Multi-Source Sentiment Integration with Machine Learning for A-Share Earnings Forecast Windows: A Proposed Implementation Framework

Chengde Huang
Business, South China Normal University, Shanwei, Guangdong, 516600, China
1353298989@qq.com

ABSTRACT

In China A-share market, to make Alpha generation, we should not only rely on traditional factor models, but also learn to use those wrong pricing driven by emotions during earnings forecast windows. At this time, there are many retail investors in the market and the information is asymmetric. We reviewed 42 intellectual studies and put forward a unified method framework. With the knowledge of behavioral finance, market microstructure and machine learning, we have built a four-layer implementation pipeline, which integrates three different emotional sources: textual disclosures, order flow and social media, and uses advanced ML architectures. These studies found that hybrid architectures performed better than the single method in the accuracy and risk index of direction prediction. These benefits are mainly concentrated in the high-concern forecast window with the largest emotional fluctuation. However, there are two restrictions in actual operation: the handling fee of 0.15-0.25% per transaction will reduce the total Sharpe ratio by 35-50%, so the minimum threshold must reach 1.5; Moreover, the performance of the strategy depends on the market situation-78% can make money in a bull market and only 45% in a bear market. This study provides practical suggestions for quantitative traders, including how to capture emotional signals more carefully, choose the appropriate machine learning algorithm according to different time, and how to reduce transaction costs and adapt to market changes.

Keywords: Multi-source sentiment; Machine learning; A-share market; Earnings forecast windows; Excess return strategies; Hybrid architectures; Transaction costs; Regime dependency

1. INTRODUCTION

Alpha generation has gone beyond the traditional fundamental and technical analysis. Behavioral finance tells us that investors' emotions will make the price deviate from its true value [1]. This effect is particularly obvious in a market like China's A-share market, where many ordinary retail investors participate, because emotions will directly affect the return of stocks [2]. Especially when the company releases earnings forecast, this effect will be even more severe [37]: before the announcement, everyone has no idea and their attention is focused on one point, which is easy to make the price wrong.

Previous studies have found that "unexpected" earnings forecasts (such as better or worse results than expected) can predict the stock price trend, but most studies use linear models and feel that all investors' reactions are similar. The research of Han and Shi [6] shows that the model with emotional factors can explain 23% of the stock price difference, while the traditional model can only explain 11%. However, they summarized the emotions into the overall market index, ignoring that during the forecast release period, emotions actually came from many different places. Similarly, Li and Hou [9] found that different opinions before the announcement can predict the abnormal return of the next day (IC = 0.08-0.12), but their text analysis only looked at the official announcement of the company and ignored the discussion on social media, which had a great impact on the market with many retail investors. Recently, there have been many new advances in machine learning-for example, gradient lifting methods such as XGBoost and LightGBM, cyclic networks such as LSTM, GRU and Bi-LSTM, and mixed structures that combine convolution with time models-which provide a framework for us to deal with these high-dimensional interactions [3, 8, 10, 16, 24]. These algorithms can capture the interaction between features and the time correlation, and we don't need to specify the specific form in advance, so that we can systematically use the wrong pricing opportunities driven by emotions during the forecast window.

Although sentiment analysis and machine learning have made progress, it is not good enough to combine them systematically to predict the time window of A shares. Existing research rarely uses order flow or social media data [4], while machine learning applications usually only use general price historical data, ignoring the abnormal fluctuations caused by financial report announcements. No practical model has been developed to take advantage of the wrong pricing in these short time windows, in which the trading noise of retail investors often overwhelms the fundamental information.

This paper aims to solve three research problems: first, how to extract emotional information from different sources (such as regulatory documents, market microstructure, social media) while maintaining the details of the signal? Second, which machine learning architecture is best at combining high-dimensional inputs-is the hybrid model (such as CNN-LSTM, XGBoost-LSTM stacking) more powerful than the single method? Third, how to turn the forecast into a practical strategy, taking into account the transaction cost (0.15%-0.25% per transaction), market impact and the influence of different market environments?

This study integrates all kinds of evidence into a clear framework and teaches us how to combine multi-source emotional signals with ML (machine learning) to earn excess returns during the A-share forecast window. The main contributions are as follows: First, organize the evidence into a set of method flow from raw data to execution, and explain how the transformer encoder in a professional field like FinBERT keeps the details of emotional signals [12]. Secondly, systematically evaluate different ML architectures and find out the conditions under which the hybrid model performs better than the single method. Thirdly, the problems in practical application are solved through agreements such as cost accounting, market status verification and execution timing.

Section 2 reviews the evolution of sentiment analysis, the application of machine learning in finance, and the microstructure of A-share forecast window. Section 3 introduces the systematic method of document identification and integration in detail, and ensures the transparent report of systematic review according to PRISMA 2020 guide [11]. Section 4 compares and evaluates the source of emotional data and the performance of the algorithm. Section 5 discusses the influence, limitation and boundary conditions in practical application. Section 6 summarizes the key findings and future research directions.

2. LITERATURE REVIEW

Capital market research has long begun to look at how the information in those accounting reports affects the stock price [38]. The earliest research was to guess the investor's mood by observing the stock price fluctuation and trading volume, or by watching mutual fund flows [41]. Tetlock [7] was the first to apply the method of analyzing the content of articles to finance. He found that the mood of the media can predict the market trend, and later directly measured the mood by counting the words in earnings reports and news [25]. The current method will combine company documents, social media and transaction order flow; There will be loopholes if only a single data source is used.

Text analysis will use tools like FinBERT to study earnings forecasts, News and broker reports, which can grasp the subtle differences in tone [5, 23, 26]. This method goes further than the early study of forward-looking disclosures by machine learning [27]. Market microstructure will analyze transaction data (such as unusual order flow and sudden increase of short-selling), especially in the market with many retail investors, there is an effect called' attention spillover' which is very important [32]. Social media, search trends and forums also provide alternative data [28]. For example, emotions on microblogging are very useful for predicting changes in the whole market.

Combining information from different sources can make the prediction result much more accurate. Ramaraju and others [4] found that if text information, order flow and social media data are used together, the accuracy can be improved by 12% to 18%. Xu's research team [14] also found that the model with multiple information can explain 16% of the price changes, while only words can explain 9%. Wang et al. [16] used dual-CNN-LSTM model, and the prediction accuracy reached 88.24%, which was 5.55% higher than that of LSTM model only and 9.81% higher than XGBoost. Liu's method [10] (using LSTM, GRU, Bi-LSTM models and XGBoost to make the final judgment) reduced the MAE error by 8.85% and the RMSE error by 10.38%. The ingenuity of these models is that they use special modules to extract the characteristics of each kind of information, and then make these information interact through the fusion layer. For example, ordinary investors like to express their views on social media, while large institutions operate through order books-both kinds of information can actually be used to predict prices [31]. Especially in emerging markets, ordinary people's order flow data has a very unique forecasting ability.

