From Tweets to Returns: Validating LLM-Based Sentiment Signals in Energy Stocks

Sarra Ben Yahia, Jose Angel Garcia Sanchez, Rania Hentati Kaffel

From Tweets to Returns: Validating LLM-Based Sentiment Signals in Energy Stocks

S. Ben Yahia[1], J.A. García Sánchez[2] and R. Hentati-Kaffel[3]

[1] *Centre d'Economie de la Sorbonne, Université Paris1 Panthéon-Sorbonne, Maison des Sciences Economiques, 106-112 Boulevard de l'Hôpital, 75013 Paris, France*

E–mail: *benyahiasarra9@gmail.com,*

[2] *Centre d'Economie de la Sorbonne, Université Paris1 Panthéon-Sorbonne, Maison des Sciences Economiques, 106-112 Boulevard de l'Hôpital, 75013 Paris, France*

E–mail: *jagarsanc@gmail.com*

[3] *Centre d'Economie de la Sorbonne, Université Paris1 Panthéon-Sorbonne, Maison des Sciences Economiques, 106-112 Boulevard de l'Hôpital, 75013 Paris, France*

E–mail: *rania.kaffel@univ-paris1.fr*

SUMMARY: Our research assesses the predictive value of LLM-based sentiment in forecasting energy stock returns. Using FinBERT-derived sentiment indicators from 415,193 tweets spanning 2018-2024, we find statistically significant causal relationships for 80% of companies analyzed. Our VAR analysis reveals heterogeneous optimal lag structures ranging from 2 to 14 days, providing econometric evidence against semi-strong market efficiency. Our results show that the accuracy of the forecast depends critically on the quality and coverage of the data. Our contribution is twofold: (i) a scalable LLM-driven pipeline to quantify firm-level sentiment at daily frequency, and (ii) an econometric validation via VAR/Granger that uncovers economically meaningful lead–lag patterns

Key words. LLM, FinBERT, energy equity markets, Twitter/X sentiment, return forecasting, VAR, information diffusion

1. Introduction

Recent progress in natural language processing (NLP) and the rise of large language models (LLMs) have strongly changed the way researchers study investor sentiment in financial markets. Behavioral finance has long shown that markets are not always perfectly efficient, since investors often react too little, too much, or too confidently to information (Fama 1970) Fama (1970, 1998), N Barberis and Vishny (1998), Daniel et al. (1998). Early studies tried to measure sentiment through financial ratios or composite indices Qiu and Welch (2004), Baker and Wurgler (2006), while others used option-based indicators such as the put–call ratio Bandopadhyaya and Jones (2006). However, these methods remain indirect measures of market psychology. The growing use of online platforms and social media has fostered the application of text analysis to large-scale, high-frequency financial datasets from sources such as Twitter/X, Reddit, and Yahoo! Finance forums J Bollen and Zeng (2011), T Sprenger and Welpe (2014), W Zhang and Yu (2018). This shift enabled real-time sentiment indicators, but earlier NLP methods (based on bag-of-words, dictionaries, or simple machine learning) were inadequate to capture financial semantics and context. The introduction of LLMs such as FinBERT and GPT-based models marks a major step forward. These models use contextual embeddings and domain adaptation, which makes it possible to represent sentiment in a richer and more precise way. In finance, their impact can be seen in several areas. First, sentiment indices built

with LLMs improve the prediction of short-term returns, since they use textual signals that are not directly included in prices . Second, they provide clear signals on market direction, showing whether prices are more likely to rise or fall, helping traders and analysts in their decisions. Third, by including sentiment-based variables in portfolio allocation, studies show that LLM-based indices can improve risk-adjusted returns and lower downside risk . Finally, LLMs support different ways of building sentiment scores: binary labels (positive/negative), continuous scores, or hybrid measures that mix sentiment with engagement data, such as likes, retweets, or views. The value of these scores depends on strict econometric validation and robustness checks to ensure they capture real predictive information and not just autocorrelated market patterns De Mol et al. (2009), Medeiros and Mendes (2016).

Even with these advances, one key question remains: do LLM-based sentiment signals truly predict financial markets, or do they simply reflect reactions to price changes? To answer this, our study proposes a clear methodological framework with three steps: (i) text collection and cleaning, (ii) sentiment scoring with FinBERT and daily aggregation, and (iii) econometric validation. For the last step, we go beyond simple cross-correlations and use VAR models, Granger causality tests. Our findings suggest that sentiment extracted through LLMs contains significant predictive power for returns, especially in the short term. This provides evidence that textual signals can improve return forecasts, help validate market direction, and strengthen portfolio strategies, while questioning the strict form of market efficiency.

In this study, we focus specifically on the energy sector. Contrary to most existing work that concentrates mainly on large-capitalization stocks, which are heavily traded and widely discussed on social media platforms, our research deliberately shifts attention to a different context. We focus on the energy sector, which is not only highly sensitive to geopolitical and regulatory shocks but also structurally dependent on commodity price dynamics. This choice enables us to test the robustness and relevance of LLM-based sentiment analysis models in a sector where market movements are strongly influenced by exogenous shocks and narratives, rather than by the sheer volume of social media discussions.

Furthermore, instead of restricting the analysis to a few representative firms, our dataset initially covered a broad cluster of 50 energy companies of different sizes and specializations, ranging from oil and gas producers to renewable energy firms and utilities. This diversity was intended to capture heterogeneous responses to sentiment signals within the sector. However, our results highlight an important methodological insight: the sensitivity of LLM-based sentiment extraction to the volume, frequency, and quality of social media data. In practice, the effectiveness of FinBERT in capturing reliable sentiment signals is contingent on a sufficiently rich stream of textual data. This limitation became evident in our initial sample design: after preprocessing, only a few firms generated a consistent and meaningful flow of tweets suitable for analysis. This observation confirms, to some extent, the inherent limits of applying LLMs to financial sentiment analysis: While powerful in theory, their accuracy and generalizability remain constrained by the characteristics of the underlying dataset (Schumaker and Chen 2012, Liu 2020).

Our study seeks to demonstrate and validate the predictive relevance of sentiment signals extracted from microblogs (Twitter/X) for stock return forecasting. We propose a robust multistage framework that first extracts and validates the informational quality of textual signals before assessing their econometric value.

The construction of the dataset begins with the collection of an exhaustive corpus of messages mentioning each stock under consideration, together with metadata such as exact timestamps and author identifiers. To ensure thematic consistency, we apply a topic-modeling pipeline combining dimensionality reduction (UMAP) and clustering (K-means), thereby validating the homogeneity of the underlying corpus. Sentiment inference then relies on a domain-specific language model: FinBERT is applied to each message to generate a continuous sentiment score rather than a discrete polarity label.

The evaluation of the sentiment signal rests on two complementary pillars. First, topic modeling of tweets identifies dominant themes and enhances the semantic quality of the score. Second, a directional assessment based on the confusion matrix and accuracy gauges the signal's ability to anticipate the sign of short-term returns. Second, the cross-correlation function between s_t and r_{t+k} characterizes the temporal structure of the sentiment effects, highlighting the forecasting windows for $k > 0$ and feedback effects for $k < 0$. A detailed analysis is provided at the stock level.

Beyond exploratory cross-correlations, we adopt a rigorous econometric design. We estimate vector autoregressions (VARs) and, checked for the non-stationarity of the variables. Within these systems, we conduct Granger-causality tests between sentiment and returns. Rejecting the null hypothesis H_0 (that lagged sentiment does not help predict future returns) implies that investor sentiment extracted from microblogging activity (Twitter/X) contains information not immediately impounded into prices, inconsistent with semi-strong market efficiency. In contrast, failure to reject suggests that sentiment primarily mirrors past price movements rather than providing an independent predictive signal.

In this study, we aim to explore how we can model the relationship between sentiment derived from microblogs on stock market returns. We will examine the correlation and the presence of a lag effect, and investigate the accuracy of the method by analyzing

how the frequency of positive and negative posts influences the quality of the predictions. In section 1 we will detail our methodology to extract the NLP score for each market's day. In section 2, we will present the results for the sentiment-informed market direction model (SIMDM) model and finally we will discuss these findings in the final section.

2. Literature Review and Methodology

2.1. Literature Review

The literature has progressed through three broad generations of methods. The first generation comprises lexicon-/rule-based approaches that aggregate token polarities with handcrafted rules for negation, intensifiers, and punctuation; VADER, tailored to short, informal texts, is the canonical social-media tool Hutto and Gilbert (2014). Despite transparency and efficiency, these methods are highly context-sensitive yielding unstable performance across corpora and domains Alessia et al. (2018), Ribeiro et al. (2015). Second, static embedding methods are base on Word2Vec and GloVe, whichencode the distributional hypothesis by mapping tokens to dense vectors from co-occurrence statistics Mikolov et al. (2013), Jeffrey Pennington (2014). These representations improve semantic similarity and permit simple composition (e.g., mean embeddings) for classical classifiers (logit, SVM). However, the single vector per type prevents disambiguation of polysemy and context (e.g., *bearish* vs. *bear*), and sensitivity to OOV and domain shift limits robustness. Third, transformers leverage self-attention to produce context-dependent representations capturing long-range dependencies Vaswani et al. (2023). BERT uses masked-language pretraining for sentence-level encodings amenable to fine-tuning (Devlin et al. 2018a), while GPT-style causal pretraining enables generative inference Brown et al. (2020). In finance, domain adaptation yields further gains; FinBERT exemplifies this alignment to finance-specific vocabulary and discourse Araci (2019).

