{\sum_{i \in I_t} w_k(i)}$$

where  $w_k(i)$  represents the query weights for the news  $i$ , which downweight overly common topics and ensure broad coverage of event types, and  $I_t$  represents the news set for date  $t$ .  $G_t$  is the weighted average tone of all news on the day  $t$ , where the query's importance weight is applied to each article. On the other hand, define the sentiment index for a given sector  $s$  as

$$S_{s,t} = \frac{\sum_{i \in I_t} (w_k(i) \times p_{i,s}) \times pol_i}{\sum_{i \in I_t} w_k(i) \times p_{i,s}}.$$

The sector sentiment index  $S_{s,t}$  is a soft sector assignment that allows one article to contribute to multiple sectors with partial probability  $p_{i,s}$ . Consequently, the effective sector news volume can be defined with

$$N_{s,t}^{(s)} = \sum_{i \in I_t} p_{i,s}.$$

If all articles are purely assigned, this is close to the number of sector articles; otherwise, if assignments are diffuse, it is fractional.

Finally, consider the sector idiosyncratic sentiment, which measures the sector sentiment net of market-wide tone. Define

$$I_{s,t} = S_{s,t} - G_t.$$

Here,  $G_t$  is a common news tone affecting all sectors and  $I_{s,t}$  is the sector-specific deviation. Positive sector idiosyncratic sentiment means sector news is better than the market's average tone that day. In contrast, negative means sector news is worse than the market's average tone that day.

### 3. Regression analysis

This section provides a detailed explanation of the regression analysis used to examine the relationship between sentiment and market returns for sector indices.

#### 3.1. Independent variable transformations

To avoid the look-ahead prediction, this section considers using the lagged sentiment index  $G_{t-1}$  and  $I_{s,t-1}$  as independent variables for the analysis of returns on day  $t$ . Additionally, to reduce the influence of outliers, one can consider the winsorized series as

$$Win(x_t) = \min\left(\max\left(x_t, Q_p(x)\right), Q_{1-p}(x)\right),$$

for any series  $x_t$ , where  $Q_p(x)$  is the  $p$ -quantile of  $x_t$  over history; in practice, one can choose  $p = 0.01$ , so that only the values between the 1st and 99th percentiles of  $x$  are used. Additionally, denote the standardized z-score series of series  $x_t$  as

$$Z(x_t) = \frac{x_t - \bar{x}}{\sigma(x)},$$

where  $\bar{x}$  is the sample mean and  $\sigma(x)$  is the sample standard deviation. Apply these standardization transformations to the lagged series and denote

$$\tilde{G}_{t-1} = Z(\text{Win}(G_{t-1})), \tilde{I}_{s,t-1} = Z(\text{Win}(I_{s,t-1})).$$

$\tilde{G}_{t-1}$  and  $\tilde{I}_{s,t-1}$  preserve the trend and direction of  $G_{t-1}$  and  $I_{s,t-1}$  but reduce the impact of outliers or high leverage observations, which help capture the main relationship between the sentiment and the index returns.

In addition to the sentiment variables, one can also consider the control variables, which include the daily percent change in volatility index ( $VIX, \Delta VIX_t$ ), daily percent change in TNX (10Y yield index proxy,  $\Delta TNX_t$ ), global news volume ( $\log, l_t^{(g)} = \log(1 + N_t^{(g)})$ ) and sector news volume ( $\log, l_{s,t}^{(s)} = \log(1 + N_{s,t}^{(s)})$ ).

