
Guba Emotion

CONTENTS

01

Introduction

Background: The Uniqueness and Challenges of China's A-Share Market

The Dominance of Retail Investors in the A-Share Market

- China's A-share market plays a **key** role in global finance but differs notably from mature markets in structure and behaviour.
- A defining trait is its retail investor dominance—despite a rise in institutional participation, individuals still made up about 55% of trading in 2019.[1]
- This retail-dominated investor structure drives high volatility and turnover, as individual investors are more prone to emotions, rumours, and short-term speculation, leading to frequent price deviations from intrinsic value.

Investor Sentiment and Asset Pricing

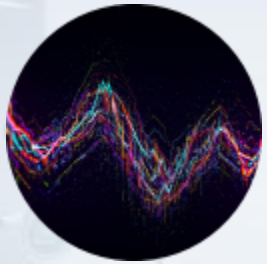
- This market environment provides an ideal setting for studying the impact of investor sentiment on asset pricing—a "quasi-natural experiment."
- Given participants' strong sensitivity to information, sentiment shifts are quickly and sharply reflected in prices, creating a complex nonlinear relationship.
- As a result, traditional forecasting models grounded in the Efficient Market Hypothesis (EMH) often fail to capture fluctuations driven by behavioural biases, underscoring the value of incorporating alternative data—especially social media signals that directly reveal investor sentiment.

[1] Luo C, Li Z, Liu L. Does investor sentiment affect stock pricing? Evidence from seasoned equity offerings in China[J]. National Accounting Review, 2021, 3(1): 115-136.
[2] Li J, Bu H, Wu J. Sentiment-aware stock market prediction: A deep learning method[C]//2017 international conference on service systems and service management. IEEE, 2017: 1-6.

Current Challenges: Traditional Methods

Statistical Models

Statistical models, represented by ARIMA and GARCH, assume at their core that a linear pattern exists in market price series. However, financial markets, especially the A-share market, are full of nonlinear and non-stationary characteristics, making it difficult for these models to effectively capture these patterns.[1]



Early Technical Indicators

With technological advancements, machine learning models using technical indicators (such as the MACD and the RSI) as features have become increasingly popular. However, these models are essentially reprocessing price and volume data, ignoring the core element driving short-term price discovery—human emotions and opinions.

[1] Jiang T, Zeng A. Financial sentiment analysis using FinBERT with application in predicting stock movement[J]. arXiv preprint arXiv:2306.02136, 2023.

Core Research Topics

Research Goal

Construct a daily A-share price fluctuation prediction framework that deeply integrates financial social media signals with traditional market data to address existing challenges.

Contribution 1: BERT Sentiment Analysis

Fine-tune the model on "stock forum" text (Eastmoney Stock Forum) to capture language paradigms, slang, and sentiment expressions in the informal Chinese financial context, achieving precise quantification of investor sentiment.

Contribution 2: Multimodal Predictor

Integrate traditional price/volume features with high-precision sentiment scores from the StockForum-BERT model and deep semantic features extracted from text reflecting discussion quality and hot topics.

Performance Improvement Proof

1. Advanced model tailored to a specific domain and data source; 2. Carefully designed fusion factor that reflects the microstructure of the market.

02

Related Work

Investor Sentiment and the Stock Market: Theory and Empirical Evidence

- Theoretical Controversy in Asset Pricing
  - There are two main academic camps regarding the role of investor sentiment in asset pricing. The Efficient Market Hypothesis (EMH) holds that all publicly available information is quickly reflected in stock prices, and irrational factors like sentiment have no lasting impact on the market.[1] In contrast, behavioral finance theory points out that under arbitrage constraints and cognitive biases, investor sentiment is a significant driver causing asset prices to deviate from their fundamental value and creating market anomalies.[2]
- Empirical Research Support for Behavioral Finance
  - A wealth of empirical research supports behavioral finance. The academic community has confirmed a significant correlation between social media public sentiment and stock market returns, volatility, and trading volume, with studies using international platforms such as Twitter[3] and StockTwits[4]. For the Chinese market, investor sentiment from Eastmoney Stock Forum has significant positive predictive power on the returns and trading volume of CSI 300 Index constituent stocks[2], and this predictive power remains robust after controlling for traditional risk factors like book-to-market ratio and beta coefficient, providing a basis for selecting "Stock Forum" as the core data source.

[1] Jiang T, Zeng A. Financial sentiment analysis using FinBERT with application in predicting stock movement[J]. arXiv preprint arXiv:2306.02136, 2023.
[2] Wang G, Yu G, Shen X. The effect of online investor sentiment on stock movements: An LSTM approach[J]. Complexity, 2020, 2020(1): 4754025.
[3] Lee J, Youn H L, Poon J, et al. Stockemotions: Discover investor emotions for financial sentiment analysis and multivariate time series[J]. arXiv preprint arXiv:2301.09279, 2023.
[4] Upadhyay A, Jain H, Dhingra P, et al. Stock Market Prediction Using Social Media Sentiments[C]//International Conference on Complex, Intelligent, and Software Intensive Systems. Cham: Springer Nature Switzerland, 2024: 14-26.