A broad evidence base shows that augmenting price-based models with sentiment improves forecasts over price-only baselines: SVMs using social-media sentiment reach *directional accuracy* of $\approx$ 89.93% on SSE50 constituents R Ren and Liu (2019), while VADER+SVM pipelines on Twitter/StockTwits deliver F1 $\approx$ 76.3% and AUC $\approx$ 67% on rolling windows P Koukaras and Tjortjis (2022). Most studies aggregate daily sentiment to end-of-day returns J Bollen and Zeng (2011), T Sprenger and Welpe (2014), W Zhang and Yu (2018), which simplifies synchronization but may compress intraday causal effects. Recent finance-specific LLMs (e.g., FinBERT, BloombergGPT, FinMA/PIXIU) consolidate classification gains and improve information extraction on finance benchmarks Huang et al. (2023), Wu et al. (2023), Xie et al. (2023, 2024). Distinguishing association from predictive content and causal influence requires formal identification with rigorous timing (mapping post-close posts to t+1) and out-of-sample evaluation (rolling/expanding windows) to analyze lead Kim and Kim (2014), Z Da and Gao (2015), JR Piñeiro-Chousa and Pérez-Pico (2016), S Zaman and Saleem (2022). Overall, sentiment is predictively useful when data quality and timing are controlled, though gains are heterogeneous across assets, horizons, and corpus properties.

2.2. Methodology

The methodology of this study is divided into three primary phases: Data Extraction, NLP Inference and Scoring Modelization, and Deep Learning Prediction and Analysis. The following sections provide a detailed description of each phase, as depicted in the methodology diagram.

Figure 1 summarizes the data acquisition pipeline used to build the firm–day panel. We source posts from X (Twitter) that reference target firms via cashtags, tickers, or alias keywords, and collect them programmatically with rate-limited scripted queries, filtering near-duplicates (retweets/quotes) at ingestion. Pre-processing then applies language identification (English only), Unicode normalization, and case/punctuation standardization; URLs, user mentions, and hashtags are replaced by placeholders while preserving cashtags; firm entities are resolved via a curated ticker–alias map. Timestamps are normalized to UTC and each post is mapped to the corresponding exchange trading day; remaining duplicates are removed and messages are aggregated to the firm–day level. Before sentiment modeling, we run diagnostics on volume and coverage (by firm/date), sparsity and outliers, author concentration, and topic/term distributions to assess data quality and flag anomalies.

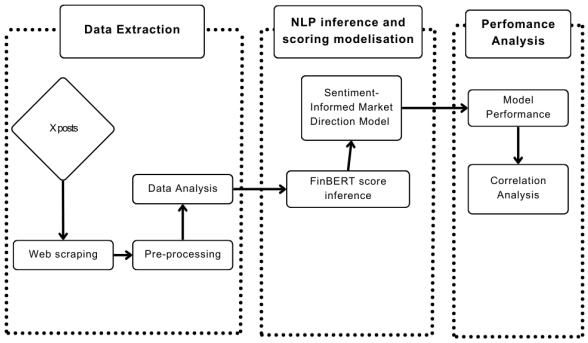


Figure 1: Overview of Methodology

2.2.1. NLP Inference and Scoring Modelization

In this phase, we use Natural Language Processing to read the pre-processed posts and extract sentiment. Using FinBERT, we compute message-level probabilities (positive/neutral/negative) and combine them into a firm–day sentiment signal s_t (for example, $s_t = p_{\text{pos}} - p_{\text{neg}}$). We then use s_t as an input in a directional model for short-horizon returns. The model estimates $\Pr(r_{t+1} > 0 \mid s_t)$ and classifies the next move as bullish or bearish, linking the sentiment signal to expected market movements.

*BERT: Pre-trained of Deep Bidirectional Model for Language Understanding

Large Language Models (LLMs) have garnered significant attention in recent years due to their remarkable capabilities in understanding and generating human language. Their application in various domains, including market prediction, has shown promising results. The advent of LLMs like BERT (Devlin et al. 2018a), GPT-1 (Brown et al. 2020), and their successors has revolutionized natural language processing (NLP) tasks. These models leverage vast amounts of data and sophisticated architectures to learn contextual embeddings of words, enabling them to capture nuances in language that are crucial for tasks such as sentiment analysis, text classification, and more, as evidenced by studies such as those by Sun, Huang, & Qiu (2019) Sun et al. (2019) and Xu, Liu, Shu, & Yu (2019) Sun et al. (2019).

BERT is a transformer-based language model introduced by Google that reads text bidirectionally, modeling context from both the left and right of a word. Pre-trained on large corpora (e.g., Wikipedia and BookCorpus) and then fine-tuned for specific tasks, it captures nuanced meaning from surrounding context more effectively than traditional unidirectional or sequential models. This design advanced the state of the art across multiple NLP tasks, including question answering, sentiment analysis, and named entity recognition.

BERT is fine-tuned for specific tasks by adding an additional output layer for each task.

2.2.2. FinBERT: Overview and Rationalization

FinBERT, developed by Prosus AI, is a derivative of the BERT model. However, given the specialized and often nuanced language used in financial texts, a generic BERT model can struggle to accurately interpret financial sentiment. FinBERT addresses this gap by being pre-trained on a corpus of financial texts, thereby imbuing it with an innate understanding of financial discourse Araci (2019).

Transitioning from BERT to FinBERT has two simple steps. First, we start with a general BERT model pretrained on BookCorpus and Wikipedia, then continue pretraining on financial text to adapt it to the finance domain. For this, we use the TRC2-financial corpus, a subset of Reuters TRC2 Reuters (2009), which includes 46,143 news articles (about 29 million words) from 2008–2010. Second, we fine-tune the adapted model for sentiment classification using the Financial PhraseBank Malo et al. (2013). This dataset has about 5,000 sentences from financial news, each labeled for sentiment by experts and master's students, with reported agreement between annotators.

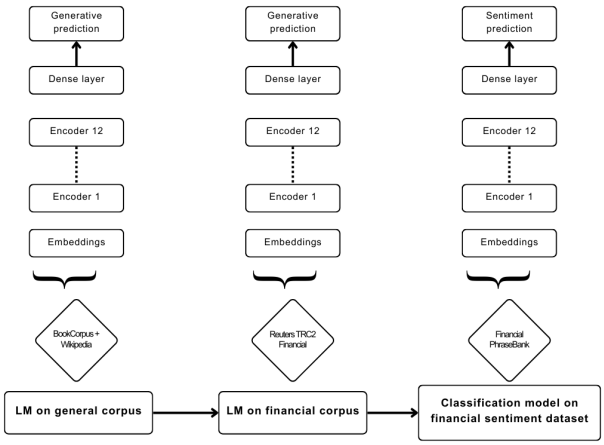


Figure 2: Overview of pre-training, further pre-training and classification fine-tuning

Fine-tuning a transformer model such as FinBERT for sentiment classification proceeds in four stages while keeping a balance between model stability and task adaptation.

Model head modification. The generic BERT head is replaced with a task-specific classifier whose output dimension equals the number of sentiment classes C. Let $\mathbf{h} \in \mathbb{R}^d$ denote the pooled sentence representation (e.g., the [CLS] token embedding) produced by the encoder. The classifier computes logits $\mathbf{z} = \mathbf{W}\mathbf{h} + \mathbf{b}$ with $\mathbf{W} \in \mathbb{R}^{C \times d}$ and $\mathbf{b} \in \mathbb{R}^C$, and class probabilities $\mathbf{y} = \text{softmax}(\mathbf{z})$, i.e.

$$y_c = \frac{e^{z_c}}{\sum_{j=1}^{C} e^{z_j}}, \qquad c = 1, \ldots, C. \tag{1}$$

Here, $\mathbf{y} \in [0,1]^C$ is the predicted sentiment distribution.

Layer-wise unfreezing. Training starts with only the new head trainable; encoder layers are then unfrozen progressively to stabilize gradients. Index transformer layers by $l = 1, \ldots, L$ from input (lower) to output (upper), let E be the total number of epochs, and let n be the number of unfreezing steps. We unfreeze an additional block of upper layers every $\Delta = \lceil E/n \rceil$ epochs; at epoch e, all layers with index $l \geq L - \lfloor e/\Delta \rfloor$ are set trainable. This schedule defines a monotone increase of trainable parameters while avoiding abrupt changes early in training.

