Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Deep Learning for Stock Market Prediction: A Systematic Review

Dean Rimmer

dean.rimmer1313@gmail.com

Northumbria University https://orcid.org/0009-0000-2658-8197

Martin Wonders

Northumbria University https://orcid.org/0000-0001-9510-2955

Systematic Review

Keywords: deep learning, real-time forecasting, sentiment analysis, stock-market prediction, trading performance evaluation, transaction-fee adjustment

Posted Date: March 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9267236/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Deep Learning for Stock Market Prediction: A Systematic Review

Dean Rimmer[1*] and Martin Wonders[1]

[1*]Department of Engineering and Environment, Northumbria University, Ellison Pl, Newcastle upon Tyne, United Kingdom.

*Corresponding author(s). E-mail(s): dean.g.rimmer@northumbria.ac.uk;
Contributing authors: martin.wonders@northumbria.ac.uk;

Abstract

Stock market fluctuations is challenging due to nonlinear price dynamics and short-term investor sentiment effects. Deep learning models address challenges by combining historical price and trading/volume features with sentiment signals from financial news, social media, or blended sources. Although studies report strong results, evaluation practices remain inconsistent, ignoring real-world viability. This systematic review analyses 22 papers published 2015-2024 that integrate historical and sentiment data for stock market forecasting. For classification, accuracy is predominant, reported in 13 of 22 studies. However, only 6 papers supplement raw accuracy with discriminatory performance measures, limiting confidence in directional reliability. Among regression papers, 4 report root mean square error (RMSE) and 4 report mean absolute error (MAE) as headline metrics. 5 studies report finance-aware metrics, including net profit or Sharpe ratio; 1 paper adjusts for transaction fees. No regression study reports finance-specific metrics such as maximum drawdown or transaction cost evaluation. Findings show sentiment integration benefits the next-day time horizon; 10 studies report a median accuracy gain of +7.4%. Weaker persistence at longer horizons suggests sentiment accuracy decay. Datasets vary from millions of sentiment records to small annotated samples, often without clear documentation, reducing reproducibility. Sentiment enhanced deep learning improves predictive metrics relative to price only baselines; however, minimal cost aware evaluation and real-time deployment analysis prevent substantiation of commercial viability. Rigorous, standardised, finance-aware evaluation frameworks are required to reconcile academic benchmarks with deployable trading systems.

Keywords: deep learning, real-time forecasting, sentiment analysis, stock-market prediction, trading performance evaluation, transaction-fee adjustment

1 Introduction

Accurately predicting stock market fluctuations is a fundamental problem in financial computing that has historically defied standard statistical analysis. Traditional statistical methods have struggled to model the nonlinearities and inconsistencies present in historical data, a range of modern techniques based on machine learning have been developed to address these challenges. More technologically advanced techniques have been implemented recently to account for this issue, providing a more robust and accurate way to predict stock market fluctuations. Although many models boast significant accuracy with historical data sets in this field, rarely is the resultant model assessed for its viability in real-world scenarios.

Deep learning models have been widely studied for their ability to process historical data with features such as stock open/close prices, trading volume, and day-to-day percentage change in price. However, market movements are influenced not only by numerical features but also by human-driven sentiment. Many papers explore this area of sentiment-driven stock market prediction as a result of exploring data extracted from financial articles, social media, financial reports, or generic news headlines.

Research papers that explore this side of prediction often argue that public perception has a direct correlation with stock prices. They also suggest that combining sentiment data with traditional historical features can improve prediction accuracy Wang et al. (2023). The addition of sentiment features has led researchers to conclude that "Combining investor sentiment with market data enhances prediction accuracy significantly" Li, Bu, and Wu (2017). Although combining historical and sentiment-based approaches shows significant promise, there are significant challenges that remain. Issues, such as the reliability and quality of the data, as well as the considerable complexity of integrating these models into a commercially viable system (transaction costs, profitability testing), are areas that remain unexplored. Real-time integration is particularly challenging, as it requires the development of automated and efficient data preprocessing pipelines to handle incoming data streams. This review seeks to provide a comprehensive analysis of current methods, datasets, results, and the commercial viability of integration, while addressing the challenges that limit real-world applicability.

1.1 Justification & Significance

There is an increasing amount of research on the use of deep learning for stock market prediction; however, few studies provide a comprehensive review that focuses on the quality of data used in training and testing sets. Most existing reviews only document the source of the data without an assessment of reliability Hu, Wang, Ho, and Tan (2021) or consistency Kilimci and Duvar (2020). Data quality varies significantly between studies, potentially skewing results and influencing the perception of the most effective predictive methodology.

The literature is also fragmented in terms of the evaluation of real-time applicability, a critical factor for commercial deployment. Many studies, such as Shilpa and Shambhavi (2023) and Khan et al. (2022) report high predictive accuracy but fail to critically assess model performance in a real-world scenario, often omitting crucial

real-world factors such as transaction costs. A systematic review is necessary to consolidate existing research and highlight gaps, particularly regarding data quality and the integration of real-time predictions, both of which are essential for the practical implementation of these models in commercial trading environments.

This review provides a structured synthesis of 22 studies published between 2015 and 2024 that integrate historical market data with sentiment signals for stock market prediction using deep learning. It systematically categorises existing approaches by model architecture, sentiment source, and prediction horizon, demonstrating a clear horizon-dependent effect in which sentiment integration delivers the most consistent performance gains at short-term (T+1) forecasts, with diminishing effectiveness at longer horizons. Beyond predictive performance, the review critically examines evaluation practices across the literature, revealing a strong reliance on raw accuracy and a widespread absence of finance-aware, cost-adjusted, and risk-sensitive metrics. By explicitly identifying this deployment-readiness gap between academic benchmarking and commercially viable trading systems, the review highlights key operational considerations that are largely neglected in prior work. To address this, it proposes a suite of evaluation metrics that includes statistical performance, trading outcomes, cost viability, and operational realism, providing concrete guidance for more transparent and viability-oriented assessment of sentiment enhanced deep learning models for market prediction.