Evolution of Financial Text Sentiment Analysis Technology and BERT Application

Phase One: Dictionary and Traditional Machine Learning

Relied on pre-built financial sentiment dictionaries for positive/negative word matching; traditional machine learning models like SVM and Naive Bayes were widely used. Limitations included lack of context understanding and difficulty handling complex linguistic phenomena (negation, contrast, irony).

Phase Two: Rise of Deep Learning Models

With increased computing power, deep learning models (RNNs, LSTMs, CNNs) dominated. They automatically learned feature representations from text sequences, significantly improving sentiment classification accuracy.

Phase Three: Transformer Revolution and BERT Application

Transformer models based on self-attention mechanisms emerged; BERT (Bidirectional Encoder Representations from Transformers) achieved unprecedented contextual understanding via bidirectional encoding, reaching state-of-the-art performance in sentiment analysis and other NLP tasks.

Evolution of Financial Text Sentiment Analysis Technology and BERT Application

Fine-tuning BERT Models in the Financial Sector

BERT models in finance fall into general-purpose and domain-specific categories (e.g., FinBERT, pre-trained on Reuters news/corporate reports). Both require domain-specific data fine-tuning for optimal downstream task performance. Academic consensus shows fine-tuning on task-specific labeled data enables learning of unique domain language patterns and sentiment expressions, significantly boosting accuracy—e.g., fine-tuning BERT/FinBERT on FiQA/Financial PhraseBank improves results.[1]

[1] Nasiopoulos D K, Roumeliotis K I, Sakas D P, et al. Financial Sentiment Analysis and Classification: A Comparative Study of Fine-Tuned Deep Learning Models[J]. International Journal of Financial Studies, 2025, 13(2): 75.

Construction of Predictive Factors through Multimodal Data

Sentiment Information Enhances Stock Prediction Model Performance

Combining sentiment information with traditional market data is an effective strategy to improve the performance of stock prediction models. Common methods use sentiment scores together with technical indicators like RSI and MACD as input features. Hybrid models integrating market data and sentiment information generally outperform single-source models.[1]

[1] Fu K, Zhang Y. Incorporating Multi-Source Market Sentiment and Price Data for Stock Price Prediction[J]. Mathematics, 2024, 12(10): 1572.
[2] Lee J, Youn H L, Poon J, et al. Stockemotions: Discover investor emotions for financial sentiment analysis and multivariate time series[J]. arXiv preprint arXiv:2301.09279, 2023.

