Time series analysis model

1.Data preparation

Using a topic classification model, the top 25 news items are categorized into 15 different topics. We count each theme for the day and remove columns other than count, date, and label. Due to the label column in the dataset being the Dow Jones Industrial Average (DJIA) for the current day, it is clearly meaningless for our prediction target. So we move the data in the Label column up one unit so that the Label column is a prediction for the next day. And the last column will be deleted due to the lack of predictable labels.

2.Granger Causal Relation Test

2.1 Basic principles

The Granger causality test, proposed by Clive Granger, is used to determine causal relationships between time series variables. The basic assumption of Granger causality test is that if time series X has predictive ability for the future values of time series Y, then X is called the Granger cause of Y. This causal relationship does not imply a direct causal connection, but rather refers to the information contained in X that helps improve the accuracy of predicting Y. In our experiment, Y represents the rise and fall of DJIA (Label), while X represents the count of each news topic.

Granger's mathematical expression takes into account the following linear model form:

Autoregression :

Union-regression :

Where ,represent intercept terms; p,q represents the maximum lag order of Y and X respectively; , represent the parameters to be estimated; , represents the error term.

2.2 Inspection standards

By comparing the sum of squared residuals (RSS) of these two models, we further use an F-test to determine whether adding the lagged term of X significantly reduces the error variance.

The formula for F-test statistic is as follows:

is the sum of squared residuals of a model without X; is the sum of squared residuals of a model containing X; T is the sample size; K is the total number of parameters in the model containing X; q is the lag order of X.

If the calculated F value is greater than the critical value, reject the null hypothesis and accept the alternative hypothesis of Granger causality between X and Y. In addition, we also referenced the p-value of the results through the built-in functions of the statsmodels library. The P-value is calculated based on the F-statistic and its degrees of freedom, reflecting the probability of observing the current or more extreme result under the null hypothesis. If the P-value is less than the selected significance level (such as 0.1), the null hypothesis can be rejected and considered as the Granger cause of Y.

3. Sentiment Analysis Framework

The objective of this framework is to systematically quantify the sentiment polarity (positive/neutral/negative) and intensity associated with each financial news headline. This enables a granular examination of the interaction between emotional tone and news categorization in influencing market movements. By leveraging domain-specific natural language processing (NLP) models, the framework aims to enhance the accuracy of sentiment detection in financial contexts, thereby supporting the downstream task of assessing the correlation between news sentiment and stock price fluctuations.

3.1 Model Selection and Customization

Model Architecture  
To address the challenges inherent in financial sentiment analysis, this study employs FinBERT, a transformer-based language model pre-trained on financial texts and fine-tuned on corpora such as earnings call transcripts, SEC filings, and Reuters financial news. FinBERT has demonstrated superior performance in interpreting domain-specific language compared to general-purpose sentiment models.

Model Adaptations  
To optimize FinBERT for headline-level sentiment analysis, the following customizations are introduced:

Input Truncation: Headlines are truncated or padded to a fixed length of 64 tokens, ensuring compatibility with FinBERT’s input limitations while preserving computational efficiency.

Entity-Aware Attention Masking: Named Entity Recognition (NER) is performed using SpaCy to identify and prioritize financial entities such as company names, stock indices, and monetary references. Attention masks are adjusted to emphasize these entities, enabling the model to focus on sentiment-bearing components of the headline.

3.2 Sentiment Quantification

Polarity Classification  
Each headline is assigned a discrete sentiment polarity based on model predictions:

Positive (+1): Indicates news associated with corporate growth, technological innovation, regulatory approval, or strategic alliances.  
*Example*: “Microsoft acquires AI startup to boost cloud capabilities.”

Neutral (0): Represents factual or balanced news with minimal directional bias.  
*Example*: “Federal Reserve maintains benchmark interest rate.”

Negative (−1): Reflects news related to financial risks, losses, legal actions, or operational disruptions.  
*Example*: “Boeing halts 737 Max production due to safety concerns.”

Intensity Scoring  
To capture the strength of sentiment, a continuous Sentiment Intensity Score s∈[−1,+1] is computed for each headline as follows:

**s=P(Positive)−P(Negative)**

Where P(Positive) and P(Negative) represent the softmax-normalized probabilities output by FinBERT for the positive and negative classes, respectively. This scoring scheme ensures comparability across headlines by standardizing outputs on a common scale. A score closer to +1 or −1 indicates higher sentiment certainty, while values near zero suggest ambiguity or neutrality.