2 Methods

This section details the process used to transform an initial pool of search results into the curated evidence set analysed in this review. First, it defines the eligibility rules that frame the scope of inquiry. Then it documents the databases consulted and the bespoke search strings used to identify relevant literature. This is followed by a description of the screening and selection workflow, including how titles, abstracts, and full texts were filtered against the pre-specified criteria. Finally, the data-extraction template is described, detailing how each eligible study was converted into a consistent set of variables for analysis. Taken together, these steps provide the transparency and reproducibility necessary for judging both the completeness of the evidence base and the rigour of the conclusions that follow.

2.1 Experimental Design

2.1.1 Eligibility Criteria

This review includes studies that meet the following criteria:

- Model Type: The study must utilise a deep learning model (e.g., CNNs, LSTMs) specifically for predicting stock market fluctuations.
- Data Type: The study must utilize both historical stock market data (e.g., stock open / close volumes, trading amounts, percentage change) and sentiment data (e.g., social media data, news articles).

- Timeframe: The review will exclusively include studies conducted between 2015 - 2024 to reflect recent advancements in deep learning and sentiment analysis artificial intelligence models.

Studies may be excluded based upon:

- Market Type: Studies focused on non-stock financial markets (e.g., cryptocurrency, FOREX).
- Data Type: Studies that use only historical data or only sentiment data, as the focus is on the integration of both data types.
- Model Type: Studies that do not use deep learning models, or that use traditional machine learning algorithms (e.g., support vector machines, decision trees).
- Market Volatility: Studies that focus on particularly volatile markets, such as niche or emerging stock markets, due to the difficulty in generalising their findings to more stable market conditions.
- Data Transparency: Studies that do not provide clear details on how data was collected, processed, and used in model training and validation to ensure transparency and replicability in the review process.

2.1.2 Information Sources

- IEEE Xplore (IEEE/IET Electronic Library): searched on 09 June 2025. This database covers electrical engineering, computer science, electronics, and related technical literature.
- Scopus (Elsevier): searched on 09 June 2025. Scopus indexes a broad range of scientific and technical research across journals, conference proceedings, and books.

No other study registries, organisational websites, or database platforms were consulted.

2.1.3 Data Refinement

The review began with an electronic search of the two databases that were most likely to index technical work on data-driven stock market prediction. The search was deliberately narrow. In Scopus, the full query as outlined Fig. 1 was applied, and the results were limited to material published between 2015 and 2024, yielding 6 results.

```
TITLE-ABS-KEY ( deep AND learning )
AND TITLE-ABS-KEY ( historical AND data )
AND TITLE-ABS-KEY ( sentiment AND analysis )
AND ALL ( data AND quality )
AND TITLE-ABS-KEY ( sentiment AND market )
AND ( TITLE-ABS-KEY ( predict )
OR TITLE-ABS-KEY ( forecast ) )
```

Fig. 1 Scopus Query

In IEEE Xplore, the string "stock market forecasting using deep learning" was paired with an index-term filter and restricted to the same 2015-2024 time window. The index-term filters were as follows: deep learning, sentiment analysis, historical data, data quality, predict, or forecast. Results were then filtered to conference papers and journal articles, yielding 192 conference abstracts and 20 journal papers, for a combined total of 218 results across both databases. No other sources or material types were queried.

All 218 records were exported to a structured Excel workbook. A single reviewer removed duplicates manually, leaving the original six Scopus papers plus the 212 IEEE records for screening. Each title and abstract were assessed against the specified eligibility criteria, with the exclusion or inclusion decisions and reasons recorded in the spreadsheet. Full texts of all studies marked for inclusion were then retrieved and read in full by the same reviewer; inclusion or exclusion at this stage was again documented in the spreadsheet. As the workflow involved only one reviewer, there was no independent duplicate screening, no reconciliation step, and no requirement for adjudication. The process was entirely manual, with no assistance from automation tools or machine-learning prioritisation.

For each study included, the reviewer then extracted data into a second spreadsheet. Alongside basic bibliographic fields, the following variables were recorded:

- Dataset Characteristics; such as price series frequency, sentiment-feature derivation, and training and testing windows.
- Model Architecture and Hyper-parameters; such as network type, batch size, epoch count, optimiser choice, and dropout use.
- Predictive Performance Outcomes; such as root mean square error (RMSE) or mean absolute error (MAE).

When multiple metrics or forecast horizons were reported, such as predictive accuracy (root-mean-square error or mean absolute error), the value highlighted by the study's authors as their principal result was selected for extraction. No risk-of-bias appraisal was undertaken, no study investigators were contacted for clarification, and no text-mining or PDF-parsing utilities were used. All data-collection was carried out manually by the lone reviewer.