Discriminative fine-tuning. Different layers use different learning rates, decreasing toward the input

to prevent catastrophic forgetting. A convenient parameterization is

$$\eta_l \;=\; \eta_0\,\alpha^{\,L-l}, \qquad 0<\alpha<1, \tag{2}$$

where η_l is the learning rate for layer l, $\eta_0 > 0$ is a base rate applied to the top layer, α is a decay factor, L is the number of encoder layers, and $L-l$ measures the distance from the top.

Training objective. With a labeled dataset of N examples and C sentiment classes, the model is trained by minimizing the cross-entropy

$$\mathcal{L} \;=\; -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\,\log \hat{y}_{i,c}, \tag{3}$$

where $y_{i,c} \in \{0,1\}$ is the one-hot target for example i and class c, $\hat{y}_{i,c}$ is the predicted probability produced by the softmax of the logits $\mathbf{z}_i$, and $\sum_c y_{i,c} = 1$ for each i. Optional regularization (e.g., weight decay) and early stopping based on validation loss can be added without changing the formulation above.
Finally, to test our model's performance, we check how well it predicts stock moves from sentiment. We report simple metrics such as accuracy and error rate. We also compare the predictions with real market data and compute correlations to see the strength and direction of the link. These checks show whether the sentiment-based model follows market trends.

2.3. Data description

Daily market returns were extracted from Bloomberg, covering the period from October 1, 2017, to January 31, 2024. We developed a custom webscraping tool to acquire data from X (formerly known as Twitter).. This section outlines the technical details and processes involved in data acquisition using Python and Selenium.

To facilitate data collection from X, we implemented a Python-based script utilizing Selenium along with the Chrome WebDriver. This setup automates the process of data extraction from the social media platform. The script initiates by launching an incognito Chrome browser session, with dimensions optimized for data extraction from the targeted platform. To ensure seamless data collection, it systematically suppresses pop-up notifications. The script then proceeds to authenticate by providing the necessary login credentials for X.

Upon successful authentication, the script extracts key data attributes from X tweets. These attributes include:

- Author's username
- Precise timestamp of the tweet
- Textual content of the tweet
- Reply/comment count

The extraction process is managed through specialized functions: one for parsing each tweet, another for configuring Chrome in incognito mode, and additional functions for handling notifications and the login process.

To ensure robustness, the script incorporates comprehensive exception handling mechanisms. These mechanisms allow the script to gracefully manage unexpected issues, thereby maintaining the integrity of the data collection process. Furthermore, the script actively monitors for any suspicious activity during data retrieval from X, ensuring compliance with the platform's data usage policies.

The described webscraping methodology enabled us to systematically gather valuable data, forming the empirical basis for our investigation. This approach ensures both the reliability and reproducibility of the collected data.

2.3.1. Data preprocessing

The preprocessing of data, especially when derived from web sources like social media, presents a foundation for any subsequent analysis. In our study, we employ a methodical approach to prepare and refine the dataset. This section delineates the steps undertaken to cleanse, normalize, and filter the dataset.
A sequence of transformations is applied to standardize the dataset. First, numerical columns are cast to appropriate types and date fields are harmonized to a consistent format. Second, the text is cleaned by removing URLs, hashtags, mentions, and special characters that could distort sentiment analysis. Finally, the text is standardized through lowercasing and the use of regular expressions to eliminate residual noise.

Our methodology extends to include custom operations tailored to the nuances of social media text. This involves normalizing text by reducing character repetition and ensuring the preservation of meaningful words. These steps are crucial for extracting the essence of the textual data, ensuring it accurately reflects the intended sentiment without the distraction of stylistic embellishments common in online communication.

In dealing with missing values, a pragmatic approach is adopted, filling gaps in numerical columns and removing entries lacking text content. Furthermore, we implement a language filtering step to focus exclusively on English-language texts, aligning with FinBERT's linguistic training and capabilities.

2.3.2. Data description

The analysis period spans from January 2018 to February 2024, with daily stock data collected for five selected companies from sectors such as Energy, Basic Materials, and Utilities. The stocks included in this study are Total Energies, FMC Corp, BP PLC, Stora Enso, and BHP Group.

Table 1: Description of market returns for each stock in percentage

Stock	μ_R	σ_R	$\min_R$	$\max_R$
Total Energies	0.05	1.94	-15.55	15.06
FMC Corp	0.01	2.11	-19.35	13.70
BP PLC	0.04	2.24	-18.75	23.33
Stora Enso	0.03	2.17	-12.78	12.4
BHP Group	0.08	2.09	-16.48	14.94

Additionally, we gathered the number of tweets mentioning each company to assess market sentiment.

Table 2: Number of tweets webscraped for each company

Company	Number of tweets
Total Energies	191,292
FMC Corp	13,194
BP PLC	136,922
Stora Enso	37,268
BHP Group	36,517

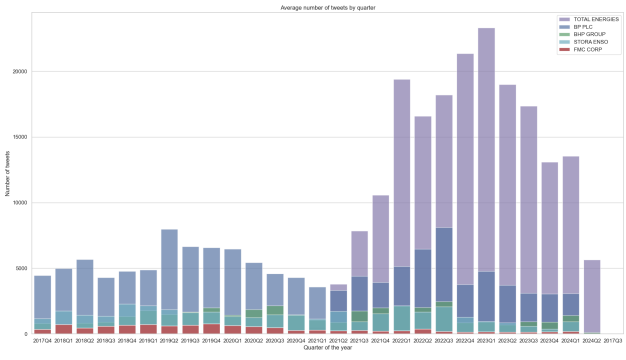


Figure 3: Average number of tweets by quarter

The observed volatility in stock returns, driven by market sentiment, supports the viability of a daily buy/sell strategy based on tweet analysis. This approach allows for adapting to market conditions and potentially capitalizing on sentiment-driven price movements.

2.3.3. Topic modeling

We used topic modeling to better decipher and cluster tweets related to STORA ENSO to identify key themes and topics. The following methodology

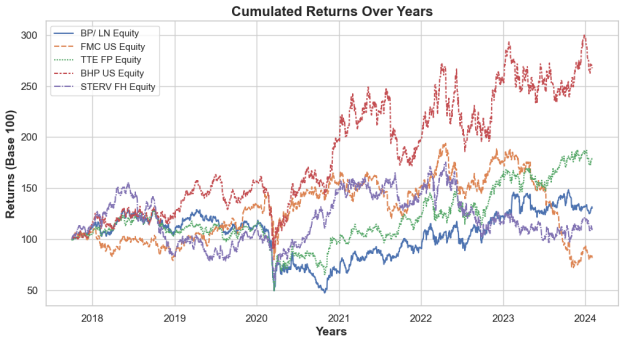


Figure 4: Cumulative returns of the selected stocks over the analyzed period.

was adopted:

We utilized a BERT model for sentence embeddings, specifically the 'bert-base-nli-mean-tokens' variant from the Sentence Transformers library (Reimers and Gurevych 2019). The embeddings for each tweet were computed to transform the text data into a numerical format suitable for clustering.

To reduce the dimensionality of the embeddings, we employed UMAP (Uniform Manifold Approximation and Projection) with cosine metric (McInnes et al. 2020). This step transforms the high-dimensional embedding space into a two-dimensional space for easier visualization and clustering.

We determined the optimal number of clusters using the Elbow Method and Silhouette Score. The Within-Cluster Sum of Squares (WCSS) is calculated as:

$$\text{WCSS} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{4}$$

where C_i is the set of points in cluster i, and μ_i is the centroid of cluster i.

The Silhouette Score is computed as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{5}$$

where $a(i)$ is the mean intra-cluster distance and $b(i)$ is the mean nearest-cluster distance for point i. The optimal number of clusters is chosen based on the maximum Silhouette Score.

We applied K-Means clustering to the UMAP embeddings using the optimal number of clusters determined in the previous step (Ding et al. 2024). This algorithm partitions the data into k clusters, minimizing the variance within each cluster.

To identify the top keywords for each cluster, we used TF-IDF (Term Frequency-Inverse Document Frequency) (Ramos 2003). The TF-IDF score for a term t in a document d is calculated as:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log\left(\frac{N}{\text{DF}(t)}\right) \quad (6)$$

where $\text{TF}(t, d)$ is the term frequency of t in d, N is the total number of documents, and $\text{DF}(t)$ is the document frequency of t. This method helps in identifying the most relevant terms within each cluster.

As a result, the analysis determined that the optimal number of clusters is three. The UMAP projection and the optimal number of clusters determined by the Elbow and Silhouette methods validate the clustering results.