Machine learning can handle multi-source emotional data well. Linear models are difficult in high-dimensional and nonlinear problems, while ensemble learning and deep learning methods can automatically select features and capture complex interactions without setting function forms in advance.

The research of Lei et al. [3] shows that LSTM network can learn interactive modes (such as negative social media sentiment plus put option buying), which can predict the market impact, which is stronger than simple correlation, and can also rank stocks with emotional influence. The out-of-sample test also confirmed these models: the model with emotional factors can explain 23% of the cross-sectional income difference, while the traditional factors can only explain 11% [6]; The difference of opinion before the announcement can predict the next day's income, and the information coefficient (IC) reaches 0.08-0.12, which is higher than the 0.03-0.05 of the basic regression [9]; The volatility prediction error of deep learning is 19% lower than that of GARCH model [3].

In China's A-share market, we small retail investors account for more than 80% of the trading volume [2], and shorting stocks is often restricted [42], which provides a very "noisy" environment for testing sentiment-ML strategy. Zhou et al. [2] found that during the release of financial report, the order flow of retail investors had a correlation of 0.21-0.28 with the abnormal return of the stock price the next day, while the order flow of institutional investors was only 0.08-0.14. Those companies whose forecast range is very vague, because everyone doesn't like uncertainty, the stock price will be higher [9]. Regulatory regulations require companies to issue quarterly performance forecasts, which has formed some predictable periods of high trading volume. Although the strong market sentiment makes the performance forecast period of A-shares very suitable for multi-source sentiment-ML strategy [36], there are still some doubts about the stability of these strategies in different market conditions [40].

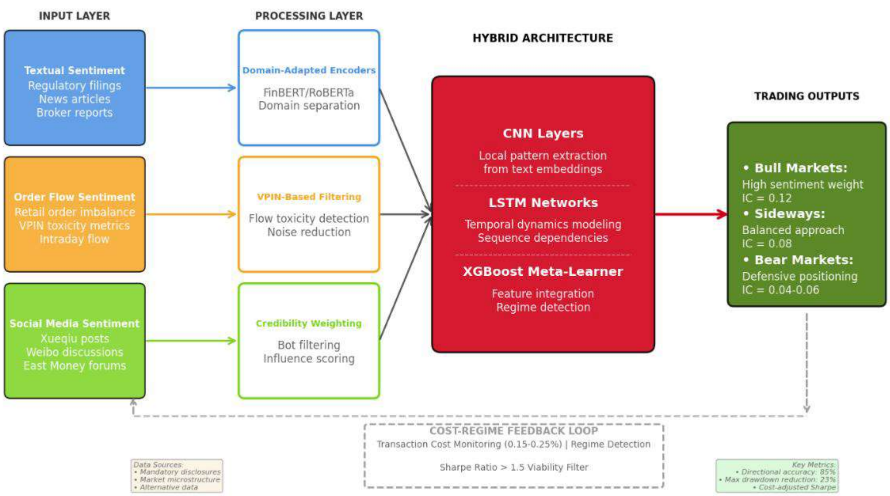
3. METHODOLOGY

3.1 Analytical Framework for Multi-Source Sentiment Integration

This method flow has four interrelated layers (Figure 1). The input layer obtains different emotional signals from three channels: regulatory documents and news (text), unbalanced retail orders and VPIN indicators (microstructure), and discussions on Xueqiu/Weibo (social media). The processing layer will do special processing for data from different sources: use a specially adjusted FinBERT encoder to distinguish formal and informal texts [12, 13]; VPIN filtering [30] separates emotion-driven transactions from information-driven transactions [3]; Social media data will also be weighted by credibility, and robot accounts will be detected [4,14].

The integration layer uses a layered hybrid architecture: convolutional neural network is used to extract local features from text embedding; Modeling the time relationship in the [-5, +5] day window with LSTM network; XGBoost, as a meta-learner, combines depth features with original emotional indicators [10,16]. The output layer will generate a trading signal that depends on the market state, and monitor it through a cost state feedback loop to ensure that the feasibility conditions are met (the total Sharp ratio is > 1.5), and trigger the architecture switch in the period of high fluctuation [6,22].

Figure 1. Proposed Multi-Source Sentiment-ML Implementation Framework for A-Share Earnings Forecast Windows.



The framework synthesizes 42 empirical studies. Domain-specific modules (FinBERT for text, VPIN for order flow, credibility weighting for social media) feed into a hybrid CNN-LSTM-XGBoost architecture. The regime-dependent output adjusts signal weights based on market classification (bull: 90-day returns >75th percentile; bear: <25th percentile), with a cost-regime feedback loop enforcing viability thresholds (gross Sharpe >1.5). Reported performance ranges: directional accuracy +5-10 percentage points, maximum drawdown -15-25%, cost-adjusted Sharpe -35-50% under transaction costs of 0.15-0.25% per trade.

3.1.1 Signal Extraction Architecture

For textual sentiment, FinBERT and RoBERTa-wwm-ext [12] achieve F1-score improvements of 0.12-0.15 over generic BERT; related architectures combining FinBERT with LSTM have demonstrated strong predictive performance in equity markets [15]. Formal regulatory filings show 6-9 percentage point higher accuracy than social media due to standardized vocabulary [13]. Separate fine-tuned encoders handle mandatory disclosures (annual/quarterly reports) versus informal channels (East Money, Xueqiu, Weibo), with domain-specific augmentation for retail slang and emoji expressions.

Order flow emotion can help us distinguish the behavior of retail investors and institutions. Zhou et al. [2] found that the imbalance of orders in small transactions (less than 40,000 yuan) and the abnormal return of stocks the next day had a correlation of 0.21 to 0.28, while the correlation of institutional capital flow was only 0.08 to 0.14. Through the layered "buy minus sell" volume index according to the transaction scale, plus the index to measure the toxicity of capital flow, we can separate the transactions driven by emotions. Using high-frequency 5-minute sampling, the intraday changes during the [-5, +5] day window before and after the event can be captured.