2.1.4 Data Items

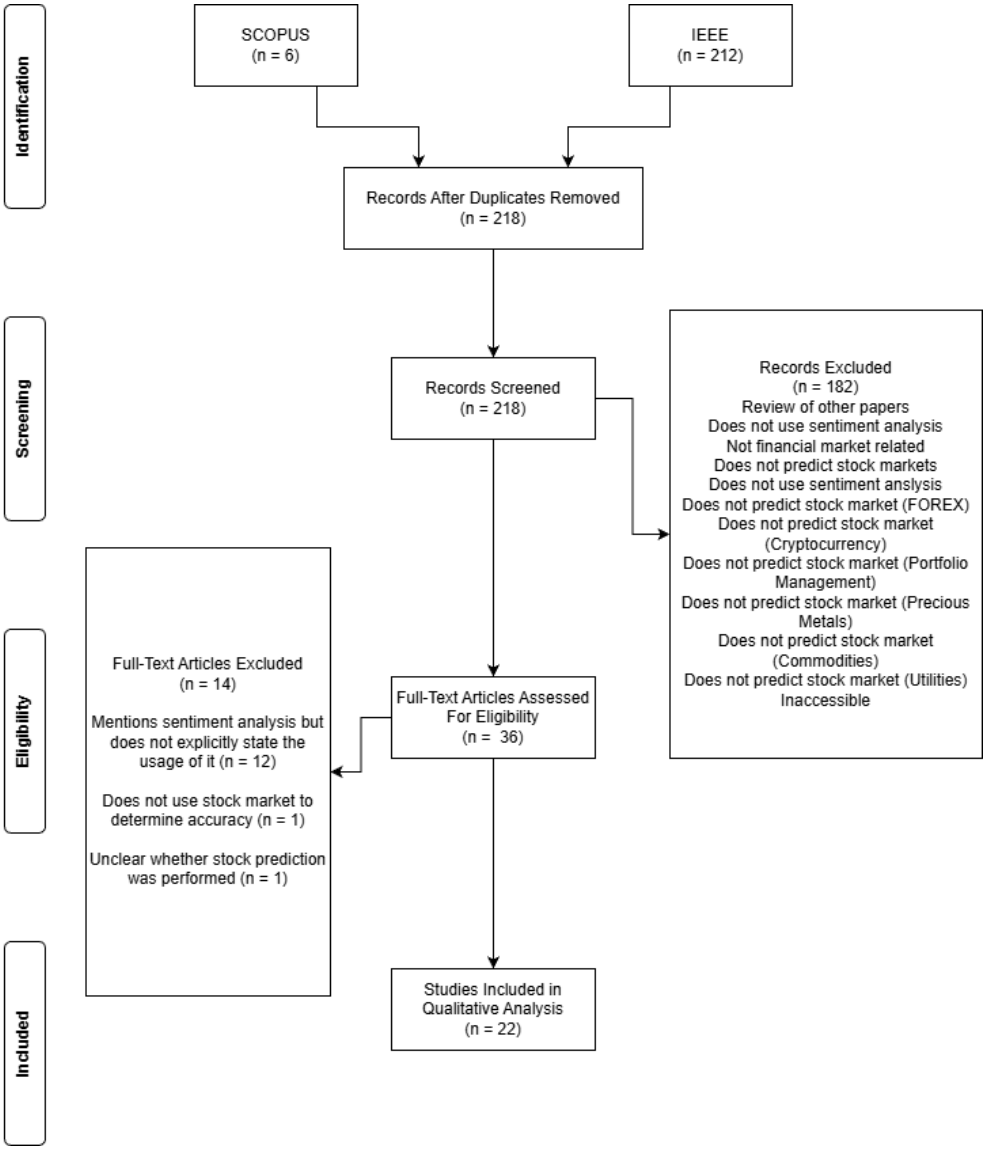


Fig. 2 PRISMA Flow Diagram

Fig. 2 shows that as result of the PRISMA process, an initial 218 documents identified as a synthesis from both research queries were filtered to 22 articles to be included in the systematic review. Most papers were excluded at the abstract screening stage,

with the primary reason for exclusion being that they did not use the technologies or market required for this review.

2.1.5 Risk of Bias

When interpreting the evidence, this review will explicitly consider how study design, data selection, and result presentation may bias observed performance. Possible distortions include preferential reporting of the best-performing model variants, omission of negative or neutral findings, selective use of sentiment or price windows that inflate accuracy, and incomplete disclosure of training/testing splits. To guard against these biases, the review will apply a consistent set of criteria to judge whether each study's data sources, preprocessing steps, and evaluation metrics are described in enough detail to be reproducible. Headline performance metrics will be compared against any secondary results reported in the paper to detect selective emphasis. In addition, the consistency of findings will be examined across studies that use similar data blends and architectures. Studies that fail to meet transparency standards will be excluded from synthesis. By making these judgments explicit and analysing how conclusions shift when less transparent studies are set aside, the review aims to temper over-optimistic claims and provide a balanced evaluation of the true predictive value of deep-learning models that fuse historical and sentiment-based information.

2.1.6 Synthesis Methods

After full-text screening, each eligible study was entered into a master Excel workbook that served as a central data store linking study characteristics to the review's prespecified synthesis plan. Each record was tagged according to the data-source blend it analysed (e.g., news, social media, blended, or other), the deep-learning model type it utilised (e.g. CNN, LSTM, hybrid), and the outcome metric it reported (root-mean-square error, mean absolute error, directional accuracy, f1-score, etc.). Studies that did not provide sufficient data for this quantitative synthesis were removed during the full-text review stage, with the reasons for exclusion documented.

Before any calculations were undertaken, the outcome dataset was harmonised. All price-prediction errors were expressed in daily units and, where necessary, converted to U.S. dollars. Descriptive characteristics of the included studies were summarised using pivot tables exported from Excel, and risk-of-bias traffic-light plots were generated with the official ROB-ME template. Heterogeneity in data sources, model architectures, and market regimes were anticipated, a holistic approach was taken to compare reported accuracy scores across all included studies. Each paper's reported accuracy metrics were collected and compared across all other included studies for synthesis. When implemented across a suitable dataset, this enables identification of the most optimal data sources, model architectures, market regimes, as well as how they individually interact with accuracy metrics.

Potential sources of heterogeneity were examined first through stratified analyses, with separate study pools organised by data type (news versus social-media versus blended) and by model class (RNN, CNN, LSTM). Given the heterogeneity of models, datasets, and evaluation metrics, a formal meta-analysis was not conducted; instead, a structured quantitative aggregation of reported headline metrics was performed.