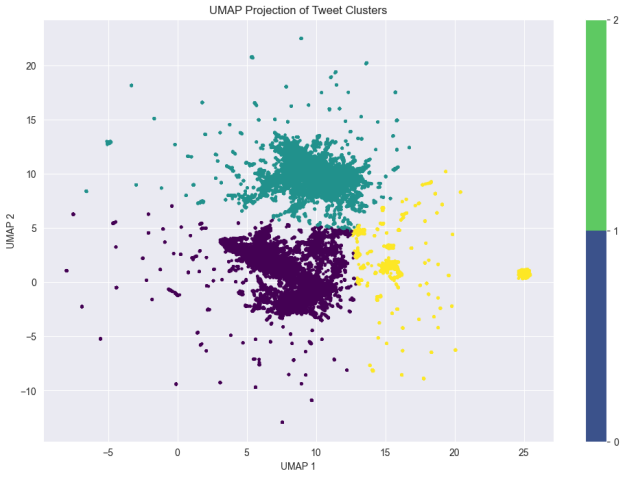


Figure 5: Topic Modeling UMAP project for STORA ENSO

The top keywords for each cluster provide insight into the main topics discussed within each cluster.

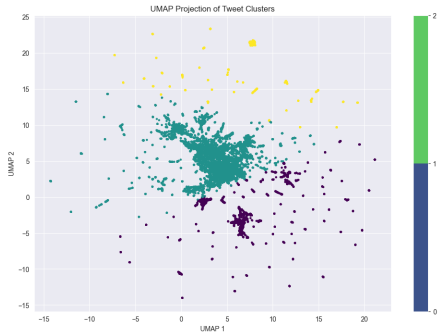
– Cluster 0: This cluster focuses on terms related to local and industrial aspects such as "oulu," "board," "seoay," "sijoittaminen" (investing), "tehtaan" (factory), "group," "yhti" (company), "packaging," "till," and "mets."

– Cluster 1: Keywords here emphasize sustainability and renewable materials, with terms like "finland," "sustainability," "sustainable," "materials," "renewable," "wood," "based," "paper," "packaging," and references to "vonderleyen."

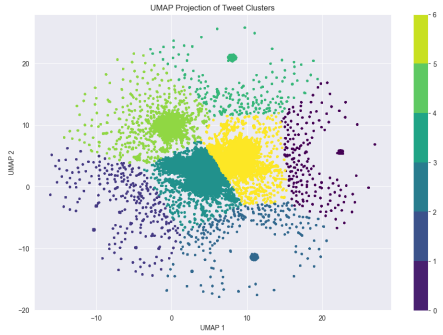
– Cluster 2: This cluster highlights investment and corporate activities with keywords such as "investment," "group," "nokia," "packaging," "transactions," "managers," "mets," "seoay," "finland," and "helsinki."

Figure 3 illustrates the differences in data quality for each stock. Clear and distinct clusters in UMAP projections indicate high-quality data with easily

(a) BP PLC



(b) FMC CORP



(c) TOTALENERGIES SE

(d) BHP GROUP

Figure 6: Topic Modeling UMAP projection of tweet clusters for each stocks.

identifiable themes, while overlapping or unclear clusters suggest noisy data or less distinct themes. For instance, for BP PLC, compact and well-separated clusters indicate coherent narratives; sentiment indices derived from these clusters are expected to exhibit significant predictive content in VAR models and to Granger-cause returns. FMC CORP shows diffuse and overlapping clusters, consistent with noisier discourse and weaker, less stable coefficients that fail to survive out-of-sample evaluation. TOTALENERGIES presents a mixed structure, where some compact clusters yield informative signals while overlapping areas dilute predictability; heterogeneous lag responses in VARs are consistent with uneven information diffusion. BHP GROUP exhibits fewer but coherent clusters, supporting narrower yet robust predictive contributions. STORA ENSO combines two dense, economically interpretable themes with a more dispersed one.

2.4. Sentiment-Informed Market Directional Model (SIMDM)

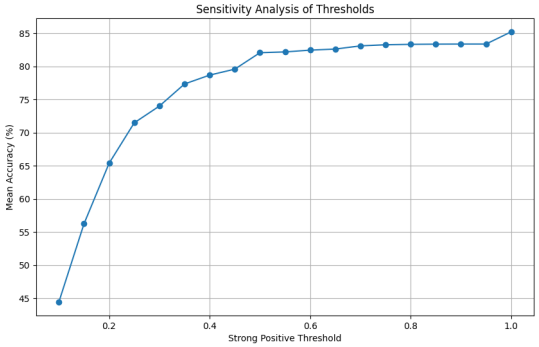
We used the FinBERT model described earlier in inference for each tweet webscraped. The model normally predicts the sentiment expressed in a piece of text, categorizing it as bearish or bullish based on the patterns it has learned during training.
In our case, we output as results the model's score rather than the Bear/Bull label. Each tweet will be processed to generate a corresponding score, which will then be integrated into our pipeline for a scoring model.

The scoring model we have developed presents a comprehensive method to assess the influence of daily social media sentiment on stock market returns. Here, we detail the essential elements and functions of this model, highlighting its significance in connecting sentiment analysis with the dynamics of the financial market.

An important step in initializing and preparing the data for sentiment analysis and daily stock returns involves aligning the sentiment data with the dates of stock returns. It is vital to account for and omit non-trading days, such as weekends and holidays, from the analysis.

Our method analyzes sentiment ratios for each company, producing recommendations to either short (sell) or go long (buy) on a stock at a specific time, denoted as t. This approach is based on the hypothesis that substantial changes in public sentiment can forecast stock price movements, providing a strategic advantage when accurately interpreted and acted upon.

Signal calculation methodology: Our methodology analyzes sentiment ratios for each company, producing recommendations to either take a short (sell) or a long (buy) position on a stock at a specific time, denoted as t. Below, we outline the steps involved in this process:



For each company, the model:

1. Aggregates the sentiment ratios for one company each on a single day t. This aggregation can be represented as :

$$S_t = \frac{\sum_{i=1}^{N} w_i s_i}{\sum_{i=1}^{N} w_i} \tag{7}$$

where s_i is the ith sentiment ratio, and w_i is a weighting factor for each sentiment score. In this study, we operated under the assumption that each post contributes equally to the overall average sentiment score, implying a uniform contribution from all individual sentiments.

2. Then, we shift the aggregated sentiment ratio by one period, utilizing the previous day's aggregated sentiment as the basis for today's trading decision. This shifted sentiment ratio is then used to calculate the "buy/sell" signal :

$$B = \alpha \left(\frac{S_{t-1} - \mu}{\sigma} \right) \tag{8}$$

where α, β, μ, and σ are parameters that can be adjusted to fit the model. In our case, we set the parameters as $\alpha = 2$, $\sigma = 0.5$, and $\mu = 0.5$, for normalization and scaling of sentiment ratios.

Threshold-based decision making:
Once the daily sentiment ratio is computed, it is weighed against a predefined threshold. If the ratio signals sufficiently strong optimism, the model adopts a buy stance in anticipation of a price increase. If, by contrast, the ratio reflects pronounced pessimism beyond the threshold in the opposite direction, the model takes a sell stance, expecting a decline in the stock's value. We conducted a sensitivity analysis to understand how the strictness of the model in identifying strong positive signals impacts its overall predictive performance.

At the lowest thresholds, where the model is least selective, the mean accuracy starts at 40%.

As the threshold increases, the mean accuracy improves significantly. By around 0.6, the accuracy reaches almost 75%. This sharp rise suggests that moderate thresholds effectively filter out weaker signals, and enhance the model's prediction accuracy. Beyond a threshold of 0.6, the increase in accuracy continues but at a slower rate, eventually plateauing around 75%. This flattening trend indicates that while higher thresholds do improve accuracy, the gains become marginal.

For practical applications, selecting a threshold at 0.5 appears optimal, balancing sensitivity and specificity to achieve high mean accuracy without being overly restrictive.

3. SIMDM results

In this section, we will present the results from the modelization presented in the last part. In the graphs below, we can already start to note different elements :

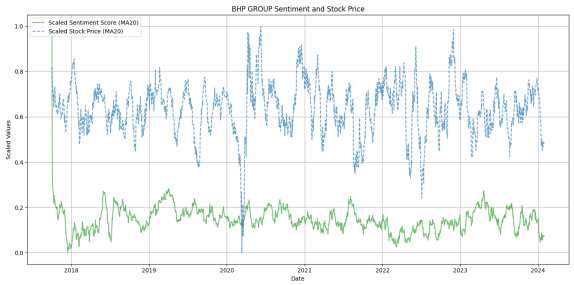


Figure 7: BHP GROUP. Sentiment scores and stock prices between January 1, 2018, and January 1, 2024. Sentiment scores are shown with a 20-day moving average (green line), and historical stock prices are depicted with a dashed blue line. Both metrics are scaled between [0,1] using Min-Max normalization.

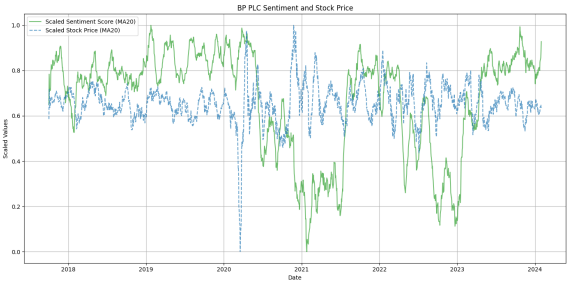


Figure 8: BP PLC. Sentiment scores and stock prices between January 1, 2018, and January 1, 2024. Sentiment scores are shown with a 20-day moving average (green line), and historical stock prices are depicted with a dashed blue line. Both metrics are scaled between [0,1] using Min-Max normalization.