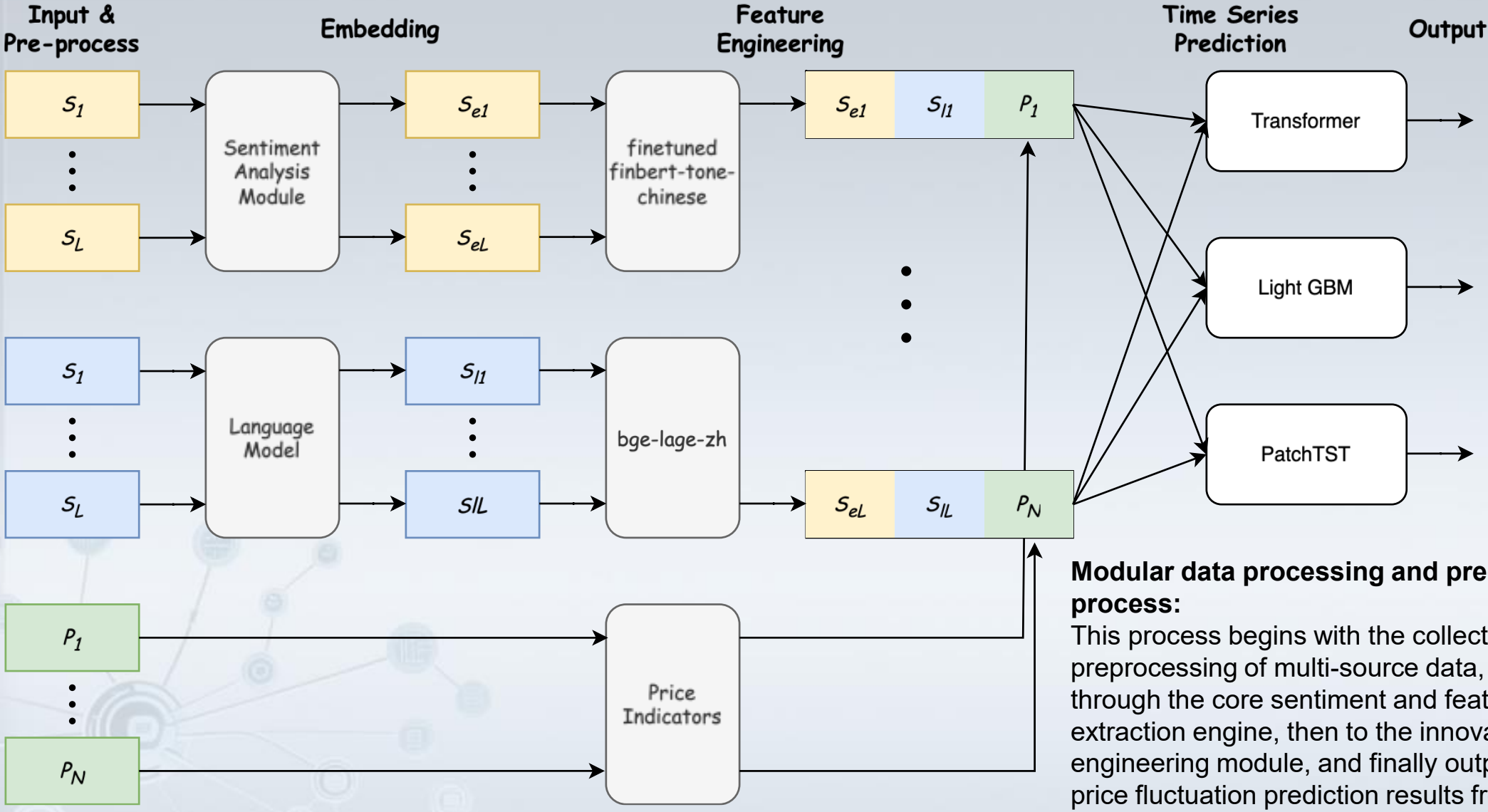
Sentiment & Price Sequential Modelling

Daily sentiment scores are treated as a new time series and input into time series prediction models (e.g., Transformer) along with technical indicator time series. For example, the stock emotions model[2] incorporates the numerical modality from the price index and the contextual modality from sentence embedding using BERT as well as emotion embedding using GloVe, then fed into the Temporal Attention LSTM. This research constructs a multi-dimensional feature system covering price, sentiment, and text semantics to provide richer input signals.

03

Method

Pipeline Outline:



Modular data processing and prediction process:

This process begins with the collection and preprocessing of multi-source data, proceeds through the core sentiment and feature extraction engine, then to the innovative factor engineering module, and finally outputs daily price fluctuation prediction results from the downstream time-series prediction model.

Pipeline Outline:

- **Data Input Layer:** Parallel acquisition of **text data (post titles, comments)** from the Eastmoney stock forum and **A-share market data** (opening price, highest price, lowest price, closing price, trading volume)
- **Data Preprocessing Layer**: Cleaning, aligning, and standardizing the raw data to prepare for subsequent model processing.
- **Sentiment Analysis Engine**: This is the core NLP module of this study. It adopts the Finbert-Tone-Chinese model and fine-tunes it on the specific domain of the "stock forum" corpus to accurately extract text sentiment.
- **Feature Engineering Module**: This module receives information from both market data and sentiment analysis results, constructing a series of price technical indicators and quantitative sentiment indicators.
- **Time Series Prediction Module:** Employs models to learn complex patterns in historical factor sequences.
- **Prediction Output Layer**: Outputs a binary classification probability of the direction of the closing price increase or decrease on the next trading day (T+1).

Data Acquisition and Preprocessing: Data Sources

Text Data

Using existing “Guba” dataset that only has post titles, we massively crawled all related comments as supplementary data on the Eastmoney Stock Forum “Guba” regarding the constituent stocks of the CSI 300 indices. We chose "Guba" as the data source because it is China's largest and most influential stock forum, providing the best window for observing retail investor sentiment.[1]

Market Data

Retrieves daily opening price, highest price, lowest price, closing price,volume (OHLCV) and turnover data for the same period and stock range as the text data.

[1] Wang G, Yu G, Shen X. The effect of online investor sentiment on stock movements: An LSTM approach[J]. Complexity, 2020, 2020(1): 4754025.