3 Results

3.1 Study Characteristics

Long-short-term memory (LSTM) networks are generally the most dominant architecture for stock-market prediction, but the selected studies demonstrate some diversity. Twelve of the twenty-two studies primarily use an LSTM variant, ranging from plain single layer models (Li et al., 2017) to convolutional LSTMs for revenue prediction (Mishev et al., 2019) and cyclic LSTMs tuned with Whale or Sparrow Search (Mu, Gao, Wang, & Li, 2023; Qinghua & Wentao, 2022). Two studies replace the LSTM with a lighter GRU or RNN architecture once the text is pre-encoded by FinBERT or TextCNN (Chen, 2023; Kilimci & Duvar, 2020). Three studies allow for an enhanced textual nuance by combining GRUs or explicit attention gates (Lai, Jiang, Ye, & Zhou, 2021; Lien Minh, Sadeghi-Niaraki, Huy, Min, & Moon, 2018; Wang et al., 2023). Feed-forward neural networks are used in two studies; once as a fully connected model incorporating topic, semantic, and sentiment features (Hu et al., 2021) and once as a ReLU stack that classifies Chinese finance headlines (Day & Lee, 2016). Some studies adopted less common approaches: one explored a deep-belief network (Shilpa & Shambhavi, 2023); another implemented a reinforcement-learning trader that debits fixed fees (Day & Lee, 2016), and another used an ensemble voter that utilises an MLP (Khan et al., 2022).

3.2 Model Type

Table 1 Deep Learning Model Families Used In Supporting Studies

Model Family	Reference	No. of Papers
LSTM & variants	[2, 7-9, 15-22]	12
GRU / plain RNN	[4, 10]	2
Hybrid / attention	[1, 11-12]	3
Feed-forward	[3, 13]	2
Deep Belief Network	[5]	1
RL trader	[14]	1
Ensemble voter	[6]	1

Table 1 summarises the deep learning model families identified across the 22 included studies. LSTM & variants (12/22): Appears to be the dominant choice for sentiment-enhanced forecasting, ranging from classical single-layer designs to ConvLSTM revenue models and nature-optimised LSTMs.

GRU / RNN (2/22): Often chosen to reduce training cost once text is pre-encoded (Chen, 2023; Kilimci & Duvar, 2020).

Hybrid / attention (3/22): Dual-stream or attention-gated architectures (Lai et al., 2021; Lien Minh et al., 2018; Wang et al., 2023) show sharper gains when multiple sentiment cues must be weighed simultaneously.

Feed-forward networks (2/22): Hu et al. (2021) notes their best configuration is a fully connected net-enriched with topic, semantic, and sentiment features, while Day and Lee (2016) uses a ReLU stack to classify Chinese-language finance headlines.

Other (1/22 each): Deep Belief Network (1), Reinforcement learning (1), Ensemble voter (1). Each represents a single-study exploration of less common approaches.

3.3 Model Evaluation Metrics

Accuracy is the most commonly reported metric for classification models, appearing in 13 out of the 22 studies (Table 2). Yet only six of those also report the F1-score or ROC-AUC, which are essential metrics for further model evaluation when data is imbalanced. Among regression-based studies, authors split evenly between RMSE and MAE, often only reporting one of them. Only five studies quote finance-related real-world metrics such as Sharpe ratio, ROI, or profit, and only one of those (Kabbani & Duman, 2022) attempts to determine the real-world viability of the model using a profit-based evaluation process.

Table 2 Distribution of Evaluation Metrics Across Reviewed Studies

Regression-based Metric	No. of Papers	Classification/Trading Metric	No. of Papers
RMSE	4	Accuracy	13
MAE	4	Precision	5
MSE	1	Recall	5
MAPE	1	F1-score	6
RMSPE	1	ROC-AUC	4
R^2	1	Specificity	1
		MCC	1
		Sharpe Ratio	1
		ROI / Profit	3
		BCE loss	1

Eleven of the thirteen classification-focused studies rely solely on raw accuracy as an evaluation metric without reporting F1-score, balanced accuracy, or ROC-AUC. This makes it difficult to assess the quality of the findings across different conditions. Only six studies (Garg, Pandey, & Tiwari, 2023; Hu et al., 2021; Kilimci & Duvar, 2020; Li et al., 2017; Mehta, Pandya, & Kotecha, 2021; Wang et al., 2023) mitigate this limitation by reporting either F1 or ROC-AUC.

Kabbani and Duman (2022); Khan et al. (2022); Shilpa and Shambhavi (2023) Are the only studies to translate predictive skill into a finance-aware evaluation metric such as Sharpe, ROI, or net profit. Among these, only Kabbani and Duman (2022) accounts for fixed transaction costs in the calculation. None of the regression studies report key financial risk metrics such as maximum draw-down, Sortino ratio, or turnover. This omission of information about extreme-loss behaviour, the concentration of negative outcomes, and the cumulative cost impact of trade frequency limits the conclusions that can be taken from these studies. Without such disclosures, impressive gains in

accuracy (median +7%) or decreases in RMSE (median -31%) remain interesting, but their practical commercial impact remains unclear without further exploration.

3.4 Model Performance

Across all analysed studies, the addition of a sentiment feature improves predictive performance (Table 3), but the scale and reliability of this improvement vary depending on model type, forecast timescale, and reporting practices.

Only three studies, Kabbani and Duman (2022); Khan et al. (2022); Shilpa and Shambhavi (2023) translate forecasts into finance-aware scores such as Sharpe ratio, ROI, or cash profit. Only Kabbani and Duman (2022) explicitly accounts for transaction costs. The remaining nineteen studies rely on statistical metrics such as accuracy or error rates, leaving important aspects of financial risk, such as extreme-loss behaviour, downside clustering, and transaction cost (fee) drag unexamined.

Eleven of the thirteen classification-based studies report only raw accuracy without accompanying metrics such as balanced-accuracy, F1-score, or ROC-AUC, making it difficult to assess whether these models outperform other models evaluated using a more robust combination of performance metrics.

Table 3 Impact of Sentiment Features on Model Performance

Paper (Author, year)	Target & horizon	Best reported metric[1]	Gain vs. price-only
Churi 2023	NIFTY 50 monthly	RMSE 0.158; MAE 0.018	−28 % RMSE
Kilic 2020	BIST 100 daily dir.	Accuracy 86.4 %	+12 ppt
Garg, 2023	NIFTY 50 daily dir.	Accuracy 96 %; ROC-AUC 0.995	+8 ppt
Hu, 2021	S&P 500 daily dir.	Accuracy 94 %	+11 ppt
Kabbani, 2022	US equities (RL)	Sharpe 3.14; Profit $115 k	+0.39 Sharpe
Lai, 2021	S&P 500 1-day dir.	Accuracy 66.3 %; AUC 0.76	+7.8 ppt
Li, 2017	CSI 300 daily dir.	Accuracy 87.9 %; AUC 94.7	+9.3 ppt
Wang, 2023	GEM Index daily dir.	Accuracy 71.6 %	+5.6 ppt

[1]Metric highlighted by the study authors as the principal result.