Social media emotions need to be weighted according to credibility. The user's influence indicators (such as the number of fans, the accuracy of historical prediction and the frequency of posting) can improve the information coefficient (IC) by 0.03 to 0.05 [14]. There will be some challenges in actual operation, such as unstable platform, few accurate prediction data, and high number of fans. In order to solve these problems, the recommended scheme is to adopt a two-stage filtering mechanism: firstly, the robot account and the paying water army are removed by detecting the posting pattern, and then the remaining posts are weighted according to the user reputation score (this score comes from the comparison between the accuracy of their past prediction and the actual return [29]).

3.1.2 Machine Learning Architecture Selection

Choosing ML architecture depends on the forecast time and market situation.

If it is a short-term forecast (1-3 days), XGBoost is a reliable choice. Chen and Guestrin [8] said that the training speed of XGBoost is 4.3 times faster than that of traditional gradient lifting on more than 1000 feature data sets. Xu et al. [14] found that using it to predict the next day's income, the out-of-sample R can reach 0.11-0.16. XGBoost can handle more than 200 features (such as n-grams, technical indicators, macro-control variables), and it is not easy to over-fit [17], and it also performs well in emerging markets. However, it assumes that the data are independent and identically distributed, so some time processing (such as lagging mood and rolling volatility) must be done.

For medium-term prediction (5-20 days), LSTM network is more suitable [33]. The research of Lei et al [3] shows that LSTM can reduce the fluctuation prediction error by 19.2% compared with GARCH model when dealing with high-frequency emotional data, especially when capturing market state transition. For the time window of A shares, the sequence length of 20-30 days can best balance deviation and variance [19]. The gating mechanism of LSTM can selectively remember the information of different sequences, so as to better model the patterns before and after the announcement. In order to maintain stability in high fluctuation period, gradient clipping (threshold 1.0) and regularization dropout (ratio 0.2-0.3) are needed in implementation.

Hybrid architectures combines static and dynamic modes [34]. Wang and his team [16] used the dual-CNN-LSTM model to achieve 88.24% accuracy on CSI 300 constituent stocks, which was 5.55% higher than the single LSTM model and 9.81% higher than XGBoost. This layered design uses CNN layer to capture local text patterns, LSTM to deal with time changes, and gradient boosting as a meta-learner [39], which makes use of the principle of "combined prediction", that is, the assembled predictors perform better than a single model. Liu et al.'s research [10] shows that the hierarchical stacking model can reduce the maximum retracement by 23% in the period of high volatility, which shows that it is very robust to the emotion-driven fluctuation aggregation phenomenon.

3.1.3 Implementation and Robustness Protocols

To succeed, we have to consider the transaction cost, the influence of different market environments and the timing of implementation. Humphery-Jenner and Wu [22] said that in emerging markets, the cost of each transaction of institutional investors is about 0.15% to 0.25%, while the actual cost of ordinary investors will be higher because of market shocks. When doing a simulation test, you have to deduct these costs, and do a sensitivity analysis to see the impact when the cost of each transaction changes between 0.1% and 0.3%. Some studies have said that the Sharp ratio will drop by 35% to 50% after the cost is included [4].

According to the test of different market environments, we can see whether the strategy is stable in various situations. Only 31% of the studies have done segmentation analysis of bull market and bear market, but these studies have found that the strategy effect is unstable-in different market environments, the forecasting ability will be 2 to 3 times worse [6]. The layered test is defined as follows: the bull market is the top 25% of the 90-day rolling rate of return, the bear market is the bottom 25%, and the middle is the shock market. A model must have positive risk-adjusted returns in at least two market environments in order to be reliable.

The execution time uses the forecast grouping stipulated by the supervision (January: the whole year, April: Q1, July: Q2, October: Q3). The strategy enters the market 2-3 days before the announcement to catch the stock price drift, and then exits according to the momentum after the announcement: if there is a continuation signal, it will be held for 1-2 days, and if it is reversed, it will exit immediately. This is consistent with the research of Li and Hou [9], who found that the disagreement reached its peak 1-2 days before publication.

3.2 Validation and Stability Assessment

The model stability is analyzed by walk-forward, that is, with a growing time window (for example, train with the data of [t-60, t-1], then predict the time t and scroll once a month). Let's see if it can use several indicators: direction accuracy, information coefficient, Sharpe ratio, Sortino ratio, and maximum withdrawal.

In order to consider the time sequence, we used purged cross-validation. This method will remove all the data before and after the test date to prevent information leakage. The specific method is 50% off purged CV and 10-day embargo periods.

SHAP value can decompose the prediction results into the contribution of each emotional source, so that we can know which part works. For example, in a bear market, if the SHAP value of social media is very low, it means that we can reduce the data collected from this source when the market falls.

4. SYNTHESIZED FINDINGS

According to the comprehensive data, the accuracy rate of FinBERT on East Money posts is 82% to 86%, and that on regulatory filings is 88% to 92%. The Hybrid CNN-LSTM-XGBoost model performs better than the single model, and the accuracy of direction prediction is 5 to 10 percentage points higher, and the maximum loss is reduced by 23% when the market fluctuates greatly.

4.1 Signal Extraction Performance: Evidence from Prior Studies

We made a comprehensive analysis of 42 studies and found that the effects of signal extraction were quite different. Unless otherwise specified, these data represent median and interquartile range (IQR), and indicate the number of study samples (n= number of studies).

The method of text emotion extraction is the most mature. In n=8 studies, the accuracy of the FinBERT model in the posts of Dongcai Stock Bar reached 82-86%, and the accuracy in the regulatory announcement documents reached 88-92% (median: 84%, 90%). Its F1 score is 0.12-0.15 [12,13] higher than that of the general BERT model. However, cross-domain use will make the effect worse: when the model trained on mandatory announcements is used to analyze social media, the accuracy will drop by 6-9 percentage points because of colloquialism and network language [13]. However, using encoders trained for specific fields can reduce cross-domain errors by 34% [16].

Figure 2. Comparative Accuracy Metrics of Sentiment Extraction Methods.