Data Acquisition and Pre-processing: Pre-processing Workflow

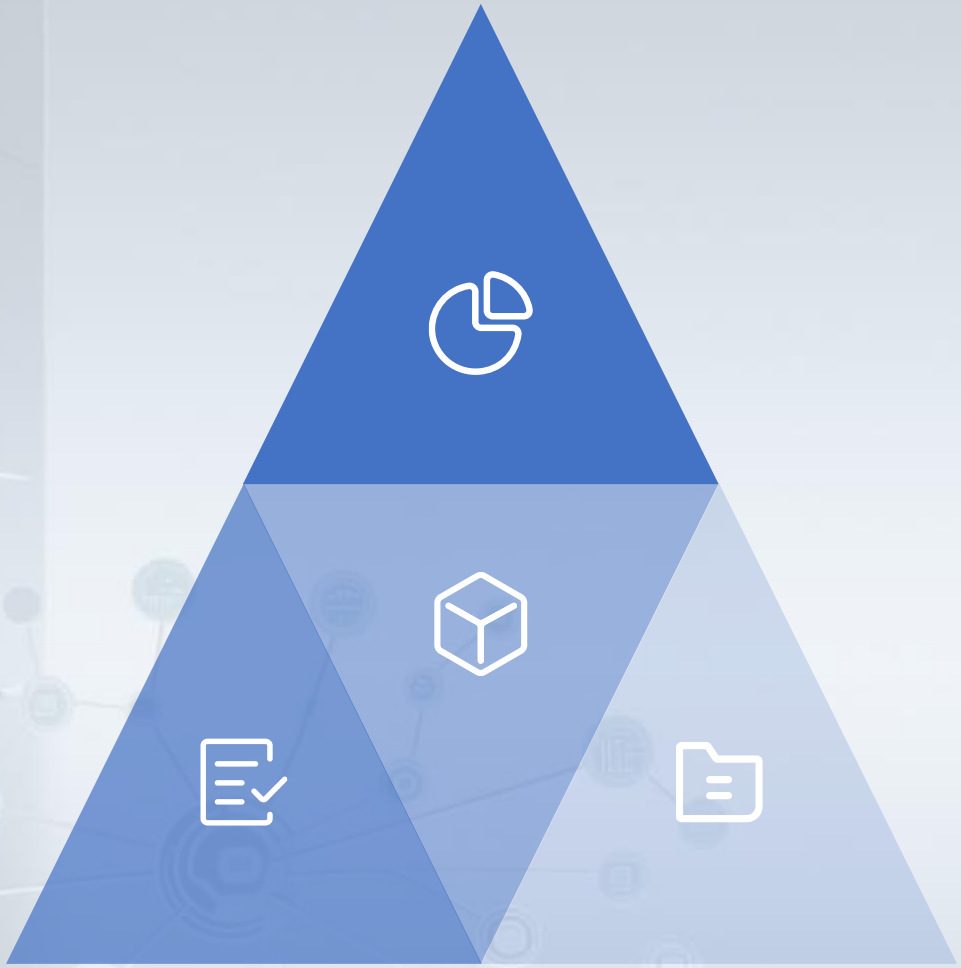
Text Cleaning and Merging

A systematic cleaning and filtering process was performed to remove common noise from stock forum texts, such as URLs, stock codes, meaningless special characters, and emoticons. To maximize contextual information, the title of each post and all its comments were merged into a single document; this approach helps to more accurately determine the overall sentiment.

Time Alignment

To ensure the timeliness of information, all text data are grouped to trading-time text and closing-time text. All posts and comments published for a specific stock within a single trading day's time window (e.g., from the close of trading on day T-1 to the close of trading on day T) are aggregated to predict stock price movements on day T+1.

Sentiment Analysis Model Based on Finbert-tone-chinese for Stock Forum



Module Objective

To create a high-precision sentiment classifier tailored to the unique Chinese financial social media context of "Stock Forum".

Base Model Selection

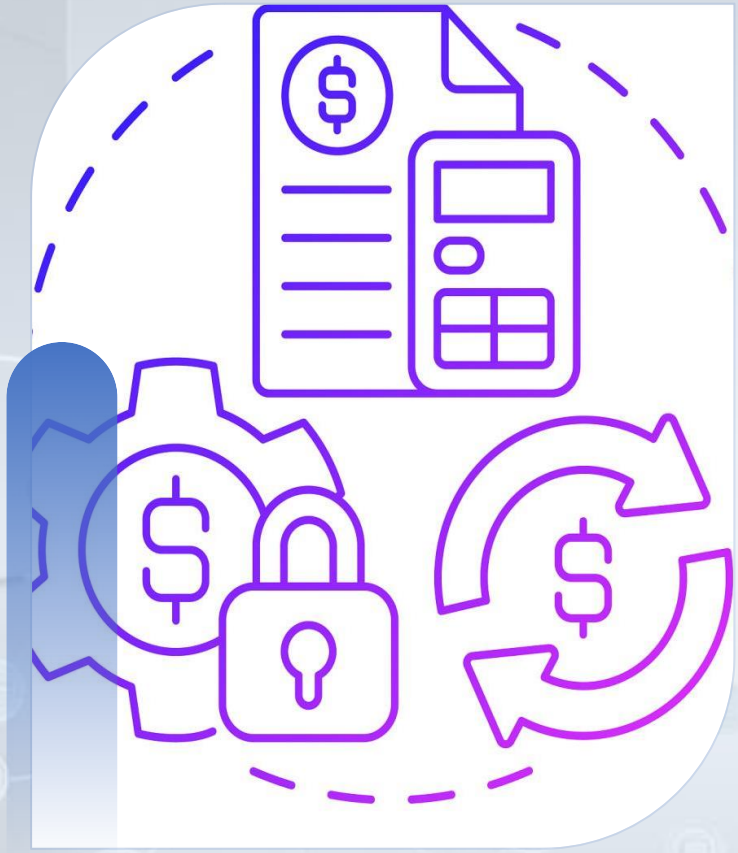
Selected yiyanghkust/finbert-tone-chinese, which is based on bert-base-chinese, fine-tuned on ~8000 analyst reports, and specifically designed for Chinese financial sentiment analysis with a strong domain foundation.

Fine-tuning for Stock Forum

Fine-tuned on 3000 manually annotated Stock Forum posts to adapt to the colloquial and informal language paradigm. The process involves adding a classification layer and optimizing with cross-entropy loss, resulting in a customized model capable of accurately judging sentiment (positive, negative, neutral) of Stock Forum posts.

Multimodal feature engineering

Price Technical Indicators in Multimodal Feature Engineering



Logarithmic Return

Widely used in financial time series analysis due to its favorable statistical properties (such as time additivity).