### 3.2. Model specification

Let  $S$  denote the set of sectors (11 GICS-style sectors). For each sector  $s$  in  $S$ , use the corresponding U.S. sector ETF as a liquid sector proxy. Let  $P_{s,t}$  denote the adjusted close price of sector ETF  $s$  at date  $t$ . Sector returns are:

$$r_{s,t} = \frac{P_{s,t} - P_{s,t-1}}{P_{s,t-1}}.$$

This study also obtains market controls using the market ETF (SPY), and denotes market return as  $r_{m,t}$ . Consequently, define excess returns as:

$$r_{s,t}^x = r_{s,t} - r_{m,t}.$$

To formally assess whether the sentiment index score has predictive power for short-term stock returns, one can estimate linear regression models for each sector in the following forms for total sector returns and excess sector returns, respectively. The specification takes the following form:

#### 1. Total return regression (panel with sector fixed effects)

$$r_{s,t} = \alpha + \beta_g \tilde{G}_{t-1} + \beta_i \tilde{I}_{s,t-1} + \gamma r_{s,t-1} + \delta_1 r_{m,t} + \delta_2 \Delta VIX_t + \delta_3 \Delta TNX_t + \delta_4 \ell_t^{(g)} + \delta_5 \ell_{s,t}^{(s)} + \eta_s + \varepsilon_{s,t}.$$

where  $\alpha$  is an intercept,  $\eta_s$  is a sector fixed effect (a constant shift for each sector),  $\varepsilon_{s,t}$  is the regression error term.

#### 2. Excess return regression (sector beyond market)

$$r_{s,t}^x = \alpha + \beta_g \tilde{G}_{t-1} + \beta_i \tilde{I}_{s,t-1} + \gamma r_{s,t-1}^e + \delta_2 \Delta VIX_t + \delta_3 \Delta TNX_t + \delta_4 \ell_t^{(g)} + \delta_5 \ell_{s,t}^{(s)} + \eta_s + \varepsilon_{s,t}.$$

Notice that in the excess return regression,  $r_{m,t}$  is removed from the model because it is already subtracted from  $r_{s,t}^x$ . Additionally, given that daily returns often have heteroskedasticity and serial correlation, the empirical study reports HAC (Newey–West) standard errors with  $L=5$  lags. Inference uses robust covariance estimates rather than assuming independent identical distributed errors.

## 4. Real data analysis

### 4.1. Data description

For the real market analysis, this section evaluated 11 key industry sectors in the US stock market using the Global Industry Classification Standard, and the ETF indices for these sectors are listed in the following table.

Table 1. Industry sectors and ETF indices

| Sector                 | ETF Index |
|------------------------|-----------|
| Communication Services | XLC       |
| Consumer Discretionary | XLY       |
| Consumer Staples       | XLP       |
| Energy                 | XLE       |
| Financials             | XLF       |
| Health Care            | XLV       |
| Industrials            | XLI       |
| Materials              | XLB       |
| Real Estate            | XLRE      |
| Technology             | XLK       |
| Utilities              | XLU       |

For each industry sector, this study selected panel data from 2024-01-01 to 2025-09-30, with a total of 425 unique trading dates and 34117 news items across all industries. The experiment constructs the sentiment dataset from global news articles retrieved via the GDELT 2.1 event/news archive using the Python package gdeltdoc (i.e., GdeltDoc with keyword-based Filters). For each query phrase in a weighted keyword bank, articles are pulled in weekly date chunks over the fixed sample window (2024-01-01 to 2025-09-30), deduplicated by URL/title/date. The article-level sentiment is computed using TextBlob for each article title, producing a per-article sentiment score in  $[-1,1]$  along with metadata, and later aggregated into daily global and sector-level sentiment indices.

#### 4.2. Regression summary

The regression models are estimated using daily returns for sector ETFs (XLRE, XLK, XLU, etc.) and sentiment indices derived from news headlines. For each sector, the study regresses daily returns on lagged sentiment index values, controlling for market returns and volatility. The models include fixed effects to account for unobserved heterogeneity across sectors and robust standard errors to address autocorrelation in the panel data. This section reports the regression coefficients, t-statistics, p-values, and the upper and lower bounds for the 95% confidence interval of each coefficient. Table 2 displays the empirical results of the total return regression, with the Consumer Services sector serving as the reference group. All other sectors are represented in the regression using sector-fixed-effect dummy variables.