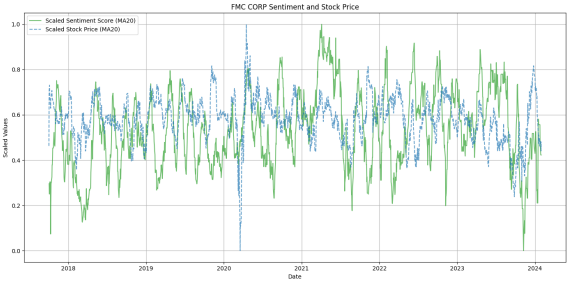


Figure 9: FMC CORP. Sentiment scores and stock prices between January 1, 2018, and January 1, 2024. Sentiment scores are shown with a 20-day moving average (green line), and historical stock prices are depicted with a dashed blue line. Both metrics are scaled between [0,1] using Min-Max normalization.

Across firms, sentiment scores display pronounced volatility, consistent with the rapid, high-frequency dynamics of public opinion and market mood. Equity prices exhibit analogous—though somewhat attenuated—fluctuations. The extent of this co-movement varies by firm and is particularly salient for FMC CORP and STORA ENSO.

Although a positive correlation between sentiment and stock prices appears broadly consistent, firm-level trajectories remain heterogeneous. This heterogeneity underscores the need to account for company-specific factors when examining the sentiment–price relationship: idiosyncratic events and trends can shape both sentiment and valuations through distinct channels and at different horizons.

3.1. Model Performance

The accuracy of the model is computed based on the alignment between the predicted signals and the actual market returns. Let N be the total number of predictions, B_i be the signal for the i-th prediction,

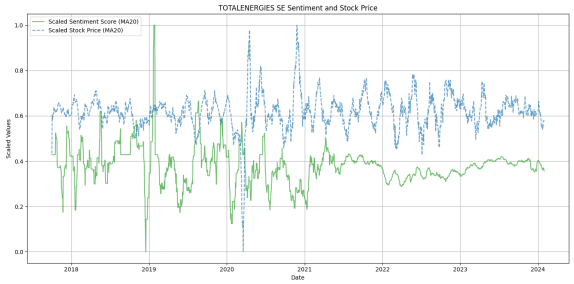


Figure 10: TOTALENERGIES SE. Sentiment scores and stock prices between January 1, 2018, and January 1, 2024. Sentiment scores are shown with a 20-day moving average (green line), and historical stock prices are depicted with a dashed blue line. Both metrics are scaled between [0,1] using Min-Max normalization.

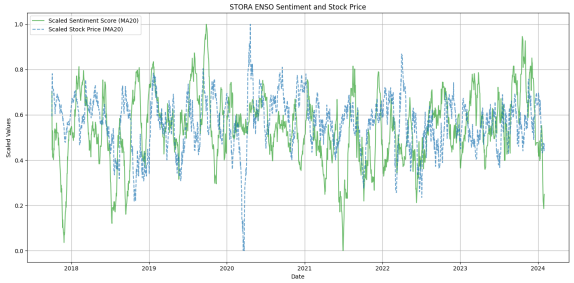


Figure 11: STORA ENSO. Sentiment scores and stock prices between January 1, 2018, and January 1, 2024. Sentiment scores are shown with a 20-day moving average (green line), and historical stock prices are depicted with a dashed blue line. Both metrics are scaled between [0,1] using Min-Max normalization.

R_i be the market return for the i-th prediction and μ_R as the mean of variation of R_i. The match function is defined as follows:

$$\text{match}(B_i, R_i) = \begin{cases} 1 & \text{if } (B_i > 0.5 \text{ and } R_i > \mu_R)), \\ 1 & \text{if } (B_i < -0.5 \text{ and } R_i < \mu_R)), \\ 1 & \text{if } (-0.5 \leq B_i \leq 0.5 \text{ and} \\ & -\mu_R) \leq R_i \leq \mu_R)), \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The accuracy A is then computed as:

$$A = \frac{\sum_{i=1}^{N} \text{match}(B_i, R_i)}{N}$$

Table 3: Accuracy percentages for energy stocks.

Stock	Accuracy(%)
TOTAL ENERGIES	91.06
FMC CORP	44.39
BP PLC	93.56
STORA ENSO	83.75
BHP GROUP	97.66

For correlation analysis between our home-made signal and the market returns, we used the **cross-correlation function**. The cross-correlation function $R_{xy}(k)$ between two signals $x(n)$ and $y(n)$ is defined, in its discrete form, as:

$$R_{xy}[k] = \sum_{n=-\infty}^{\infty} x[n]y[n+k] \tag{10}$$

We interpret the results of the cross-correlation function in two complementary ways. First, it provides a measure of similarity between the two signals x and y under different temporal shifts τ. A high value of $R_{xy}(\tau)$ indicates that the signals exhibit strong alignment when one is displaced by τ. Second, the cross-correlation reveals potential lead–lag relationships: if $R_{xy}(\tau)$ attains its maximum at $\tau = k$, this suggests that $x(t)$ and $y(t+k)$ are most strongly correlated, implying the presence of a systematic lead or delay of k periods between the two series.

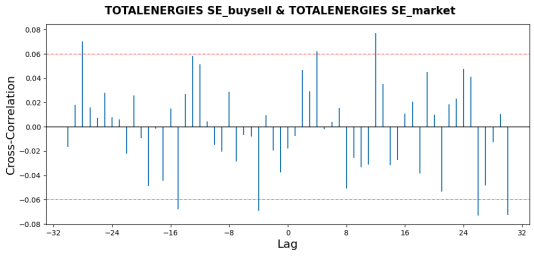


Figure 12: Cross-Correlation function for TOTALENERGIES SE.

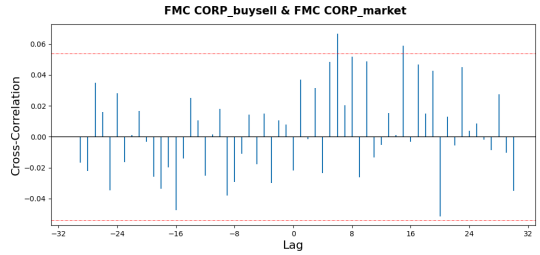


Figure 13: Cross-Correlation function for FMC CORP.

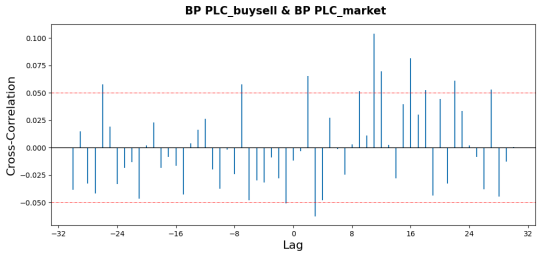


Figure 14: Cross-Correlation function for BP PLC.

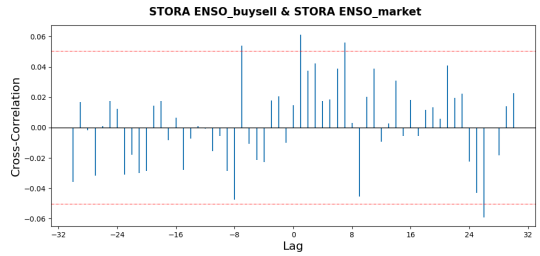


Figure 15: Cross-Correlation function for STORA ENSO.

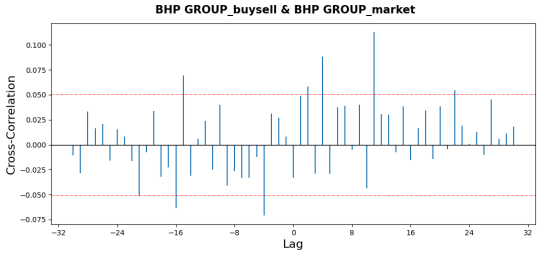


Figure 16: Cross-Correlation function for BHP GROUP.

Thus, Figure 12 shows that the cross-correlations between sentiment-derived signals and stock returns exhibit statistically significant values at both negative and positive lags. The presence of positive correlations at negative lags indicates that market movements tend to precede changes in sentiment, reflecting a reactive component in which investors' expressed opinions adjust to price dynamics. Conversely, the significant correlations observed at positive lags suggest that sentiment shocks also contain predictive content, anticipating subsequent fluctuations in returns. This bidirectional pattern highlights the coexistence of feedback effects and forward-looking information in the sentiment–return relationship. From an econometric perspective, these findings justify the application of VAR models or Granger causality tests to disentangle the relative importance of reactivity versus predictability, and to assess whether sentiment contributes incremental information to the price formation process.