Amplitude

A direct indicator for measuring intraday price volatility.

Relative Strength Index (RSI)

A momentum indicator used to determine overbought or oversold conditions in the market.

Moving Average Convergence Divergence (MACD)

A classic indicator used to identify trend changes and momentum.

Bollinger Band Width

Used to measure the contraction and expansion of market volatility.

On-Balance Volume (OBV)

An indicator that combines price movements and trading volume to determine the strength of a market trend.

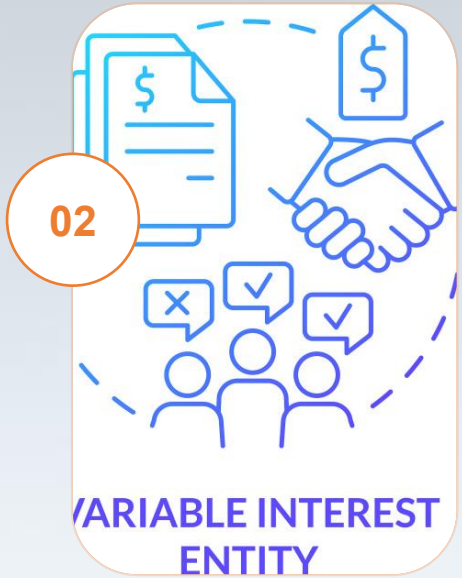
Multimodal feature engineering

Sentiment and Attention Indicators in Multimodal Feature Engineering

Sentiment Index Calculation Method

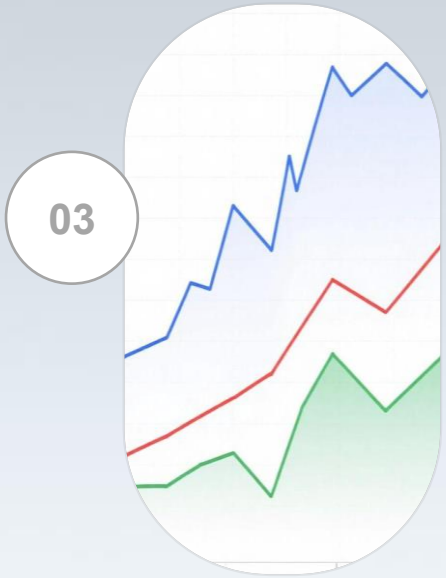
Emotion Index: Aiming to quantify the daily net sentiment tendency.

$$\frac{Total\_Positive - Total\_Negative}{Total\_Positive + Total\_Negative}$$



Investor Attention Indicators

Total Posts Count and Total Click Count serve as proxy variables for investor attention; changes in these indicators are significantly correlated with stock trading volume, volatility, and future returns.



Sentiment Momentum Indicator

Emotion Momentum is the 3-day or 5-day moving average of the sentiment index, aiming to capture short-term trend changes in investor sentiment, as sentiment may exhibit a momentum effect similar to price.

Downstream time series prediction model

Model Selection

This study evaluates three distinct architectures: **LightGBM (LGBM)**, a high-performance gradient boosting model for tabular data; the **Vanilla Transformer**, which uses self-attention for long-range dependencies; and **PatchTST**, an advanced Transformer variant that uses patching to model local and long-term time-series patterns.

General Model Strategy

We first train a single **General Model** on the aggregated dataset of all 230 stocks. This model aims to learn universal, market-wide patterns and serves as our baseline for generalized predictive accuracy.

Finetuned Model Strategy

Our second approach is a "General-to-Specific" strategy. We first pre-train the model on the full dataset, and then **finetune**this model on each individual stock. This adapts the model's broad market knowledge to unique, stock-specific dynamics.

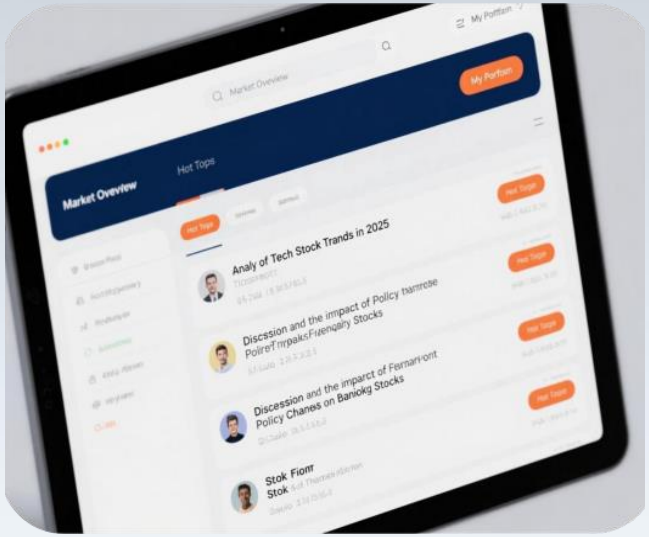
04

Experiments

Dataset Details

Stock Data:

110k daily stock data of 230 constituent stocks of the CSI 300 indices.