Table 2. Total return regression with sector fixed effects and lagged sentiment (HAC standard errors)

|                                    | Coef           | Std err       | t-stats        | p-value       | Lower Bound    | Upper Bound   |
|------------------------------------|----------------|---------------|----------------|---------------|----------------|---------------|
| Intercept                          | 0.0015         | 0.0020        | 0.7400         | 0.4590        | -0.0020        | 0.0050        |
| C(sector)[T.ConsumerDiscretionary] | 0.0005         | 0.0010        | 0.5260         | 0.5990        | -0.0010        | 0.0020        |
| C(sector)[T.ConsumerStaples]       | -0.0007        | 0.0000        | -1.5500        | 0.1210        | -0.0020        | 0.0000        |
| C(sector)[T.Energy]                | -0.0013        | 0.0010        | -1.4250        | 0.1540        | -0.0030        | 0.0000        |
| C(sector)[T.Financials]            | 0.0003         | 0.0010        | 0.4110         | 0.6810        | -0.0010        | 0.0020        |
| C(sector)[T.HealthCare]            | <b>-0.0015</b> | <b>0.0010</b> | <b>-2.3970</b> | <b>0.0170</b> | <b>-0.0030</b> | <b>0.0000</b> |
| C(sector)[T.Industrials]           | -0.0005        | 0.0000        | -1.3050        | 0.1920        | -0.0010        | 0.0000        |
| C(sector)[T.Materials]             | <b>-0.0017</b> | <b>0.0010</b> | <b>-2.0950</b> | <b>0.0360</b> | <b>-0.0030</b> | <b>0.0000</b> |
| C(sector)[T.RealEstate]            | -0.0004        | 0.0010        | -0.7040        | 0.4820        | -0.0020        | 0.0010        |
| C(sector)[T.Technology]            | -0.0008        | 0.0010        | -0.9900        | 0.3220        | -0.0020        | 0.0010        |
| C(sector)[T.Utilities]             | 0.0005         | 0.0010        | 0.5490         | 0.5830        | -0.0010        | 0.0020        |
| sent_global_l1_z                   | 0.0002         | 0.0000        | 1.3260         | 0.1850        | -0.0001        | 0.0000        |
| sent_idio_l1_z                     | 0.0000         | 0.0000        | -0.2330        | 0.8160        | 0.0000         | 0.0000        |

Table 2. (continued)

|              |                |               |                |               |                |                |
|--------------|----------------|---------------|----------------|---------------|----------------|----------------|
| ret_l1       | -0.0047        | 0.0140        | -0.3380        | 0.7350        | -0.0320        | 0.0220         |
| mkt_ret      | <b>0.7075</b>  | <b>0.0450</b> | <b>15.7090</b> | <b>0.0000</b> | <b>0.6190</b>  | <b>0.7960</b>  |
| vix_chg      | -0.0075        | 0.0030        | -2.3030        | 0.0210        | -0.0140        | -0.0010        |
| tnx_chg      | <b>-0.0734</b> | <b>0.0120</b> | <b>-6.0180</b> | <b>0.0000</b> | <b>-0.0970</b> | <b>-0.0490</b> |
| log_n_global | 0.0003         | 0.0010        | 0.3840         | 0.7010        | -0.0010        | 0.0020         |
| log_n_sector | -0.0009        | 0.0010        | -1.0530        | 0.2920        | -0.0020        | 0.0010         |