Accuracy metrics, while widely reported, capture just one facet of a model's effectiveness. Evaluating broader methodologies shows that model usage varies: hybrid/attention networks show the largest average performance increase, yet LSTM variants remain by far the most commonly deployed approach in the literature.

- Hybrid / attention (3 / 22): Dual-stream and attention-gated designs Lai et al. (2021); Lien Minh et al. (2018); Wang et al. (2023) report the highest median accuracy improvement at +9%.
- LSTM & variants (12 / 22): Classical or convolutional/cyclic LSTMs achieve a median accuracy gain of +6%.
- GRU / plain RNN (2 / 22): Lightweight models Chen (2023); Kilimci and Duvar (2020) provide a +4% median accuracy gain while reducing training cost.

- Feed-forward nets (2 / 22): Topic and sentiment-enriched FNNs Day and Lee (2016); Hu et al. (2021) lift accuracy by +3.8%.
- Deep-belief network (1): The SIWOA-optimised DBN Shilpa and Shambhavi (2023) improves median accuracy by +3.2%.
- Ensemble voter (1): The MLP ensemble Khan et al. (2022) reports a gain of +2.9% median accuracy.
- Reinforcement learning (1): Kabbani Kabbani and Duman (2022) reports Sharpe, not accuracy, and therefore is excluded from accuracy ranking.

Table 4 Horizon-Level Impact of Sentiment Features on Reported Metrics

Horizon	Metric	Median Gain	No. of Studies
Intraday ($\leq$1 day)	Accuracy	+5.2%	2
T+1 (daily)	Accuracy	+7.4%	10
5–10 days	Accuracy	+4.6%	3
Monthly	RMSE	−28%	2
Quarterly	RMSE	+12%	1

Sentiment features deliver the most consistent benefit at the next-day (T + 1) horizon (Table 4), with ten studies reporting a median uplift of +7.4% in accuracy. Performance gains are smaller for intraday forecasts and diminish further when the window extends to five or more days. This suggests that sentiment features are mostly indicative of short-term market momentum. While for monthly predictions, error drops sharply (-28% RMSE), though monthly predictions also often use smoothing techniques. Quarterly horizons see only mild improvement, suggesting most sentiment information has become irrelevant by the time the longer horizon is reached.

3.5 Dataset Sizes

The studies included in this review demonstrated substantial variability in dataset size and composition, reflecting differences in data availability, market focus, and design (Table 5). Across studies, dataset construction combined numerical stock market features (e.g., open, high, low, close prices and volumes) with sentiment-derived features extracted from either financial news, social media streams, or a blend of both. This diversity in scale is an important factor when interpreting reported performance metrics. Several studies leveraged large sentiment corpora. Li et al. (2017) utilised over 18 million forum posts from guba.eastmoney.com (January 2009 – October 2014) aligned with daily CSI300 data from a popular Chinese data platform, 'Wind Information'. Similarly, Wang et al. (2023) incorporated 7.7 million tweets alongside macroeconomic indicators and historical series covering 42 blue-chip stocks from June 2020 to June 2023. Other studies worked with moderately sized but richly annotated datasets. Hu et al. (2021) combined Reuters news headlines and bodies with S&P trading data (October 2006 – November 2013), applying a 70% training / 30% testing split. Mishev et al. (2019) built a dataset of 435 labelled revenue blocks from U.S. companies, again

using a 70/30 split. Time-segmented approaches were also common; Lien Minh et al. (2018) conducted their study based on a Two-Stream Gated Recurrent Unit Network, dividing data into training (2006–2012), validation (2013 H1), and testing (2013 H2). In contrast, studies focusing on specific securities often used shorter horizons. Liashenko, Kravets, and Plushchov (2024) analysed Tesla (TSLA) data spanning 30 September 2021 to 29 September 2022, while Mu et al. (2023) applied their framework to a one-month Google (GOOGL) trading window to simulate practical deployment. Other research, such as Qinghua and Wentao (2022), explicitly reported 1,312 rows of data (1,180 training and 132 validation) generated from their historical dataset.

Table 5 Dataset Size and Composition Across Included Papers

Study Title	Data Sources	Train/Test Split
Sentiment-aware stock market prediction: A deep learning method	18 million forum posts (guba.eastmoney.com) + CSI300 daily data (2009–2014)	Not specified
Stock Movement and Volatility Prediction from Tweets, Macroeconomic Factors and Historical Prices	7.7 million tweets + 42 blue-chip stocks (2020–2023)	Not specified
Stock Market Trend Forecasting Based on Multiple Textual Features: A Deep Learning Method	Reuters news + S&P historical data (2006–2013)	70 % training / 30 % testing
Forecasting Corporate Revenue by Using Deep-Learning Methodologies	435 labelled revenue blocks from US SEC filings (2010–2016)	70 % training / 30 % testing
Deep Learning Approach for Short-Term Stock Trends Prediction Based on Two-Stream GRU	Financial news and indicators (2006–2013)	Train: 2006–2012; Validation: 2013 H1; Test: 2013 H2
Stock Price Forecasting Using Sentiment Analysis of Stock Tweets	Tesla (TSLA) OHLCV data + Twitter (2021–2022)	Not specified
A Stock Price Prediction Model Based on Investor Sentiment and Optimized Deep Learning	Google (GOOGL) OHLCV data	One-month trading simulation
An Efficient Deep Learning Algorithm for Predicting Market Trends Using Cyclic LSTM	Historical OHLC sequences	1,180 train / 132 validation

3.6 Data Types

The included studies consistently integrated two categories of data: historical market data and sentiment-derived data. Precise features varied, with each study combining these data types in unique ways in an attempt to enhance predictive performance.