Text Data:

7.3M posts and 8M crawled comments from 230 stocks' forum

Time span:

January 1, 2023 to December 31, 2024

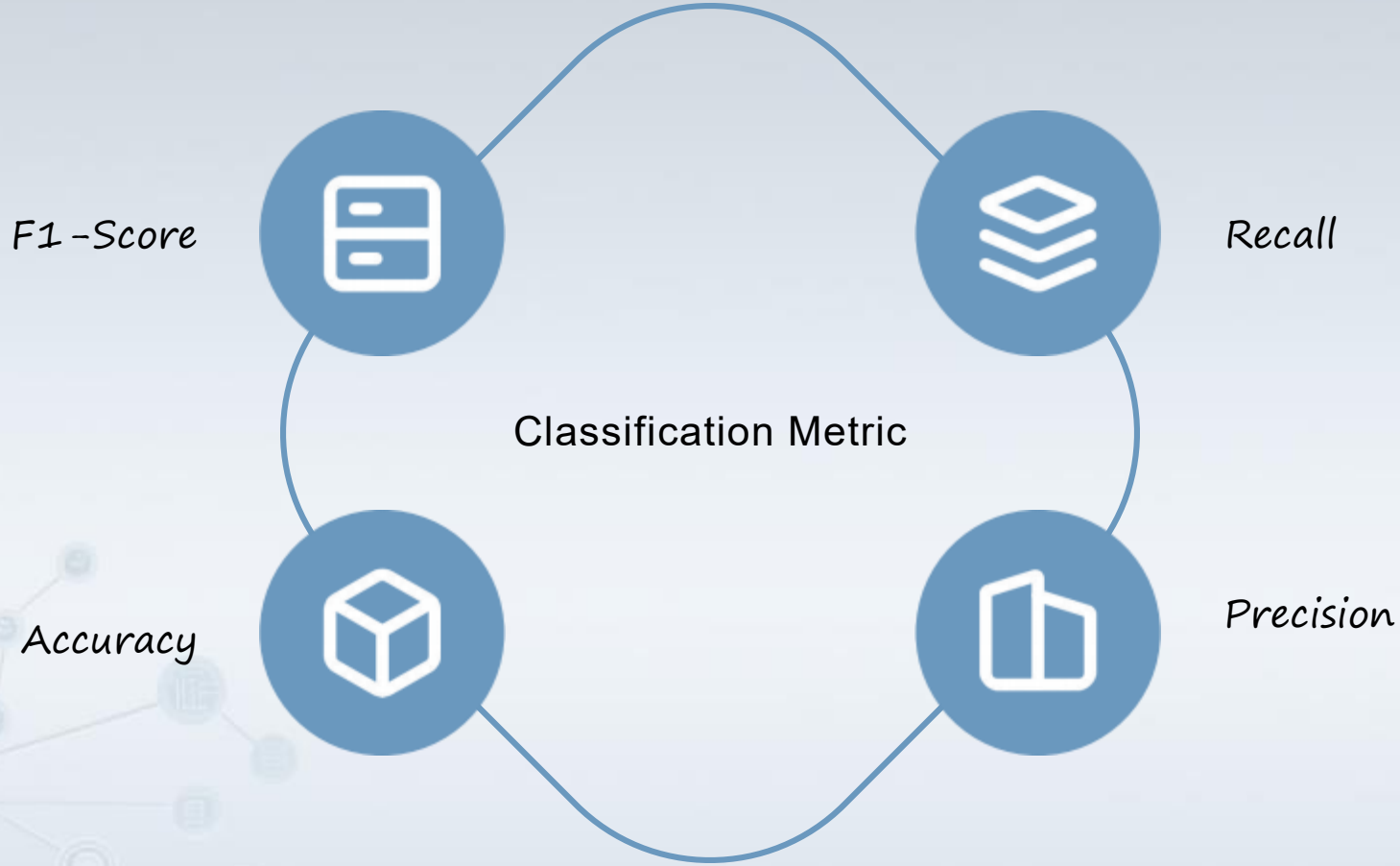
Sentiment Finetuning Details

neu_ori	neg_ori	pos_ori	content	neu	neg	pos
0.996906	0.001208	0.001886	PingAn Bank is my new favorite.	0.663775	0.107033	0.229192
0.986771	0.003113	0.010116	The main force is fleeing, starting to liquidate. You'll always make money if you're not greedy.	0.166202	0.548753	0.285046
0.951209	0.002636	0.046155	In a bull market, buying bank stocks is a sure win.	0.171968	0.344553	0.483478
0.983702	0.010840	0.005458	Selling and leaving, securing profits.	0.052462	0.931736	0.015802
0.973929	0.013783	0.012288	Last place, awesome.	0.256179	0.687294	0.056527
0.975063	0.010480	0.014457	How about a limit-up?	0.053515	0.049577	0.896908
0.952097	0.022017	0.025886	Trash PingAn.	0.129868	0.838603	0.031529

Evaluation Metric

To comprehensively evaluate the model's performance, this study employs an evaluation system that combines classification task indicators and empirical financial indicators.