Table 2 reports the estimated coefficients from a sector-level total return regression that includes sector fixed effects, lagged global and idiosyncratic sentiment, lagged returns, and standard market controls. The dominant driver is the market return: mkt\_ret loads strongly and positively (0.7075,  $t=15.71$ ,  $p<0.001$ ), indicating sector ETF returns co-move tightly with the broad market. Among macro controls, increases in VIX are associated with lower sector returns (vix\_chg =  $-0.0075$ ,  $p=0.021$ ), and rising Treasury yields also predict significantly lower sector returns (tnx\_chg =  $-0.0734$ ,  $p<0.001$ ). The sentiment variables are not statistically significant at conventional levels in this specification (sent\_global\_l1\_z:  $p=0.185$ ; sent\_idio\_l1\_z:  $p=0.816$ ), suggesting that once broad market and macro conditions are controlled, lagged news sentiment adds limited incremental explanatory power for next-period total sector returns in this run. Finally, relative to the omitted baseline sector, the sector fixed-effect estimates are mostly insignificant. Still, Health Care ( $p=0.017$ ) and Materials ( $p=0.036$ ) show significantly lower average returns over the sample after controlling for the included covariates.

Table 3 summarizes the excess-return regression results. VIX changes show a strong positive effect (coef = 0.0191,  $t = 7.20$ ,  $p < 0.001$ ), while 10-year yield changes have a strong negative impact (coef =  $-0.0811$ ,  $t = -6.23$ ,  $p < 0.001$ ). Lagged global sentiment is positive but not significant (coef = 0.0002,  $p = 0.178$ ), and lagged idiosyncratic sentiment is negligible ( $p = 0.813$ ). Health Care and Materials sectors report notably lower excess returns (Health Care:  $-0.0016$ ,  $p = 0.014$ ; Materials:  $-0.0018$ ,  $p = 0.030$ ); other sector effects are insignificant. The lagged return term is marginally negative (coef =  $-0.0296$ ,  $p = 0.118$ ), indicating limited short-term reversal.

Table 3. Excess return regression with sector fixed effects and lagged sentiment (HAC standard errors)

|                                    | Coef           | Std err       | t-stats        | p-value       | Lower Bound    | Upper Bound    |
|------------------------------------|----------------|---------------|----------------|---------------|----------------|----------------|
| Intercept                          | 0.0016         | 0.0020        | 0.7750         | 0.4380        | -0.0020        | 0.0060         |
| C(sector)[T.ConsumerDiscretionary] | 0.0006         | 0.0010        | 0.6340         | 0.5260        | -0.0010        | 0.0020         |
| C(sector)[T.ConsumerStaples]       | -0.0007        | 0.0010        | -1.4640        | 0.1430        | -0.0020        | 0.0000         |
| C(sector)[T.Energy]                | -0.0014        | 0.0010        | -1.4660        | 0.1430        | -0.0030        | 0.0000         |
| C(sector)[T.Financials]            | 0.0004         | 0.0010        | 0.5060         | 0.6130        | -0.0010        | 0.0020         |
| C(sector)[T.HealthCare]            | <b>-0.0016</b> | <b>0.0010</b> | <b>-2.4530</b> | <b>0.0140</b> | <b>-0.0030</b> | <b>0.0000</b>  |
| C(sector)[T.Industrials]           | -0.0006        | 0.0000        | -1.3720        | 0.1700        | -0.0010        | 0.0000         |
| C(sector)[T.Materials]             | <b>-0.0018</b> | <b>0.0010</b> | <b>-2.1700</b> | <b>0.0300</b> | <b>-0.0030</b> | <b>0.0000</b>  |
| C(sector)[T.RealEstate]            | -0.0004        | 0.0010        | -0.6200        | 0.5350        | -0.0020        | 0.0010         |
| C(sector)[T.Technology]            | -0.0009        | 0.0010        | -1.1180        | 0.2630        | -0.0020        | 0.0010         |
| C(sector)[T.Utilities]             | 0.0006         | 0.0010        | 0.6420         | 0.5210        | -0.0010        | 0.0020         |
| sent_global_l1_z                   | 0.0002         | 0.0000        | 1.3470         | 0.1780        | -0.0001        | 0.0000         |
| sent_idio_l1_z                     | 0.0000         | 0.0000        | -0.2370        | 0.8130        | 0.0000         | 0.0000         |
| ret_l1                             | -0.0296        | 0.0190        | -1.5630        | 0.1180        | -0.0670        | 0.0080         |
| vix_chg                            | <b>0.0191</b>  | <b>0.0030</b> | <b>7.2010</b>  | <b>0.0000</b> | <b>0.0140</b>  | <b>0.0240</b>  |
| tnx_chg                            | <b>-0.0811</b> | <b>0.0130</b> | <b>-6.2330</b> | <b>0.0000</b> | <b>-0.1070</b> | <b>-0.0560</b> |
| log_n_global                       | 0.0003         | 0.0010        | 0.3450         | 0.7300        | -0.0010        | 0.0020         |
| log_n_sector                       | -0.0010        | 0.0010        | -1.1710        | 0.2410        | -0.0030        | 0.0010         |