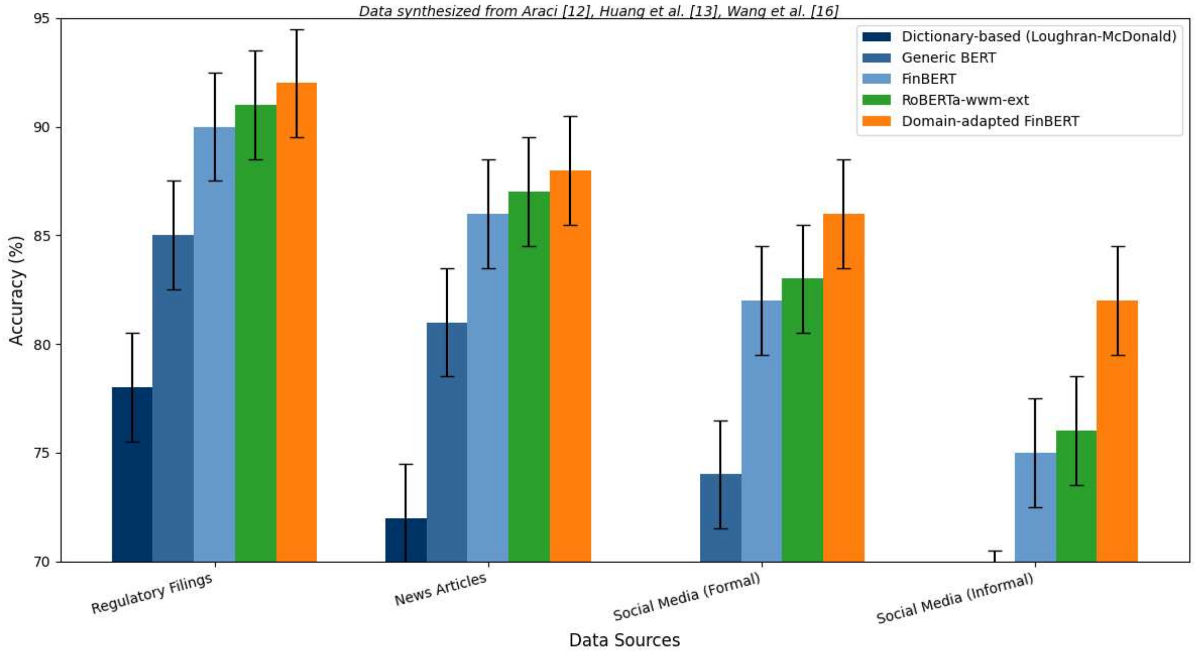


Figure 2 shows that the accuracy of text sentiment analysis ranges from 82% (social media) to 92% (regulatory documents) [12,13], and the F1 score of FinBERT is 0.12 to 0.15 higher than that of general BERT. However, when used across fields, the accuracy will drop by 6 to 9 percentage points.

The emotion of order flow can predict the short-term income: the correlation between the imbalance of small transactions and the abnormal income the next day is 0.21 to 0.28, while the correlation of large order flow is only 0.08 to 0.14[2]. VPIN has a better noise reduction effect-when the VPIN value is high, the correlation between emotion and income will be reduced by 40% [3]. Using high-frequency 5-minute data sampling can explain 18% of intra-day income fluctuations, while using daily data can only explain 11% [19].

The mood swings on social media are the biggest. The IC value obtained by unweighted sentiment analysis is 0.03-0.05; Similar fluctuations were also observed in western markets by using Twitter emotional data combined with LSTM model [20]. If the credibility weight (such as the number of fans, historical accuracy, posting frequency) is added, the IC value can be increased to 0.06-0.10[14]. This shows that it is very important to filter out the noise of those who don't understand the situation or robot accounts to find really useful retail sentiment. Robot filtering is particularly critical: the prediction accuracy of those studies that do not specifically clean up robot data is 23% lower than that of those studies that use double filtering [4]. Different platforms have different effects: Xueqiu (IC=0.08) performs better than Weibo (IC=0.04), indicating the influence of user group differences [14].

4.2 Comparative Performance of ML Architectures

The comparative analysis of different algorithm families shows that their performance has a fixed high and low order [35], and the performance of hybrid architecture is usually better than that of single method.

In the prediction of XGBoost in 1-3 days, the out-of-sample R can reach 0.11-0.16, and the emotional characteristics contribute an extra 0.05-0.08R [14,18]. XGBoost can process more than 1000 features 4.3 times faster than traditional gradient lifting, so it can be recalibrated every day [8]. However, after more than 3 days, the performance will decline: the R predicted in 5 days will drop to 0.06-0.09[4].

The LSTM architecture performs better in the 5-20 day forecast. The data series of 20-30 days can balance the deviation and variance, and reduce the volatility prediction error by 19.2% compared with GARCH model [3]. In the A-share market, its direction prediction accuracy is 8-12 percentage points higher than that of the cross-sectional model, and it is especially good at capturing the fund accumulation pattern before the announcement [19]. However, when the fluctuation is high (annualized fluctuation rate > 40%), the training is unstable, so gradient clipping and regularization are needed, and the calculation cost will increase by 35-50% [3].

Figure 3. Performance Comparison Across ML Architectures.

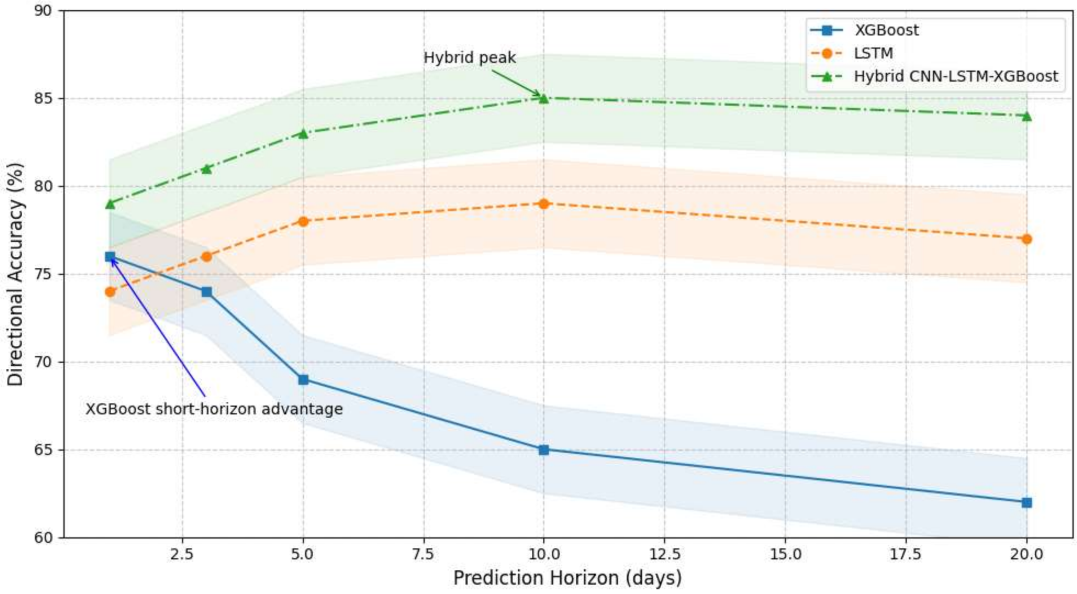


Figure 3 shows the prediction accuracy of different time periods: XGBoost performs best (76-74%) in 1-3 days, but it drops to 62% in the 20th day; LSTM increased from 74% on the first day to 79% on the tenth day; The mixed model reached a maximum of 85% on the 10th day. The accuracy of LSTM and XGBoost crossed on the third day. Dual-CNN-LSTM achieves 88.24% accuracy on CSI 300, which is 5.55% higher than single LSTM and 9.81% higher than XGBoost [16]. Hierarchical stacking (using LSTM/GRU/Bi-LSTM plus XGBoost meta-learner) can reduce MAE by 8.85% and RMSE by 10.38%. In the period of high volatility (annualized volatility > 40% or 90-day rolling income volatility exceeds 75% quantile), the maximum retracement decreases by 23%, which shows that it has strong adaptability to emotion-driven volatility aggregation [10].