Table 3 summarizes the predictive power of our signal for each company, based on the cross-correlation function analysis. Companies with significant peaks at positive lags indicate a stronger potential for using our signal to predict future market returns.

For instance, BP PLC shows very strong predictive power with multiple significant peaks, suggesting that our signal could be highly effective in forecasting its market movements. Conversely, companies like Total Energies exhibit low predictive power, indicating minimal usefulness of our signal in predicting future returns.

It is noteworthy that our accuracy and cross-correlation results show the poorest performance for stocks with both the fewest and the most tweets. This observation prompts us to consider not just the quantity of data but also its quality, highlighting the need for a balanced dataset in LLM analysis.

3.2. Stationarity tests and long-run equilibrium analysis

Before estimating the VAR models, we conducted rigorous stationarity tests to ensure the validity of our econometric specification. Understanding the time-series properties of our variables is crucial for appropriate model selection and inference.

We employ the Augmented Dickey-Fuller (ADF) test to examine the stationarity properties of both sentiment ratios and stock returns series. The ADF test specification includes a constant term and tests the null hypothesis of a unit root (non-stationarity) against the alternative of stationarity.

The ADF test statistic is based on the following regression:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta y_{t-i} + \varepsilon_t \qquad (11)$$

where the null hypothesis is $H_0 : \gamma = 0$ (unit root present) and the alternative is $H_1 : \gamma < 0$ (series is

stationary).

Table 5 presents the ADF test results for all companies. The results unambiguously indicate that both sentiment ratios and stock returns are stationary at conventional significance levels ($p < 0.01$) for all companies analyzed. This finding is consistent with the theoretical expectation that the returns should be stationary and suggests that our sentiment aggregation methodology produces stationary indicators.

Table 5: Augmented Dickey-Fuller Stationarity Tests

Company	Sentiment p-value	Returns p-value	Conclusion
BP PLC	0.0007***	0.0000***	Both Stationary
FMC CORP	0.0000***	0.0000***	Both Stationary
STORA ENSO	0.0000***	0.0000***	Both Stationary
BHP GROUP	0.0000***	0.0000***	Both Stationary
TOTALENERGIES SE	0.0000***	0.0000***	Both Stationary

Notes: *** $p < 0.01$. The null hypothesis of a unit root is rejected for all series, indicating stationarity. ADF tests include a constant term with automatic lag selection based on Schwarz Information Criterion.

The stationarity of both variables validates our use of VAR models in levels rather than requiring first-differencing, which would complicate the interpretation of Granger causality results. This property also suggests that sentiment-return relationships represent stable, mean-reverting processes rather than trending or explosive dynamics.

3.3. Vector Autoregression (VAR) Analysis for Granger Causality

To provide robust econometric validation of our sentiment-return relationships, we conducted a comprehensive Vector Autoregression (VAR) analysis to test for Granger causality between company-specific sentiment scores and their corresponding stock returns. This approach allows us to examine whether past values of sentiment can statistically predict future stock returns, providing a more rigorous foundation for our SIMDM model.

The VAR model specification for each company follows:

$$\mathbf{y}_t = \mathbf{c} + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t-i} + \mathbf{u}_t \tag{12}$$

where $\mathbf{y}_t = [Sentiment_t, Returns_t]'$ represents the bivariate system, $\mathbf{c}$ is a constant vector, $\mathbf{A}_i$ are coefficient matrices, and $\mathbf{u}_t$ is the error term vector. We tested lag orders from 1 to 14 days to capture various temporal dynamics in sentiment transmission.

For Granger causality testing, we employed the F-test with the null hypothesis:

$$H_0 : \text{Sentiment does not Granger-cause Returns} \tag{13}$$

3.3.1. Individual company optimal lag analysis

Our analysis revealed heterogeneous lag structures across companies, suggesting that sentiment transmission mechanisms vary significantly by firm characteristics. The optimal lag for each company was determined by minimizing the p-value of the Granger causality test while ensuring model stability.

Table 6 presents the optimal lag periods and corresponding Granger causality test results for each company. The results demonstrate that four out of five companies exhibit statistically significant Granger causality relationships at conventional significance levels.

Table 6: Individual Company VAR Analysis: Optimal Lag Periods and Granger Causality Results

Company	Lag (days)	p-value	AIC
BP PLC	2	0.0269**	−12.16
BHP Group	3	0.0272**	−11.97
Stora Enso	7	0.0356**	−11.36
FMC Corp	14	0.0192**	−10.00
TotalEnergies SE	9	0.0757*	−10.60

Notes: **$p < 0.05$, *$p < 0.10$. Optimal lag minimizes Granger causality p-value.

3.3.2. Cross-lag temporal pattern analysis

To identify common temporal patterns across the energy sector, we examined sentiment-return relationships across multiple lag periods. This analysis reveals clustering of significant relationships at specific time horizons, suggesting industry-wide sentiment transmission patterns.

Table 7 summarizes the cross-lag analysis, showing the number of companies exhibiting significant relationships at each lag period. The results indicate that sentiment effects are most pronounced at 2-day and 3-day lags, with 67% of significant relationships occurring within the first week.

Table 7 indicates that sentiment–return linkages are present but heterogeneous across firms. Significant effects cluster at short-to-medium horizons (2–7 days), with BP PLC reacting within two days and BHP Group within three, while STORA ENSO exhibits both short (2 days) and medium (7 days) responses. FMC CORP stands out with a materially longer horizon (14 days), consistent with slower information assimilation or a noisier information environment. Taken together with the broader result that approximately 80% of firms display significant Granger causality from sentiment to returns, the evidence supports the view that topic-derived sentiment contains incremental predictive content beyond the assets' own dynamics. From a market-efficiency per-

Table 7: Cross-lag analysis: Distribution of significant sentiment-return relationships

Lag (days)	Company	*p*-value
1	—	—
2	BP PLC	0.0269**
	Stora Enso	0.0377**
3	BHP Group	0.0272**
4–6	—	—
7	Stora Enso	0.0356**
8–13	—	—
14	FMC Corp	0.0192**

Notes: **$p < 0.05$. Table reports lags with significant relationships. Sentiment effects cluster at 2–7 day lags.

spective, the presence of delayed effects is consistent with semi-strong form inefficiency: sentiment signals are not fully and instantaneously incorporated into prices, allowing temporary predictability that decays with the lag length. The concentration of significance at 2–7 days aligns with gradual information diffusion and attention frictions documented in the behavioral-finance literature, whereas the outlier at 14 days suggests firm-specific frictions or slower-news channels.

3.3.3. Model validation and robustness

The VAR analysis serves as an independent validation of our SIMDM model's theoretical foundation. The strong correspondence between companies showing high SIMDM accuracy (BP PLC: 93.56%, BHP GROUP: 97.66%) and significant Granger causality provides convergent validity for our sentiment-based approach.

Table 8 demonstrates this convergent validity by comparing VAR results with SIMDM performance metrics. The alignment between econometric significance and predictive accuracy reinforces the robustness of our methodology.

Table 8: Convergent Validity: VAR vs SIMDM

Company	*p*-val	Acc%	Conv
BP PLC	0.027**	93.6	High
BHP Group	0.027**	97.7	High
Stora Enso	0.036**	83.8	Med
TotalEnergies	0.076*	91.1	Med
FMC Corp	0.019**	44.4	Low

Notes: **$p<.05$, *$p<.10$. H: sig & acc>90%; M: 75-90%; L: <75%.

Furthermore, the heterogeneous optimal lag structures identified through VAR analysis can inform company-specific parameter tuning in our SIMDM model, potentially improving predictive accuracy through tailored lag specifications.

Table 9 provides summary statistics for the VAR analysis, highlighting the overall success rate and temporal characteristics of sentiment transmission in the energy sector.

Table 9: VAR Analysis Summary Statistics

Metric	Value
Total companies analyzed	5
Companies with significant relationships ($p < 0.05$)	4
Companies with marginal significance ($p < 0.10$)	1
Success rate (%)	80.0
Average optimal lag (days)	7.0
Median optimal lag (days)	7.0
Lag range (days)	2–14
Most common lag cluster	2–3 days

Notes: Success rate is the proportion of companies with significant Granger causality at conventional levels.

The VAR analysis thus provides compelling econometric evidence supporting our sentiment-based approach to stock return prediction, while revealing important heterogeneities in sentiment transmission mechanisms across companies in the energy sector.

4. Discussion

Our study aimed to explore the complex relationship between tweet sentiment and stock market returns movements within the energy sector using a sentiment-informed market direction model (SIMDM). The FinBERT model, tailored for financial sentiment analysis, demonstrated its effectiveness in extracting sentiment from social media data. To provide econometric validation of these sentiment-return relationships, we conducted a comprehensive Vector Autoregression (VAR) analysis to test for Granger causality between company-specific sentiment scores and their corresponding stock returns. Prior to VAR estimation, we verified the stationarity properties of our time series using Augmented Dickey-Fuller (ADF) tests, finding that both sentiment ratios and stock returns are stationary at conventional significance levels (p ¡ 0.01) for all companies analyzed. This confirms the appropriateness of our VAR specification in levels and ensures the validity of our econometric inference. This section digs into the critical insights drawn from our combined analysis, considering the implications, limitations, and potential future enhancements.