The regression is repeated for down-market periods ( $\text{mkt\_ret} < 0$ ), focusing on industry excess returns explained by global sentiment, industry-idiosyncratic sentiment, macro controls, and sector fixed effects. Table 4 shows the regression results, which indicate that idiosyncratic sentiment is significant during downturns:  $\text{sent\_idio\_1l\_z} = -0.0006$  ( $p = 0.022$ ). A one-standard-deviation rise in sector-specific sentiment predicts lower next-day sector performance relative to the market. This suggests sector sentiment drives cross-industry dispersion mainly during stress periods. These results differ from the full-sample results, in which sentiment had little impact. These results tell that the impact of sentiment on returns is sector-specific and not equivalent across all market regimes. During downturns, the impact of sentiment is more significant.

Table 4. Down-market regime: excess industry returns and news sentiment (HAC standard errors)

|                                    | Coef           | Std err       | t-stats        | p-value       | Lower Bound    | Upper Bound    |
|------------------------------------|----------------|---------------|----------------|---------------|----------------|----------------|
| Intercept                          | 0.0071         | 0.0060        | 1.2360         | 0.2160        | -0.0040        | 0.0180         |
| C(sector)[T.ConsumerDiscretionary] | -0.0024        | 0.0030        | -0.9130        | 0.3610        | -0.0070        | 0.0030         |
| C(sector)[T.ConsumerStaples]       | <b>0.0060</b>  | <b>0.0020</b> | <b>3.1570</b>  | <b>0.0020</b> | <b>0.0020</b>  | <b>0.0100</b>  |
| C(sector)[T.Energy]                | 0.0015         | 0.0020        | 0.6450         | 0.5190        | -0.0030        | 0.0060         |
| C(sector)[T.Financials]            | 0.0022         | 0.0020        | 0.9910         | 0.3210        | -0.0020        | 0.0070         |
| C(sector)[T.HealthCare]            | 0.0016         | 0.0020        | 0.7550         | 0.4500        | -0.0030        | 0.0060         |
| C(sector)[T.Industrials]           | -0.0008        | 0.0010        | -0.7880        | 0.4310        | -0.0030        | 0.0010         |
| C(sector)[T.Materials]             | 0.0005         | 0.0020        | 0.2220         | 0.8240        | -0.0040        | 0.0050         |
| C(sector)[T.RealEstate]            | 0.0022         | 0.0020        | 1.2860         | 0.1990        | -0.0010        | 0.0050         |
| C(sector)[T.Technology]            | <b>-0.0055</b> | <b>0.0030</b> | <b>-2.1650</b> | <b>0.0300</b> | <b>-0.0110</b> | <b>-0.0010</b> |
| C(sector)[T.Utilities]             | 0.0051         | 0.0030        | 1.8570         | 0.0630        | 0.0000         | 0.0100         |
| sent_global_1l_z                   | -0.0002        | 0.0000        | -0.5440        | 0.5870        | -0.0010        | 0.0000         |
| sent_idio_1l_z                     | <b>-0.0006</b> | <b>0.0000</b> | <b>-2.2940</b> | <b>0.0220</b> | <b>-0.0010</b> | <b>-0.0001</b> |
| ret_1l                             | 0.0375         | 0.0610        | 0.6140         | 0.5390        | -0.0820        | 0.1570         |
| vix_chg                            | 0.0091         | 0.0080        | 1.2030         | 0.2290        | -0.0060        | 0.0240         |
| tnx_chg                            | <b>-0.1178</b> | <b>0.0320</b> | <b>-3.6400</b> | <b>0.0000</b> | <b>-0.1810</b> | <b>-0.0540</b> |
| log_n_global                       | -0.0008        | 0.0020        | -0.3330        | 0.7390        | -0.0050        | 0.0040         |
| log_n_sector                       | -0.0007        | 0.0020        | -0.2880        | 0.7740        | -0.0060        | 0.0040         |