3.6.1 Historical Market Data

All studies incorporated numerical time-series data, typically drawn from public equity indices or individual stock records. Common features included:

- OHLCV data (open, high, low, close prices, and trading volume).
- Derived indicators such as moving averages (MA), exponential moving averages (EMA), relative strength indices (RSI), Bollinger Bands, and stochastic oscillators.
- Log-returns and percentage changes to stabilize variance and stationarize the input series.

Li et al. (2017) converted raw stock prices into log-returns and merged these with quarterly revenue growth labels. Similarly, Shilpa and Shambhavi (2023) experimented with input feature sets (Open/Close only; High/Low/Close; High/Low/Open/Close), identifying the four-feature set as optimal for RMSE reduction.

3.6.2 Sentiment Data

Sentiment features were derived from either financial news sources, social media platforms, or a blend of both:

- News sources: Hu et al. (2021) integrated Reuters titles and bodies, applying a VADER compound sentiment score and topic modeling with LDA. Day and Lee (2016) aggregated sentiment across NowNews, AppleDaily, Liberty Times Net, and MoneyDJ, creating lexicon-based polarity features.
- Social media sources: Li et al. (2017) processed over 18 million posts from guba.eastmoney.com, constructing daily sentiment indices aligned with CSI300 index data. Kilimci and Duvar (2020) classified tweets using the VADER toolkit, aggregating daily scores for Tesla (TSLA) predictions.
- Blended approaches: Kilimci and Duvar (2020) fuses Twitter streams with financial news headlines for BIST100 direction. Wang et al. (2023) combines tweets, macroeconomic indicators, and historical prices in a framework, and Lien Minh et al. (2018) integrates news text with market indicators via a GRU. Each study represents a different way in which text and prices can be blended and approached jointly.

Sentiment features were pre-processed using standard natural language processing pipelines (tokenization, stop-word removal, and vectorization), with some employing advanced embeddings, such as BERT (Othan & Kilimci, 2021), Word2Vec (Day & Lee, 2016), or domain-specific models like FinBERT (Chen, 2023; Kabbani & Duman, 2022). These were then integrated with market data either through concatenation layers or attention mechanisms within LSTM, GRU, or hybrid deep learning architectures.

3.7 Limitations of Current Metrics

There is a consistent overreliance on basic statistical metrics such as raw accuracy. Of the 22 included studies, 13 (59%) reported accuracy as the principal evaluation metric for classification models. However, only six studies (27%) also reported more

informative measures such as the F1-score or ROC-AUC. These supplementary metrics are essential in contexts where data is imbalanced, yet they were largely absent in most evaluations. Only 5 studies (23%) reported any real-world financial performance indicators, such as return on investment (ROI), realised profit, or the Sharpe ratio. Of these, only one study, Kabbani and Duman (2022), explicitly accounted for transaction costs in the performance evaluation. No study reported metrics such as drawdowns, trade turnover, cost-adjusted returns, or bid-ask spread costs despite their relevance in practical financial settings. The absence of domain-specific evaluation frameworks significantly limits the ability to assess whether a model's reported performance is meaningful in real-world trading environments. High accuracy or reduced error rates may appear promising in theoretical contexts, but without consideration of latency, market impact, or operational costs, these results cannot be interpreted as commercially viable. The widespread neglect of such finance-relevant metrics restricts the applicability of many published models and contributes to the disconnect between academic research and industry deployment. Only through comprehensive evaluation can the practical value of sentiment-enhanced prediction models be accurately assessed.

3.8 Intended Readers

The majority of included studies appear academically oriented, with few explicitly targeting industry implementation or financial audiences. For example, Hu et al. (2021) and Day and Lee (2016) are structured as theoretical contributions, focusing on comparative model performance and reporting metrics such as RMSE and AUC rather than finance-related metrics, namely, ROI or Sharpe ratio. A minority of papers - Kabbani and Duman (2022) - attempt to simulate real-world trading (e.g., reporting a $531 profit over eight trades). Even in these cases, real-time constraints, transaction costs, and deployment challenges are often acknowledged but not thoroughly addressed. This pattern suggests that while the research community is actively exploring sentiment-enhanced prediction models, the primary audience remains other researchers, with an emphasis on methodological innovation and benchmark accuracy. Industry-oriented considerations, such as regulatory constraints, model latency, and operational feasibility, are rarely the central focus. As such, there is a clear opportunity for future work to bridge this gap by translating academically robust models into commercially viable systems.

4 Discussion

This review indicates that incorporating sentiment-derived features into deep learning architectures enhances predictive performance over models relying solely on historical price data. Across the included studies, sentiment integration yielded consistent gains in classification accuracy, reductions in error metrics such as RMSE and MAE, and where accounted for, improvements in finance-aware measures such as Sharpe ratio and realised profit. These results are aligned with prior evidence in the wider machine learning literature, which has long suggested that market movements are not purely mathematical, but are driven by investor morale and collective sentiment signals (Baker & Wurgler, 2006; Hu et al., 2021; Li et al., 2017).