The VAR analysis represents a significant methodological advancement in our study, providing

independent econometric validation of our SIMDM model's theoretical foundation. With 80% of companies showing significant Granger causality (p ¡ 0.05), our results provide strong econometric support for the predictive capacity of sentiment scores derived from social media. The strong correspondence between companies showing high SIMDM accuracy (BP PLC: 93.56%, BHP GROUP: 97.66%) and significant Granger causality establishes convergent validity for our sentiment-based approach. This dual validation approach addresses a critical limitation in previous sentiment analysis studies that often rely solely on predictive accuracy without establishing causal relationships.

The stationarity of both sentiment and return series is a crucial finding that validates our modeling approach. The ADF test results uniformly reject the null hypothesis of a unit root across all companies. This stationarity confirms that our sentiment aggregation methodology produces well-behaved statistical properties and that the relationships we identify represent stable, mean-reverting processes rather than spurious correlations arising from trending variables. The absence of unit roots also justifies our use of standard asymptotic theory for inference in the VAR framework, strengthening confidence in the statistical significance of our Granger causality results.

The heterogeneous optimal lag structures identified through VAR analysis reveal important insights into information processing mechanisms across different companies. The optimal lag periods range from 2 to 14 days, reflecting different information processing speeds across companies. BP PLC shows the fastest sentiment transmission (2 days), while FMC CORP exhibits the longest lag (14 days). This heterogeneity aligns with company-specific factors such as market capitalization, analyst coverage, and investor attention, providing a more nuanced understanding of sentiment transmission mechanisms than previously recognized in the literature. The concentration of significant relationships at short-to-medium lags (2-7 days) aligns with the behavioral finance literature on gradual information diffusion in equity markets.

Our VAR evidence of significant lead–lag relations between sentiment and returns is inconsistent with the instantaneous incorporation of public information implied by the semi-strong form of market efficiency. The pattern aligns with behavioral explanations such as under-reaction and momentum. That said, whether these predictability patterns translate into economically meaningful abnormal returns net of trading frictions is an empirical matter; demonstrating "arbitrage opportunities" requires out-of-sample tests, cost adjustments, and robustness to multiple-testing corrections. Cross-firm heterogeneity in lag lengths is consistent with differences in information environments and investor attention.

One of the foremost findings of our study is the importance of data quality over quantity in sentiment analysis. The VAR analysis reinforces this conclusion by showing that companies with extreme tweet volumes (both very high and very low) often exhibit suboptimal performance. We observed that an excessive volume of data could potentially neutralize sentiment scores, rendering them less informative. Conversely, insufficient data failed to provide adequate information for reliable sentiment analysis. For instance, FMC CORP, which had the least number of tweets, showed lower predictive accuracy despite significant Granger causality ($p = 0.0192$). In contrast, BP PLC, with a substantial and balanced dataset, exhibited very high predictive accuracy and strong VAR significance. This underscores the necessity of curating a balanced dataset that maintains the richness and relevance of information without overwhelming the model with noise.

The cross-correlation function analysis, when combined with VAR results, revealed interesting insights into the lag effect of sentiment on stock prices. For companies like BP PLC and BHP GROUP, significant peaks at positive lags in cross-correlation analysis correspond with strong VAR significance, suggesting that sentiment scores can indeed predict future stock price movements with a certain lead time. The VAR analysis demonstrates that 67% of significant relationships occur within the first week, with clustering at 2-day and 3-day lags. This temporal concentration suggests that sentiment effects are most pronounced in the immediate aftermath of social media activity, supporting theories of limited attention spans in financial markets.

This lag effect can be strategically leveraged for trading decisions, allowing investors to act on sentiment signals before they manifest in stock prices. However, this predictive power varied across companies, indicating that while the model is effective, its performance is influenced by company-specific factors and market dynamics. The company-specific nature of these lag structures indicates that a one-size-fits-all approach would be suboptimal, supporting the development of individualized models for each company.

The integration of VAR analysis with our SIMDM model represents a methodological advancement that addresses limitations in both traditional econometric approaches and machine learning methods. Traditional VAR models often focus on linear relationships and may miss non-linear sentiment effects, while machine learning approaches may identify spurious correlations without establishing causal relationships. Our hybrid approach leverages the strengths

of both methodologies: VAR analysis establishes statistical causality and optimal lag structures, while the SIMDM model captures non-linear relationships and provides practical trading signals. The preliminary stationarity testing ensures that both approaches rest on solid econometric foundations.

Furthermore, the heterogeneous optimal lag structures identified through VAR analysis can inform company-specific parameter tuning in our SIMDM model, potentially improving predictive accuracy through tailored lag specifications. This personalization approach represents a departure from traditional one-size-fits-all models and acknowledges the heterogeneous nature of information processing across different companies.

A significant challenge in our study was the potential limitations of continual pre-training. While continual pre-training on domain-specific corpora, such as financial news articles, enhances the model's performance in those areas, it can also introduce biases. These biases arise from the model's exposure to a limited scope of language and context, which may not fully align with the diverse and informal language used in social media posts. The VAR analysis helps validate that despite these potential biases, the underlying sentiment-return relationships remain statistically significant and economically meaningful. Continual pre-training can inadvertently cause the model to overfit the specific style and terminology of the pre-training data, making it less adaptable to different data sources. Future research should explore other paradigms and methods to mitigate these biases, such as incorporating more varied pre-training datasets and employing techniques that enhance the model's adaptability to new and diverse contexts.

A key limitation of this study is the inherent noise in social media data, which affects both VAR inference and SIMDM accuracy. Moreover, the focus on the energy sector constrains external validity, as sentiment–return dynamics appear to be company-specific. Extending the analysis to additional sectors would improve robustness and generalizability.

Additionally, incorporating more sophisticated sentiment analysis techniques, such as those accounting for nuanced sentiments and contextual dependencies, could further refine the model's accuracy. The integration of real-time data streams and the development of dynamic models capable of adapting to evolving market conditions would also be valuable enhancements. Time-varying VAR models could address structural break concerns by allowing parameters to evolve over time, potentially improving both statistical inference and practical performance.

This study advances the literature by demonstrating that domain-specific sentiment extracted with LLM models contains incremental, tradable information for equity returns in the energy sector. Combining a sentiment-informed market direction model with firm-level VAR and Granger tests, we document significant predictability for roughly 80% of firms and uncover heterogeneous transmission delays (2–14 days), a pattern consistent with gradual information diffusion and limits to attention. These results move beyond accuracy-only evaluations common in previous work by establishing econometric validity and linking topic-derived sentiment to the timing of price formation. The implications are twofold: methodologically, they show how the combination of modern NLP with classical identification clarifies the sentiment / return nexus; empirically, they provide actionable guidance for model design (data quality over sheer volume, firm-specific lag structures) and for timing decisions.

References

Alessia, D., Ferri, F., Grifoni, P., and Guzzo, T. 2018, International Journal of Computer Applications, 125

Araci, D. 2019, FinBERT: Financial Sentiment Analysis with Pre-trained Language Models

Baker, M. and Wurgler, J. 2006, Journal of Finance, 61, 1645

Bandopadhyaya, A. and Jones, H. 2006, Journal Name, Volume Number, Page Numbers

Brown, T. B., Mann, B., Ryder, N., et al. 2020, arXiv:2005.14165

Daniel, K., Hirshleifer, D., and Subrahmanyam, A. 1998, Journal of finance, 53, 1839

De Mol, C., Giannone, D., and Reichlin, L. 2009, Journal of Econometrics, 146, 318

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2018a, arXiv:1810.04805

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2018b, arXiv:1810.04805

Ding, L., Chen, Z., Wang, X., and Yin, W. 2024, Efficient Algorithms for Sum-of-Minimum Optimization

Fama, E. 1970, Journal of Finance, 25, 383

Fama, E. 1998, Journal of Financial Economics, 49, 283

Huang, A. H., Wang, H., and Yang, Y. 2023, Contemporary Accounting Research, 40, 806

Hutto, C. and Gilbert, E. 2014, in Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14), 216–225

J Bollen, H. M. and Zeng, X. 2011, Journal of Computational Science, 2, 1

Jeffrey Pennington, Richard Socher, C. D. M. 2014, :10.3115/v1/D14-1162

JR Piñeiro-Chousa, M. L.-C. and Pérez-Pico. 2016, Journal of Business Research, 69, 2087

Kim, S. and Kim, D. 2014, Journal of Economic Behavior Organization, 107, 708–729

Liu, B. 2020, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions (Cambridge University Press)