Evaluation Metric (ML)



Evaluation Metric (Vanilla Transformer)

Metric	precision	recall	f1-score	support
0 (down)	0.42	0.69	0.52	3902
1 (unchanged)	0.32	0.34	0.33	2533
2 (up)	0.35	0.09	0.15	3670
accuracy			**0.39**	**10105**
macro avg	0.36	0.37	0.33	10105
weighted avg	0.37	0.39	0.34	10105

Evaluation Metric (Pretrained PatchTST with Addition PEFT)

Model – Pretrained PatchTST (IBM Granite: ibm-granite/granite-timeseries-patchtst), an encoder-only transformer tailored for time series with patch tokenization, where segment of the statements become tokens) and also shared weights across the attention layers.

Key Differences Compared to Vanilla Transformer – Vanilla treats each time step (or token) as a single token; PatchTST treats windows (patches) as tokens to capture local patterns and scale to longer horizons with lower compute, improving stability/generalization on time series.

Addition PEFT – Via Houlsby-style Adapter with backbone frozen, with various regularisers: Dropout, Weight Decay, Mix-Up, Gradient Clipping to minimize overfitting

Key Hyperparameters – (1) window_size = 30, (2) batch_size = 64, epochs = 100, learning_rate = 1e-5, weight_decay = 1e-3, dropout_rate = 0.3, grad_clip = 1.0

Evaluation Metric (Pretrained PatchTST with Addition PEFT)

Metric	precision	recall	f1-score	support
0 (down)	0.40	0.26	0.32	5405
1 (unchanged)	0.26	0.29	0.27	3152
2 (up)	0.37	0.49	0.42	4766
accuracy			**0.35**	**13323**
macro avg	0.35	0.35	0.34	13323
weighted avg	0.36	0.35	0.35	13323

Evaluation Metric (Pretrained PatchTST with Addition PEFT)T

The model's accuracy was further improved through fine-tuning on individual stock codes, resulting in an approximate 20% increase in accuracy.

Stock Code	Accuracy	Support	Weighted Precision	Weighted Recall	Weighted F1
00301	0.52	58	0.46	0.52	0.47
002460	0.52	58	0.49	0.52	0.49
603833	0.52	58	0.42	0.52	0.45
600023	0.50	58	0.52	0.50	0.49
00338	0.50	58	0.53	0.50	0.48
300433	0.48	58	0.54	0.48	0.49
601799	0.48	58	0.50	0.48	0.44
002252	0.47	57	0.46	0.47	0.45
300059	0.47	58	0.41	0.47	0.42
002179	0.47	58	0.57	0.47	0.43

Evaluation Metric (LGBM General Model)

Metric	precision	recall	f1-score	support
0 (down)	0.39	0.74	0.51	1851
1 (unchanged)	0.33	0.13	0.19	1371
2 (up)	0.36	0.17	0.23	1601
accuracy			**0.38**	**4823**
macro avg	0.36	0.35	0.31	4823
weighted avg	0.36	0.38	0.33	4823

Evaluation Metric
(LGBM Finetuned Model)

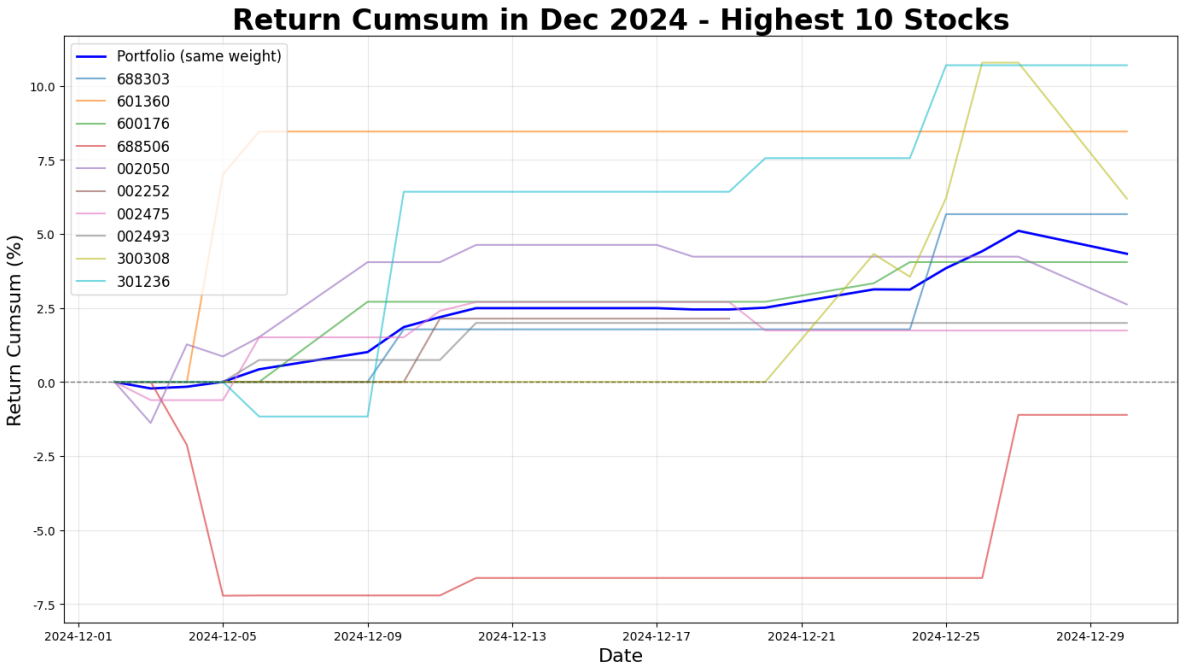
Stock Code	Accuracy	Support	Weighted Precision	Weighted Recall	Weighted F1
688303	0.71	21	0.81	0.71	0.65
601360	0.67	21	0.60	0.67	0.58
600176	0.62	21	0.80	0.62	0.54
688506	0.62	21	0.52	0.62	0.56
002050	0.57	21	0.46	0.57	0.51
002252	0.57	14	0.68	0.57	0.54
002475	0.57	21	0.51	0.57	0.53
002493	0.57	21	0.61	0.57	0.53
300308	0.57	21	0.52	0.57	0.53
301236	0.57	21	0.50	0.57	0.50