## 5. Conclusion

This study examines whether emotion embedded in news headlines has predictive power for next-day sector index returns and sector excess returns, using a GDELT-based sentiment pipeline. This paper constructs daily global sentiment and sector-level sentiment, decomposes sentiment into a market-wide component and a sector-idiosyncratic component, and then estimates HAC-robust panel regressions with sector fixed effects and standard market controls. The results show that, in the full sample, global and idiosyncratic sentiment are generally weak predictors once market and macro controls are included, suggesting that headline emotion does not yield a uniform return premium across sectors. However, the relationship becomes notably stronger in down-market regimes: when the market return is negative, sector-idiosyncratic sentiment significantly predicts sector excess returns with a negative coefficient. Overall, the findings indicate that news emotion is regime-dependent: it is not a universal predictor in regular times, but it becomes informative during market stress, helping explain cross-industry dispersion beyond aggregate market moves. Future research could improve sector coverage by reducing classification imbalance and exploring firm-level heterogeneity or higher-frequency returns to better capture short-lived sentiment transmission.

## References

- [1] Gruhl, Daniel, et al. "The predictive power of online chatter." Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. 2005.

- [2] Asur, Sitaram, and Bernardo A. Huberman. "Predicting the future with social media." 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology. Vol. 1. IEEE, 2010.
- [3] Mishne, Gilad, and Natalie S. Glance. "Predicting movie sales from blogger sentiment." AAAI spring symposium: computational approaches to analyzing weblogs. 2006.
- [4] Schumaker, Robert P., and Hsinchun Chen. "Textual analysis of stock market prediction using breaking financial news: The AZFin text system." *ACM Transactions on Information Systems (TOIS)* 27.2 (2009): 1-19.
- [5] Wysocki, Peter D. "Cheap talk on the web: The determinants of postings on stock message boards." University of Michigan Business School Working Paper 98025 (1998).
- [6] Liu, Yang, et al. "ARSA: a sentiment-aware model for predicting sales performance using blogs." *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 2007.
- [7] Das, Sanjiv R., and Mike Y. Chen. "Yahoo! for Amazon: Sentiment extraction from small talk on the web." *Management science* 53.9 (2007): 1375-1388.
- [8] Nguyen, Thien Hai, Kiyoaki Shirai, and Julien Velcin. "Sentiment analysis on social media for stock movement prediction." *Expert Systems with Applications* 42.24 (2015): 9603-9611.
- [9] Maragoudakis, Manolis, and Dimitrios Serpanos. "Exploiting financial news and social media opinions for stock market analysis using MCMC Bayesian inference." *Computational Economics* 47.4 (2016): 589-622.
- [10] Sehgal, Vivek, and Charles Song. "Sops: stock prediction using web sentiment." *Seventh IEEE international conference on data mining workshops (ICDMW 2007)*. IEEE, 2007.
- [11] Bahceci, Oktay, and Oscar Alsing. "Stock market prediction using social media analysis." (2015).
- [12] De Smedt, T., & Daelemans, W. (2012). "Pattern for python." *The Journal of Machine Learning Research*, 13(1), 2063-2067.