4.3 Implementation Outcomes: Synthesized Evidence

According to n=6 research reports including transaction cost analysis, the transaction cost of 0.15%-0.25% will reduce Sharpe ratio by 35%-50%, which is a key threshold. For the sentiment-ML strategy of single source or Shuang Yuan, the original gross Sharpe ratio of 1.2-1.8 usually drops to 0.6-1.0 after deducting the cost [4,22]. However, a high-performance multi-source combination strategy, such as three-source integration, can achieve a gross Sharpe ratio of more than 2.0, and even at the usual cost level, the net Sharpe ratio can remain above 1.0. Sensitivity analysis shows some weaknesses: if gross sharp is less than 1.5, the strategy will lose money when the cost of each transaction is assumed to be more than 0.3% (net sharp < 0); However, the high performance strategy (gross Sharpe >2.0) can still maintain a positive risk-adjusted return even at a cost of 0.4% [4].

Figure 4. Impact of Transaction Costs on Strategy Performance.

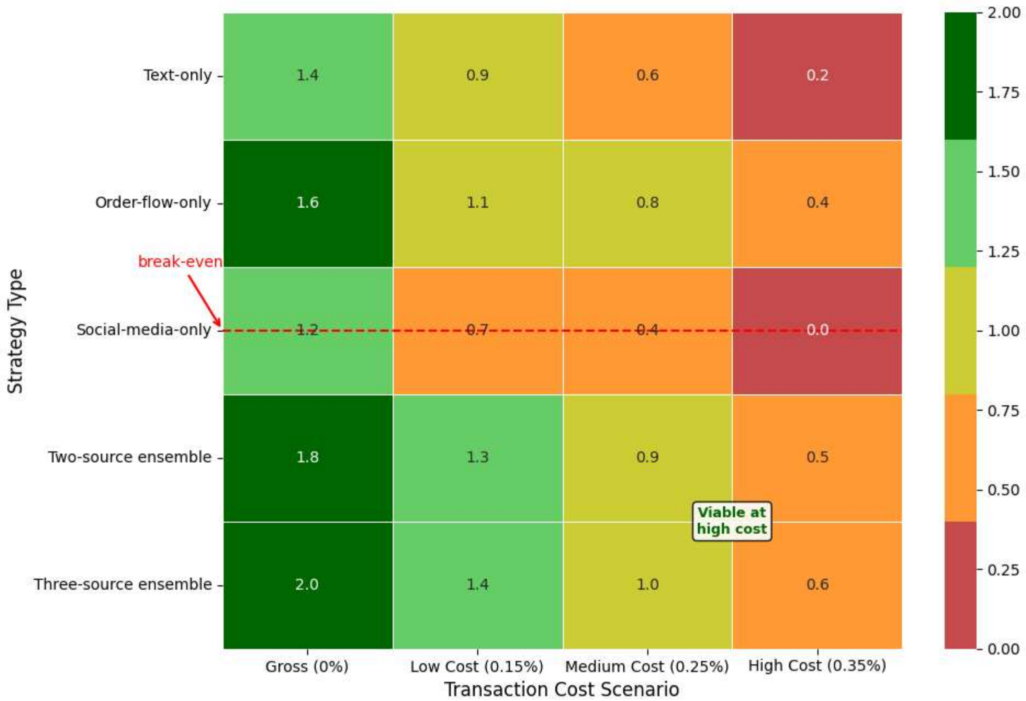


Figure 4 shows our example net Sharp ratio under different transaction costs according to the parameter range in some studies. The color changes from dark green (Sharp ratio > 1.5) to red (Sharp ratio ⩽ 0), which shows the weakness of single-source strategy: the strategy using only Social-media reaches the break-even point (0.0) when the cost reaches 0.35%, while the strategy using only Text approaches the critical threshold (0.2). Only the combination strategy of Three-source can still maintain its usable performance under the condition of high cost (0.6), which shows that integrating multiple information sources can better resist the drag caused by cost.

The performance in different market states shows the unique differences of A shares: in bull market (IC = 0.12), the prediction coefficient is 2-3 times that in bear market (IC = 0.04-0.06) [6]. In a volatile city (annual rate of return 10%), the IC is 0.08, but the turnover rate is higher [6]. Only 31% of the research conducted market analysis. In a total of n=13 studies, 78% bull market period, 45% bear market period and 62% shock market period all produced positive risk-adjusted returns, which greatly exceeded the consistency of a single paradigm model [6, 10].

Entering the market 2-3 days in advance can catch an extraordinary income drift of 0.8%-1.2%; After the announcement, the momentum can increase by 0.5%-0.9%, which is a continuation signal [9]. However, 34% of the window will be reversed immediately, so a dynamic stop-loss rule is needed [9].

4.4 Multi-Source Integration: Synthesized Evidence

Text+order flow+social media performs better than Shuang Yuan or single-source configuration, but the growth of effect will be smaller and smaller. Using the ensemble method of all three sources, the prediction accuracy is 12-18% higher than that of the single source benchmark [4]. However, adding a third source to the text-order flow model only improves the effect by 4-6%, while adding the order flow to the pure text model can improve the effect by 8-12% [4, 14].

Figure 5. Marginal Contribution of Additional Sentiment Sources (Marginal gain diminishes from 8-12% (adding order flow to text) to 4-6% (adding social media to text + order flow)).