The predominance of LSTM-based approaches within the evidence base reflects trends seen in other systematic assessments of financial time-series forecasting, where recurrent neural networks are widely recognised for their capacity to capture temporal dependencies (Hu et al., 2021; Kilimci & Duvar, 2020; Mu et al., 2023). However, this review also highlights that alternative architectures such as attention-based GRUs (Lai et al., 2021) and hybrid frameworks combining convolutional and recurrent layers (Chen, 2023) often deliver incremental gains, particularly when tasked with extracting nuanced information from sentiment data. This is because the included sentiment features, particularly from informal or unstructured sources, are often context-sensitive and noisy. Attention mechanisms enable models to focus dynamically on the most relevant parts of the input sequence, while convolutional layers help extract local semantic patterns from textual data. These architectural adaptations appear better suited for capturing subtle sentiment signals than LSTM models alone, particularly in cases where linguistic nuance or topic relevance plays a critical role in market interpretation (Chen, 2023; Lai et al., 2021). While the improvements reported are promising, there are significant gaps in the quality and transparency of data across reviewed studies. Many papers did not fully document their data collection or preprocessing procedures, particularly for sentiment data scraped from social media platforms. A small number provided clear descriptions of noise reduction steps, sampling strategies, or lexicon validation. This lack of detail limits reproducibility and makes it difficult to assess the robustness of reported gains. In some cases, studies relied on single news providers or geographically narrow datasets, introducing potential sampling biases that may not generalise to broader markets. These issues mirror concerns explicitly acknowledged in several of the included studies. Li et al. (2017) noted that sentiment indices constructed from a single forum might not generalize across different market regimes. Shilpa and Shambhavi (2023) discussed how inconsistent preprocessing of historical features affected model stability, and Chen (2023) observed that reliance on limited news sources introduced variability in sentiment quality. Inconsistencies in data provenance and preparation, which are rarely addressed in studies reviewed, can substantially skew reported results. Few studies evaluate these models in settings that reflect real-world trading conditions. While accuracy and error metrics are widely reported, few studies, most notably Kabbani and Duman (2022), simulate trading outcomes. Fewer account for practical factors such as latency, transaction costs, slippage, or regulatory constraints. Studies predominantly focus on theoretical predictive performance without considering operational deployment challenges. The heterogeneity observed in data sources, time horizons, and performance metrics also complicates cross-study comparisons. Sentiment derived from financial news (Day & Lee, 2016; Hu et al., 2021) tends to capture macro-level events, while social media sources (Li et al., 2017) often reflect more immediate, micro-level investor reactions. This duality is echoed in broader sentiment analysis literature, which consistently finds that data provenance significantly shapes predictive utility. Performance improvements in the studies reviewed are most prominent at shorter forecast horizons (T+1 day) and tend to diminish for longer-term horizons, indicating that sentiment effects tend to be temporary rather than long-term. This pattern is consistent with prior work in behavioural finance. Baker and Wurgler (2006) demonstrate that investor sentiment

affects stock prices through a mispricing channel; however, the effect is stronger over short horizons and tends to revert as prices return to their fundamental values.

Although many studies report improvements when sentiment is included, these findings are not consistent and are often undermined by issues in data quality (Li et al., 2017), methodological transparency (Day & Lee, 2016; Hu et al., 2021), and real-world evaluation (Akita, Yoshihara, Matsubara, & Uehara, 2016; Churi et al., 2023; Garg et al., 2023; Khandelwal, Marfatia, & Venkataramanan, 2023; Lai et al., 2021; Liashenko et al., 2024; Lien Minh et al., 2018; Othan & Kilimci, 2021). Addressing these limitations, through clearer documentation of data pipelines, the use of diverse and transparent sentiment sources, and more rigorous evaluation under real-world conditions will be essential for advancing this field from proof of concept studies toward models with genuine commercial applicability.

4.1 Limitations

Despite the observed improvements in predictive performance, several limitations have been identified. The review reveals inconsistencies in data reporting, methodological transparency, and evaluation practices that hinder replication and undermine generalisability. A key concern is the widespread use of pre-trained sentiment models without domain-specific fine-tuning. Approximately 14% of the studies employed FinBERT as their primary sentiment encoder. While FinBERT is a strong baseline, it was trained primarily on financial English text in 2019. It may not generalise well to non-English markets or informal communication styles found in social media. Few studies assessed the appropriateness of these embeddings for their specific dataset, and none validated sentiment labels against human annotation. This reliance on pre-trained models without adaptation introduces potential noise and may limit the true contribution of sentiment features. There is also significant variation in data provenance and preprocessing documentation. Several papers lacked clear descriptions of how sentiment scores were generated or aligned with trading windows, making it difficult to assess internal validity. The absence of standardised evaluation pipelines contributes to this limitation: studies reported different time horizons, target variables (directional classification versus price prediction), and accuracy metrics, complicating any meta-analytic synthesis or model comparison. Most classification-based studies relied on raw accuracy as their sole performance measure, despite limitations in skewed datasets. This reliance is particularly problematic in stock market prediction tasks, where class distributions are often unbalanced. Upward movements may significantly outnumber downward ones. In such cases, a model can achieve high accuracy by predicting the majority class without capturing directional shifts. Without complementary metrics such as precision, recall, F1-score, or ROC-AUC, these results risk overstating predictive success and masking weaknesses in minority-class performance. Real-world considerations were largely absent. Aside from Kabbani and Duman (2022), none of the studies simulated end-to-end trading, incorporated latency, or adjusted for transaction costs. This limits the extent to which the results can be interpreted as commercially viable or actionable in applied finance contexts.

4.2 Implications

This review identifies several areas in which future research can enhance the robustness, relevance, and real-world applicability of sentiment-augmented deep learning models for stock market prediction. Future studies should prioritise transparency and reproducibility. Clear documentation of data sources, sentiment processing pipelines, and evaluation metrics would enable more reliable benchmarking. Black-box sentiment scores should be replaced with access to annotated datasets or code repositories where possible. Greater emphasis on aligning sentiment analysis tools with domain-specific contexts would enhance sentiment accuracy. Liashenko et al. (2024) struggled with prediction when using the VADER pre-trained lexicon, and Chen (2023) similarly struggled with long-form financial texts when using the FinBERT pre-trained model. Generic models like FinBERT and VADER, while popular, may not perform optimally across diverse financial markets, especially those using non-English content or informal investor communication (e.g., social media). Model evaluation should move beyond statistical accuracy toward finance-aware metrics. As demonstrated by Kabbani and Duman (2022), integrating evaluation methods such as Sharpe ratio, ROI, or maximum drawdown offers more realistic insights into potential trading value. Future studies should simulate end-to-end trading environments, including transaction costs, slippage, and liquidity constraints. To improve comparability and ensure real-world relevance, a standardised suite of evaluation metrics that includes both statistical and financial performance measures should be adopted. This should include accuracy and F1-score to address classification balance, alongside Sharpe ratio and maximum drawdown to capture financial risk and return characteristics. Additionally, bid-ask spread and transaction costs must be considered to assess model profitability. Table 6 notes the proposed suite of evaluation metrics, grouped by which assessment vector that they target.