Evaluation Metric (Finance)

Model Prediction Strategy Evaluation

Based on the model's daily forecast signals (bullish/bearish), a long-short hedged investment portfolio is constructed for backtesting. The cumulative return, Sharpe ratio, and maximum drawdown of this strategy on the test set are reported to evaluate the model's economic value.

We choose 10 highest accuracy stocks in Dec 2024 and evaluate using the lgbm model.



Stock Code	Return	Sharpe Ratio	Maximum Drawdown
688303	5.66%	4.7987	0.00%
601360	8.45%	4.2397	0.00%
600176	4.04%	5.1440	0.00%
688506	-1.11%	-0.4932	-7.22%
002050	2.62%	2.0909	-1.92%
002252	2.14%	4.4028	0.00%
002475	1.73%	2.3471	-0.94%
002493	1.99%	4.9770	0.00%
300308	6.19%	2.6284	-4.14%
301236	10.69%	4.6099	-1.17%

05

Analysis

Comparison: LGBM and Transformer

Extreme Signal-to-Voice Ratio

- Financial sequences are dominated by **random, unpredictable "noise."**
- A Transformer, with its massive capacity, is designed to find complex patterns. When given a raw, noisy sequence, it **dramatically overfits the noise**, learning non-repeatable, false patterns.
- The LightGBM approach (using features) *first* reduces this noise via feature engineering (e.g., using moving *averages* which smooth data), making the problem easier.

Implicit vs. Explicit Signal Representation

- The **LightGBM** path relies on **Feature Engineering** (e.g., Momentum, RSI, Volatility). These features are *explicit* signals derived from financial domain knowledge. LGBM is state-of-the-art at selecting which of these strong, pre-processed signals are most predictive.
- The **Transformer** path gives the model a **raw sequence** and asks it to *implicitly learn* the concepts of "momentum" or "volatility" from scratch. This is a *much* harder task, and the model almost always gets lost in the noise (see Point 1) before it can discover these complex, subtle signals.

06

Conclusion

Conclusion

Proposed Prediction Framework

A daily price fluctuation prediction framework based on a fine-tuned Finbert-Tone-Chinese model and multimodal features is proposed and validated, targeting the retail investor-dominated and sentiment-driven characteristics of the Chinese A-share market.

Necessity of Domain Fine-Tuning

The Finbert-Tone-Chinese model, specifically fine-tuned for the Chinese financial social media context, significantly outperforms general models or unadapted financial models on sentiment classification tasks, confirming the necessity of domain adaptation in applying NLP to specific financial scenarios.

Comparison between Different Models

LightGBM's superior performance in daily financial forecasting stems from its robustness to the market's extremely low signal-to-noise ratio. Transformers, when given raw sequential data, tend to overfit this dominant noise and learn false patterns. The LightGBM approach, however, succeeds by using feature engineering to create explicit, denoised signals (like momentum). This transforms the task into a tabular problem that LightGBM is designed to solve, proving more reliable than a Transformer's struggle to implicitly find signals within raw noise.

Practical significance and application value

Implications for Quantitative Investing

The framework proposed in this study provides quantitative investment institutions with a specific methodology for mining excess returns (Alpha) in the Chinese market using alternative data (especially social media text).

Reference for Market Regulation

Regulatory agencies can draw on the indicators and framework proposed in this study to identify potential market manipulation or irrational fluctuations caused by rumors by monitoring indicators, thus providing early warning signals for maintaining market stability.

Future Work Outlook

Extending to higher frequency data

Apply the current framework to high-frequency data at the minute or even tick level to capture more granular sentiment changes and market reactions.

Exploring More Complex Model Architectures

Such as utilizes GNN etc. to model the co-occurrence relationships of stocks in online stock forums, exploring the effects of sector rotation and cross-stock sentiment transmission.

Deepening Causal Analysis

Moving from current predictive models towards more rigorous causal inference methodologies to more clearly identify the causal path between investor sentiment and stock price changes using finance knowledge.

Integrating information from multiple dimensions

The model incorporates characteristics of stock forum users (such as historical post accuracy and influence) and weights sentiment information from different sources to improve the signal-to-noise ratio.

Thanks for listening