3.3 Implementation Pipeline

Batch Processing  
Headlines are processed in mini-batches of 32 items to maximize GPU throughput and reduce

inference latency. Intermediate model outputs are cached to prevent redundant computations for duplicate headlines, thereby improving overall pipeline efficiency.

Output Integration  
Each news item is enriched with two additional columns:

Sentiment\_Polarity: A discrete variable taking values +1, 0, or −1, reflecting the headlines’ emotional direction.

Sentiment\_Intensity: A continuous score s∈[−1,+1], reflecting the magnitude of sentiment.

These outputs serve as key features in subsequent correlation analysis between news sentiment and stock price movements, facilitating both qualitative and quantitative evaluations.

Below is an expanded and refined version of the "Constructing Category-Sentiment Joint Features" section, formatted in an academic style suitable for a research paper.

4. Constructing Category-Sentiment Joint Features

Objective

The primary aim of this section is to develop interaction terms that combine news categories (e.g., Technology, Politics) with sentiment polarity. These joint features are designed to capture asymmetric effects—recognizing that the impact of sentiment on market dynamics may differ across various news categories. By integrating both categorical and sentiment information, the model can better assess how specific types of news influence market movements.

4.1 Feature Engineering

To quantify the combined effect of news category and sentiment polarity, we aggregate daily sentiment scores using a weighted summation. For a given news category cc and sentiment polarity ss (where ss can be either Positive or Negative), we compute an aggregated score for each day tt as follows:

文本

AI 生成的内容可能不正确。

where:

I(⋅)is the indicator function, which equals 1 if the condition is true and 0 otherwise.

ci and si represent the category and sentiment polarity of the ii-th headline, respectively.

si,scoresis the sentiment intensity of headline ii (as derived from the sentiment analysis framework).

ranki indicates the ranking of the headline, with Top1 =1, Top2 =2, …, Top25= 25, this serves as a decay factor, giving higher weight to more prominent (i.e., top-ranked) news items.

NN denotes the total number of headlines processed for a given day (in our example, N=25).

The use of the decay function ranki ensures that headlines deemed more important by their ranking contribute proportionately more to the aggregated score, thereby reflecting their presumed greater influence on market sentiment.

4.2 Feature Matrix Example

The computed aggregated scores for each category-sentiment pair are organized into a feature matrix. Each row of the matrix corresponds to a specific date, while each column represents a unique combination of news category and sentiment polarity. An illustrative example of the feature matrix is provided below:

| Date | Technology\_Positive | Technology\_Negative | Politics\_Positive | Politics\_Negative | ... |
| --- | --- | --- | --- | --- | --- |
| 2023-01-01 | 0.85 | -0.12 | 0.31 | -0.64 | ... |
| 2023-01-02 | 0.72 | -0.91 | 0.05 | -0.23 | ... |

In this example, each entry Sc,s,t reflects the cumulative effect of both the sentiment intensity and the importance (as determined by ranking) of news headlines within a specific category and polarity for the day.

4.3 Validation Steps

To ensure the robustness and relevance of the constructed joint features, several validation steps are undertaken:

Decay Function Sensitivity:  
The choice of the decay function is critical. While the current formulation uses to assign

weights, alternative weighting schemes—such as will be evaluated. This sensitivity analysis helps determine whether the selected decay function appropriately captures the influence of headline ranking on the aggregated score.

Nonlinear Effects:  
The relationship between sentiment and market impact may not be strictly linear. To capture

potential nonlinear dynamics, squared terms are introduced. These additional terms allow the model to account for diminishing or accelerating effects of sentiment, thereby improving the flexibility and explanatory power of the feature set.

Cross-Validation and Performance Metrics:  
To further validate the effectiveness of these joint features, cross-validation techniques will be applied. Performance metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R2 will be used to evaluate the contribution of these features to predictive accuracy in subsequent econometric or machine learning analyses.

By rigorously testing different decay functions and incorporating nonlinear effects, this framework seeks to ensure that the constructed features robustly capture the nuanced interplay between news categories and sentiment polarity.