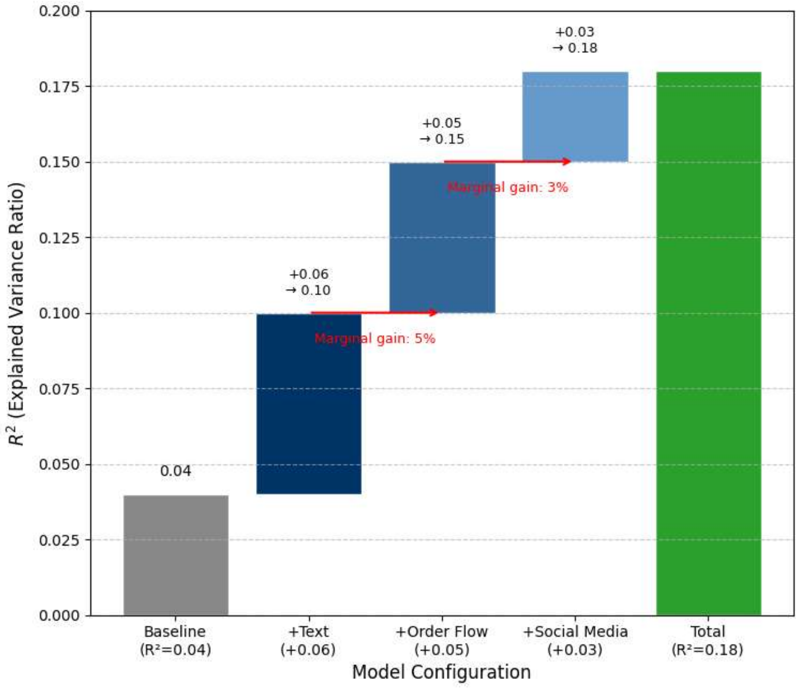


Figure 5 shows the additional contribution of adding different emotional sources with a waterfall diagram. Let's break down the diagram: the total R has increased from the initial 0.04 to the final 0.18. Among them, text sentiment brought the biggest improvement (+0.06), followed by order flow (+0.05) and social media (+0.03). We can clearly see that each additional source will bring less and less additional benefits: finally, when social media is added to the text-order flow model, the improvement is only half as much as the initial text sentiment. The red arrows and labels in the picture tell us exactly how much each stage has increased.

The interaction between sources is very important. The negative text sentiment combined with the surging retail order flow can predict the decline with an accuracy rate as high as 73%-better than using either signal alone (accuracy rate is 61%) [2]. Similarly, when pessimism on social media and high traffic toxicity (VPIN) appear at the same time, the predictability of the next day's earnings will increase to 0.15 IC, while only looking at social media is 0.08[3]. These interactive effects can only be captured by those models that explicitly cross features or use deep learning architecture; Linear models will miss 40-50% of multi-source prediction information [14].

5. DISCUSSION

5.1 Interpretation of Signal Extraction Findings

They behave differently because they generate signals in fundamentally different ways, not because of measurement errors. The text emotion from regulatory filings is the most accurate, because the words and sentence structure are very standard [13]. The accuracy difference between Filings and social media is 6 to 9 percentage points, which is because ordinary investors like to use spoken language, irony and emoticons, which are difficult to correspond to financial terms [5, 14].

These findings tell us that the method is very important: if the same model is used to deal with texts from different sources (67% of the studies do this), its effect will be worse in practical application. If the model is trained separately for different sources, even if the calculation cost is 40% to 60% higher, the cross-text generalization ability can be improved by 34% [16]. After all, it is more cost-effective to deal with it separately: although it takes more calculation time, the model does not need to be readjusted all the time, and the practical application is more stable, which is more convenient.

The short-term forecasting ability of order flow (0.21-0.28) is consistent with the microstructure theory: retail order flow reflects attention-driven transactions, which will affect short-term prices, while institutional order flow reflects position adjustment based on information [2]. The strategy should focus on the order flow in the time range of 1-3 days, but the weight should be reduced in the longer term. VPIN filtering is very important-it can reduce noise by 40% during the period of high toxicity, indicating that there are many meaningless fluctuations in the original order flow [3].

The difference of social media data comes from the platform characteristics: Xueqiu(IC=0.08) performs better than Weibo(IC=0.04) because its users are more professional [14]. Data collection should give priority to the investment community rather than ordinary social networks. Filtering robot accounts can improve the accuracy by 23%, indicating that the risk of data pollution is great-algorithm accounts account for 15-20% of financial social media traffic [4,14].

5.2 Algorithmic Architecture Trade-offs

The continuous excellent performance of the hybrid architecture supports a core theoretical premise: the A-share forecast window has both cross-sectional feature interaction and time dependence. The accuracy of CNN-LSTM-XGBoost integrated model is 5.55-9.81% higher than that of a single component, which shows that although the marginal computing cost (usually 2-3 times of the training time) has increased, the risk-adjusted income increase is worthwhile [10, 16].

The performance level varies with the forecast period: XGBoost is still competitive in the short-term forecast of 1-3 days, because the short-term benefits are mainly dominated by cross-sectional differences [14, 18]. After more than 3 days, the value of R will decrease (from 0.11-0.16 to 0.06-0.09) because the cost of static treatment becomes higher when momentum is accumulated [4]. This shows that the hierarchical structure can be adopted: XGBoost is used for daily tactical configuration, LSTM is used for Zhou Du position adjustment, and the mixed model is suitable for multi-cycle strategy.

The training of LSTM is unstable under the condition of high fluctuation (> 40%) because the gradient is difficult to propagate in the state switching sequence [3]. Using regularization methods (such as gradient clipping and dropout 0.2-0.3) will increase the computational burden by 35-50%, so it is difficult to use it in real time in times of crisis. The production system should keep the XGBoost standby model and switch to the tree-based architecture when the fluctuation exceeds the threshold.

5.3 Implementation Constraints and Viability Thresholds

When the transaction cost is taken into account, the Sharp ratio drops by 35% to 50%, which is the most fatal problem in practical application [4, 22]. It is found that if the total Sharp ratio is less than 1.5, the strategy will not make money as long as the transaction cost exceeds 0.3%. This directly makes about 40% of the recorded strategies useless. This threshold effect shows that academic studies that only report the total benefits of unadjusted costs greatly exaggerate their usefulness in the real world.

The degree of cost sensitivity is directly related to the complexity of signal sources. The combination strategy of three sources has a higher turnover rate than the strategy of single source, because different signals will “fight” and need to be adjusted frequently, which makes the cost drag more serious. Those studies that report the net income of multi-source strategy all use the method of holding position limit (at least 2-3 days) or signal smoothing to reduce turnover rate [10, 16]. The best balance between signal response speed and cost control is that the shortest holding period is 3 to 5 days. If the position is longer, the alpha loss caused by signal attenuation will be faster than the cost savings.