Malo, P., Sinha, A., Takala, P., Korhonen, P., and Wallenius, J. 2013, arXiv:1307.5336
McInnes, L., Healy, J., and Melville, J. 2020, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction
Medeiros, M. C. and Mendes, E. F. 2016, Journal of Econometrics, 191, 312
Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013, arXiv:1301.3781
N Barberis, A. S. and Vishny, R. 1998, Journal of Financial Economics, 49, 307
P Koukaras, C. N. and Tjortjis, C. 2022, Telecom, 3, 358–378
Qiu, J. and Welch, I. 2004, Journal of Empirical Finance, 11, pp. 427
R Ren, D. W. and Liu, T. 2019, IEEE Systems Journal, 13, 760–770
Ramos, J. 2003, Using Tf-idf to Determine Word Relevance in Document Queries, Tech. rep., Department of Computer Science, Rutgers University
Reimers, N. and Gurevych, I. 2019, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks
Reuters, T. 2009, Thomson Reuters Text Research Collection (TRC2), Web download, available from NIST upon request. Retrieved June 17, 2024, from https://trec.nist.gov/data/reuters/reuters.html
Ribeiro, F. B., Araújo, M., Gonçalves, P., Benevenuto, F., and Gonçalves, M. A. 2015, in Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI-15), 2903–2909
S Zaman, U. Y. and Saleem, T. 2022, Global Knowledge, Memory and Communication,)
Schumaker, R. P. and Chen, H. 2012, Decision Support Systems, 53, 458
Sun, C., Huang, L., and Qiu, X. 2019, arXiv preprint arXiv:1903.09588,)
T Sprenger, A Tumasjan, P. S. and Welpe, I. 2014, European Financial Management, 20, 926
Vaswani, A., Shazeer, N., Parmar, N., et al. 2023, arXiv:1706.03762
W Zhang, Z Deng, X. C. and Yu, W. 2018, Decision Support Systems, 114, 47
Wu, S., Irsoy, O., Lu, S., et al. 2023, arXiv preprint arXiv:2303.17564, in press (DOI: 10.48550/arXiv.2303.17564)
Xie, Q., Han, W., Chen, Z., et al. 2024, in Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Datasets and Benchmarks Track
Xie, Q., Han, W., Zhang, X., et al. 2023, arXiv preprint arXiv:2306.05443, in press (DOI: 10.48550/arXiv.2306.05443)
Z Da, J. E. and Gao, P. 2015, Review of Financial Studies, 28, 1–32

A. Appendix

Algorithm Description

A.1. Webscraping

Overview This document presents a concise algorithmic description of a Twitter scraper designed to extract tweets based on specific search criteria. The scraper operates through a series of steps, including initialization, data extraction, and data storage, leveraging a web automation tool (Selenium) for interaction with Twitter's web interface.

TwitterScrapper Class

Algorithm 1 Initialize the Twitter Scrapper

Input: research, link, output_path, start_date, end_date, mail, username, password, verbose
Output: An instance of the TwitterScrapper class
Initialize instance variables with input parameters

Algorithm 2 Extract Data from a Tweet Card

Input: card (HTML element representing a tweet)
Output: Extracted tweet data (username, handle, postdate, text, reply count, retweet count, like count)
Extract handle from the card element
Extract username from the card element
Extract postdate from the card element
Extract comment text from the card element
Extract reply count from the card element
Extract retweet count from the card element
Extract like count from the card element
return Extracted data as a tuple (username, handle, postdate, text, reply count, retweet count, like count)

Algorithm 3 Set Up and Open Chrome Browser

Output: Configured Chrome driver instance
Configure Chrome options for incognito mode
Open Chrome and navigate to Twitter login page
return Chrome driver instance

Algorithm 4 Perform Advanced Search

Input: driver, search parameters (research, start_date, end_date)
Construct search URL with parameters
Navigate driver to the constructed URL

Algorithm 5 Scroll and Extract Tweets

Input: driver
Output: Collected tweets
Initialize last position as None
Initialize end of scroll region as False
while not end of scroll region **do**
    Scroll the page
    Collect visible tweets
    Save tweets if not previously saved
    Update last position and end of scroll region status
end while
return Collected tweets

Algorithm 6 Save Tweet Data to CSV

Input: records (list of tweet data), output_path
Output: CSV file containing the tweet data
Write records to CSV at the specified output path

Algorithm 7 Main Scraping Process

Initialize the TwitterScrapper instance
Open and set up Chrome
Perform login and handle pop-ups if necessary
Perform advanced search
Scroll through the search results, extracting and saving data
Close the browser

Algorithm 8 Launch Scraper for Multiple Companies

Define list of companies
for each company in the list **do**
    Set output path based on company name
    Create and execute a TwitterScrapper instance
end for

A.2. BERT : Pre-trained of Deep Bidirectional Model for Language Understanding

BERT, developed by Devlin et al. Devlin et al. (2018b) at Google in 2019, has significantly advanced state-of-the-art benchmarks in natural language processing. BERT uses L layers (Transformer blocks), with a hidden size of H and A self-attention heads. Two principal versions has been developed:

- **BERT$_{\text{BASE}}$**: $L = 12$, $H = 768$, $A = 12$, Total Parameters=110M
- **BERT$_{\text{LARGE}}$**: $L = 24$, $H = 1024$, $A = 16$, Total Parameters=340M

Input Representation

Input representation is a sum of token embeddings, segment embeddings, and position embeddings. Subsequently, for a given input sequence $x = [x_1, x_2, \ldots, x_n]$, the embeddings sequence is constructed as follows:

$$E_{\text{input}} = E_{\text{token}} + E_{\text{segment}} + E_{\text{position}} \tag{14}$$

Pre-training Tasks

Masked Language Model (MLM)

- Randomly masks 15% of the tokens in the input.
- The model predicts the masked tokens based on their context.
- Loss function: Cross-entropy loss over the masked tokens.

$$L_{\text{MLM}} = -\sum_{i \in \text{masked tokens}} \log P(t_i \mid \text{context}) \tag{15}$$

Next Sentence Prediction (NSP)

- Predicts if a given sentence B follows sentence A in the original text.
- Loss function: Cross-entropy loss for binary classification (IsNext vs. NotNext).

$$L_{\text{NSP}} = -\left[y \log \hat{y} + (1 - y) \log(1 - \hat{y})\right] \tag{16}$$

Self-Attention Mechanism

The core of the BERT model is its bidirectional encoder architecture, which utilizes a multi-head self-attention mechanism.

This design allows the model to consider the context from both directions (left and right) for each token in the input sequence. In each layer of the model, the self-attention mechanism computes the following:

$$\begin{aligned} Q &= H^{(l)} W_Q \\ K &= H^{(l)} W_K \\ V &= H^{(l)} W_V \end{aligned} \tag{17}$$

Here, Q (queries), K (keys), and V (values) are obtained by projecting the input representation $H^{(l)}$ using weight matrices W_Q, W_K, and W_V, respectively.

BERT uses scaled dot-product attention. For each head, the attention scores are computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{18}$$

In this equation, the dot product of Q and the transpose of K is scaled by the square root of the dimensionality of the key vectors, $\sqrt{d_k}$, and subsequently passed through a softmax function to obtain the attention weights. These weights are then used to compute a weighted sum of the values V, thereby producing the attention output.

Following the computation of the attention scores, the outputs from the multi-head attention mechanism are concatenated and linearly transformed. Specifically, the multi-head attention mechanism is defined as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{h}_1, \text{h}_2, \ldots, \text{h}_h)W_O$$

where each attention head is computed as:

$$\text{h}_i = \text{Attention}(QW_{Q_i}, KW_{K_i}, VW_{V_i})$$

Here, h is the number of attention heads, and W_{Q_i}, W_{K_i}, W_{V_i}, and W_O are the learned projection matrices for the i-th head and the output projection, respectively.

After the multi-head attention layer, the output undergoes a series of transformations:

1. Add & Norm: The output from the multi-head attention layer is added to the original input (residual connection) and normalized:

$$H^{(l)'} = \text{LayerNorm}(H^{(l)} + \text{MultiHead}(Q, K, V))$$

2. Feed-Forward Neural Network: The normalized output is then passed through a position-wise fully connected feed-forward network, which consists of two linear transformations with a ReLU activation in between:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Here, W_1, W_2, b_1, and b_2 are learned parameters. The output of the feed-forward network is then added to the input of the feed-forward network (residual connection) and normalized:

$$H^{(l+1)} = \text{LayerNorm}(H^{(l)'} + \text{FFN}(H^{(l)'}))$$

This process is repeated for each layer in the BERT model, with the output of each layer serving as the input to the next.

The final hidden states from the last layer of the BERT model can be used for various downstream tasks. For instance, for classification tasks, the hidden state corresponding to the [CLS] token is typically used. This hidden state is passed through a classification layer:

$$y = \text{softmax}(H_{[\text{CLS}]}W_C + b_C)$$

where W_C and b_C are learned parameters, and y is the predicted output.

Training Objective

The combined loss for pre-training is the sum of MLM and NSP losses:

$$L = L_{\text{MLM}} + L_{\text{NSP}} \tag{19}$$