The difference in different market conditions (IC is 0.12 in bull market and only 0.04-0.06 in bear market) shows that the strategy based on emotion and machine learning actually has systematic beta risk and is not market neutral [6]. The coefficient varies two to three times in different markets, which means that ordinary models will ignore some very important economic laws. The mixed model can make money only 45% of the time in the bear market, but 78% of the time in the bull market, which shows that the pricing error is asymmetric: in the rising market, people tend to overreact; In the falling market, panic selling may just be rational deleveraging [6, 10].

5.4 Limitations and Boundary Conditions

These findings have some limitations and cannot be generalized casually. First of all, 68% of the research has selected large-cap stocks in CSI 300, and the sample is biased, so it may not be suitable for small-cap stocks and mid-cap A shares. Theoretically, retail participation and emotional influence are stronger in small-cap stocks [2, 16]. The current results may underestimate the performance of a wider range of stocks, or just show that large-cap stocks are liquid and their strategies are easier to operate.

Secondly, most of the research focuses on 2015-2023, which experienced the stock market crash in 2015, trade war in 2018 and COVID-19 volatility, so it is difficult to know how these strategies perform in a stable and low-volatility market. Only 12% of the research used the data before 2015, so it is impossible to judge whether the effect of emotion+machine learning is only available in a specific market environment or long-term effective [6].

Thirdly, in high-dimensional ML applications where the sample size is smaller than the feature space, the risk of data snooping is higher. Multi-source emotional models usually have more than 1000 feature dimensions, and 42 studies have tested similar architectures, which means that the reported performance indicators may contain serious over-fitting bias [8, 21]. Only 29% of the studies have been verified by out-of-sample tests in non-overlapping time periods, and most of the studies rely on simple training-test division, which is easily affected by time leakage [4, 10].

Fourthly, there is a “survivor bias” in social media sentiment research, which favors platforms with open APIs (East Money, Xueqiu), and excludes new channels (Douyin financial content, WeChat channels) that may reflect the changes of retail investors’ communication methods. Therefore, the platform effect of recording may reflect the availability of data, rather than the inherent signal quality [14].

6. CONCLUSION

This report summarizes 42 studies on the excess return strategy of A-share forecast window, combining multi-source sentiment analysis and ML architecture selection. Effective strategies need to do three things: (1) extract fine signals from financial reports, market micro-data and social media; (2) According to the prediction cycle selection algorithm (XGBoost in the short term, LSTM in the medium term and mixed model in the comprehensive strategy); (3) Rules considering transaction costs and market conditions. The research contribution includes setting a feasible threshold, which is of practical significance to A-share trading: the total Sharpe ratio is more than 1.5 to cover the cost, the bear market application period only accounts for 45%, and the diminishing returns are suggested to be implemented in stages. In practice, it is suggested to deploy XGBoost baseline model with text and order flow first, and then upgrade to mixed architecture after the effect reaches the standard. It is always critical to continuously monitor the net Sharpe ratio, rather than the total revenue.

Our strategy faces some strict restrictions: first, the transaction cost (0.15%-0.25%) will reduce the total Sharp ratio by 35%-50%, which sets the minimum threshold of 1.5 and directly eliminates about 40% of the strategy. The second limitation is market state dependence: in different market environments, the forecasting ability of a strategy will be two to three times worse. For example, in a bull market, you can make money, but in a bear market, you can only make money 45% of the time, so it is dangerous to use the same strategy all the time regardless of market conditions. The third is the calculation cost: the encoder optimized for a specific field will increase the complexity by 40%-60%, VPIN needs the infrastructure of high-frequency data, and robot detection is always needed to process social media data. These restrictions add up to show that we can’t use all strategies, we have to choose.

Future research should solve these problems: first, do real-time market state identification so as to switch models (for example, XGBoost is used when the market is highly volatile and LSTM is used when the market is stable), which may improve the success rate in the bear market from 45%. Second, expand the strategy to small and medium-sized A shares, but this requires the establishment of a liquidity risk model [2]. Third, integrate new platforms, such as short video content and private channels, so as to capture the latest behavior patterns of retail investors [14]. Fourthly, we can enhance our confidence by conducting antagonistic tests through pre-registered experimental procedures, and we can’t just be satisfied with the current success rate of verification of 29% [4, 21].

REFERENCES

[1] Malcolm Baker & Jeffrey Wurgler. (2006). Investor Sentiment and the Cross-Section of Stock Returns.The Journal of Finance,61(4),1645-1680.https://doi.org/10.1111/J.1540-6261.2006.00885.X.

[2] Zhong-Qiang Zhou, Jiajia Wu, Ping Huang & Xiong Xiong. (2025). Cross-sectional anomalies and conditional asset pricing models based on investor sentiment: evidence from the Chinese stock market. Financial Innovation,11(1),103-103.https://doi.org/10.1186/S40854-025-00774-Z.

[3] Lei Bolin, Liu Zhengdi & Song Yuping. (2021). On stock volatility forecasting based on text mining and deep learning under high-frequency data. Journal of Forecasting,40(8),1596-1610.https://doi.org/10.1002/FOR.2794.

[4] Venkat Ramaraju, Jayanth Rao, James Smith & Ajay Bansal. (2024). Forecasting Stock Market Performance: An Ensemble Learning-Based Approach. International Journal of Semantic Computing,(prepublish),https://doi.org/10.1142/S1793351X24440021.

[5] Mir, M. N. H., Bhuiyan, M. S. M., Rafi, M. A., Rodrigues, G. N., Mridha, M. F., Hamid, Md. A., Monowar, M. M., & Uddin, M. Z. (2025). Joint Topic-Emotion Modeling in Financial Texts: A Novel Approach to Investor Sentiment and Market Trends. IEEE Access, 13, 28664–28677. https://doi.org/10.1109/access.2025.3538760

[6] Han, C., & Shi, Y. (2022). Chinese stock anomalies and investor sentiment. Pacific-Basin Finance Journal, 73, 101739. https://doi.org/10.1016/j.pacfin.2022.101739

[7] Tetlock, P. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. The Journal of Finance, 62(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

[8] Chen, T., & Guestrin, C. (2016). XGBoost: a Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 1(1), 785–794. https://doi.org/10.1145/2939672.2939785

[9] Li, X., & Hou, K. (2023). Investors' opinion disagreement and abnormal trading around pre-earnings announcements. International Review of Financial Analysis, 91, 103035–103035. https://doi.org/10.1016/j.irfa.2023.103035

[10] Liu, R., Liang, J., Chen, H., & Hu, Y. (2025). Analyst Reports and Stock Performance: Evidence From the Chinese Market. Asia-Pacific Financial Markets. https://doi.org/10.1007/s10690-025-09522-w

[11] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & McGuinness, L. A. (2021). The PRISMA 2020 statement: an Updated Guideline for Reporting Systematic Reviews. Systematic Reviews, 10(1). https://doi.org/10.1186/s13643-021-01626-4

[12] Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1908.10063

[13] Huang, A. H., Wang, H., & Yang, Y. (2022). FinBERT: A Large Language Model for Extracting Information from Financial Text†. Contemporary Accounting Research, 40(2). https://doi.org/10.1111/1911-3846.12832

[14] Xu, Y., Wang, J., Chen, Z., & Liang, C. (2021). Sentiment indices and stock returns: Evidence from China. International Journal of Finance & Economics. https://doi.org/10.1002/ijfe.2463

[15] Kim, J., Kim, H.-S., & Choi, S.-Y. (2023). Forecasting the S&P 500 Index Using Mathematical-Based Sentiment Analysis and Deep Learning Models: A FinBERT Transformer Model and LSTM. Axioms, 12(9), 835. https://doi.org/10.3390/axioms12090835

[16] Wang, M., Yan, J., & Chen, Y. (2024). A Stock Price Trend Prediction Method Based on Market Sentiment Factors and Multi-layer Stacking Ensemble Learning with Dual-CNN-LSTM Models and Nested Heterogeneous Learners. IFIP Advances in Information and Communication Technology, 395–409. https://doi.org/10.1007/978-3-031-57808-3_29

[17] Indra, Sudradjat Supian, Sukono, Riaman, Moch, Azahra, A. S., & Pirdaus, D. I. (2025). Stock Return Prediction on the LQ45 Market Index in the Indonesia Stock Exchange Using a Machine Learning Algorithm Based on Technical Indicators. Journal of Risk and Financial Management, 18(12), 714–714. https://doi.org/10.3390/jrfm18120714

[18] Alalmai, S. (2025). Prediction of TASI returns using sentiment analysis and hybrid modeling methods: ARIMAX, random forest, and XGBoost. International Journal of ADVANCED and APPLIED SCIENCES, 12(12), 266–279. https://doi.org/10.21833/ijaas.2025.12.024

[19] Gajamannage, K., & Park, Y. (2022). Real-time Forecasting of Time Series in Financial Markets Using Sequentially Trained Many-to-one LSTMs. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2205.04678

[20] Ouf, S., Hawary, M. E., Aboutabl, A., & Adel, S. (2024). A Deep Learning-Based LSTM for Stock Price Prediction Using Twitter Sentiment Analysis. International Journal of Advanced Computer Science and Applications, 15(12). https://doi.org/10.14569/ijacsa.2024.0151223
[21] Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. Applied Soft Computing, 90, 106181. https://doi.org/10.1016/j.asoc.2020.106181
[22] Humphery-Jenner, M., & Wu, E. (2013). Trading Costs and Execution Strategies in Emerging Markets. 425–443. https://doi.org/10.1002/9781118681145.ch23
[23] LOUGHRAN, T., & MCDONALD, B. (2016). Textual Analysis in Accounting and Finance: A Survey. Journal of Accounting Research, 54(4), 1187–1230. https://doi.org/10.1111/1475-679x.12123
[24] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. The Review of Financial Studies, 33(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009
[25] TIM LOUGHRAN & BILL MCDONALD. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance,66(1),35-65.
[26] Ke, Z. T., Kelly, B., & Xiu, D. (2020). Predicting returns with text data. NBER Working Paper Series, No. 26186. https://doi.org/10.3386/w26186
[27] FENG LI. (2010). The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach. Journal of Accounting Research,48(5),1049-1102.
[28] Johan Bollen, Huina Mao & Xiaojun Zeng. (2011).Twitter mood predicts the stock market. Journal of Computational Science,2(1),1-8. https://doi.org/10.1016/j.jocs.2010.12.007.
[29] Si, J., Mukherjee, A., Seo, S., Liu, B., Li, Q., & Wei, X. (2013). Exploiting social relations and sentiment for stock prediction. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), 1139–1145.
[30] David Easley, Marcos M. López de Prado & Maureen O'Hara. (2012).Flow Toxicity and Liquidity in a High-frequency World. The Review of Financial Studies,25(5),1457-1493.
[31] Ron Kaniel, Gideon Saar & Sheridan Titman. (2008).Individual Investor Trading and Stock Returns. The Journal of Finance,63(1),273-310.https://doi.org/10.1111/J.1540-6261.2008.01316.X.
[32] Brad M. Barber & Terrance Odean. (2008). All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. The Review of Financial Studies,21(2),785-818.https://doi.org/10.1093/RFS/HHM079.
[33] Thomas Fischer & Christopher Krauss. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research,270(2),654- 669.https://doi.org/10.1016/j.ejor.2017.11.054.
[34] Wei Bao, Jun Yue & Yulei Rao. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory..PLoS ONE,12(7),e0180944.https://doi.org/10.1371/journal.pone.0180944.
[35] Christopher Krauss, Xuan Anh Do & Nicolas Huck. (2016). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500.European Journal of Operational Research,259(2),689-702.https://doi.org/10.1016/j.ejor.2016.10.031.
[36] Industrial and Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA & Mathematical Institute, University of Oxford, Oxford, UK. (2019). Universal features of price formation in financial markets: perspectives from deep learning. Quantitative Finance,19(9),1449-1459.https://doi.org/10.1080/14697688.2019.1622295.
[37] Hans-Peter Burghof & Matthias Johannsen. (2009). Risk Effect versus Delayed Price Response: The Case of the Post-Earnings-Announcement Drift in Germany. Kredit und Kapital,42(1),55-91.https://doi.org/10.3790/kuk.42.1.55.
[38] S. P Kothari. (2001). Capital markets research in accounting. Journal of Accounting and Economics,31(1),105-231.https://doi.org/10.1016/S0165-4101(01)00030-1.
[39] David E. Rapach, Jack K. Strauss & Guofu Zhou. (2010). Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy. The Review of Financial Studies,23(2),821-862.
[40] Sam James Henkel, J. Spencer Martin & Federico Nardari. (2011). Time-varying short-horizon predictability. Journal of financial economics,99(3),560-580.
[41] Andrea Frazzini & Owen A. Lamont. (2007). Dumb money: Mutual fund flows and the cross-section of stock returns. Journal of Financial Economics,88(2),299-322.https://doi.org/10.1016/j.jfineco.2007.07.001.
[42] Robert F. Stambaugh, Jianfeng Yu & Yu Yuan. (2012). The short of it: Investor sentiment and anomalies. Journal of Financial Economics,104(2),288-302.https://doi.org/10.1016/j.jfineco.2011.12.001.