Table 6 Proposed Suite of Evaluation Metrics

Assessment Category	Proposed Metrics
Model Accuracy	Accuracy + F1 / ROC-AUC
Value Prediction Performance	RMSE + MAE
Trading Performance	Sharpe Ratio, ROI, max drawdown
Cost Viability	Transaction fees, bid-ask spread
Operational Realism	Latency, data freshness

Real-time deployment remains underexplored. Only a handful of studies even consider latency or scalability in their design. To be commercially viable, predictive systems must ingest, process, and respond to both numerical and textual data streams in near real-time. Research focused on building modular, low-latency pipelines, event-driven architecture, or real-time data ingestion frameworks would offer significant practical benefit. While current research confirms the potential of sentiment-enhanced deep learning, the field must now turn toward resolving the challenges of interpretability, standardisation, and operationalisation. Bridging this gap will be essential

for translating academic interest to commercially relevant applications in financial forecasting.

Declarations

Funding

No funding was received to assist with the preparation of this manuscript.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

References

Akita, R., Yoshihara, A., Matsubara, T., Uehara, K. (2016). Deep learning for stock prediction using numerical and textual information. *2016 ieee/acis 15th international conference on computer and information science (icis)* (pp. 1–6).

Baker, M., & Wurgler, J. (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

Chen, D. (2023). Deep learning-based investor sentiment indicator construction and stock prediction research. *3rd international signal processing, communications and engineering management conference (ispcem)* (Vol. 8, pp. 31–35).

Churi, A., Chakraborty, D., Khatwani, R., Pinto, G., Shah, P., Sekhar, R. (2023). Stock price prediction using deep learning and sentiment analysis. *2nd international conference on futuristic technologies* (pp. 1378–1384).

Day, M., & Lee, C. (2016). Deep learning for financial sentiment analysis on finance news providers. *2016 ieee/acm international conference on advances in social networks analysis and mining (asonam)* (pp. 1127–1134).

Garg, D., Pandey, R.K., Tiwari, P. (2023). A systematic deep learning approach to forecast nifty50 index trend. *2023 11th international conference on intelligent systems and embedded design (ised)* (pp. 1–5).

Hu, Z., Wang, Z., Ho, S.B., Tan, A.H. (2021). Stock market trend forecasting based on multiple textual features: A deep learning method. *2021 ieee 33rd international conference on tools with artificial intelligence (ictai)* (pp. 1002–1007).

Kabbani, T., & Duman, E. (2022). Deep reinforcement learning approach for trading automation in the stock market. *IEEE Access*, *10*, 93564–93574, https://doi.org/10.1109/ACCESS.2022.3203697

Khan, W., Ghazanfar, M., Azam, M., Karami, A., Alyoubi, K., Alfakeeh, A. (2022). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing*, *13*, 3433–3456, https://doi.org/10.1007/s12652-020-01839-w

Khandelwal, V., Marfatia, P., Venkataramanan, V. (2023). Implementing various deep learning models to predict stock price movement. *4th international conference for emerging technology (incet).*

Kilimci, Z.H., & Duvar, R. (2020). An efficient word embedding and deep learning based model to forecast the direction of stock exchange market using twitter and financial news sites: A case of istanbul stock exchange (bist 100). *IEEE Access*, *8*, 188186–188198, https://doi.org/10.1109/ACCESS.2020.3029860

Lai, S., Jiang, H., Ye, C., Zhou, H. (2021). Chinese stock market prediction based on multifeature fusion and textcnn. *2021 international conference on service science (icss)* (Vol. 1844, pp. 279–285).

Li, J., Bu, H., Wu, J. (2017). Sentiment-aware stock market prediction: A deep learning method. *2017 international conference on service systems and service management* (pp. 1–6).

Liashenko, O., Kravets, T., Plushchov, V. (2024). Stock price forecasting using sentiment analysis of stock tweets. *2024 14th international conference on advanced computer information technologies (acit)* (Vol. 10, pp. 351–354).

Lien Minh, D., Sadeghi-Niaraki, A., Huy, H., Min, K., Moon, H. (2018). Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. *IEEE Access*, *6*, 55392–55404, https://doi.org/10.1109/ACCESS.2018.2868970

Mehta, P., Pandya, S., Kotecha, K. (2021). Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. *PeerJ Computer Science*, *7*, e476, https://doi.org/10.7717/peerj-cs.476

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L., Souma, W., Trajanov, D. (2019). Forecasting corporate revenue by using deep-learning methodologies. *2019 international conference on control, artificial intelligence, robotics & optimization (iccairo)* (pp. 115–120).

Mu, G., Gao, N., Wang, Y., Li, D. (2023). A stock price prediction model based on investor sentiment and optimized deep learning. *IEEE Access*, *11*, 51353–51367, https://doi.org/10.1109/ACCESS.2023.3278790

Othan, D., & Kilimci, Z.H. (2021). Stock market prediction with new generation deep contextualized word representations and deep learning models using user sentiments. *2021 international conference on innovations in intelligent systems and applications (inista)* (pp. 1–7).

Qinghua, Y., & Wentao, Z. (2022). An efficient deep learning algorithm for predicting market trends using cyclic long short-term memory network. *2022 ieee conference on telecommunications, optics and computer science (tocs)* (pp. 1436–1442).

Shilpa, B.L., & Shambhavi, R.B. (2023). Deep learning models for stock market prediction using optimization approach. *2023 international conference on network, multimedia and information technology (nmitcon)* (Vol. 1, pp. 1–6).

Wang, S., Bai, Y., Ji, T., Fu, K., Wang, L., Lu, C. (2023). Stock movement and volatility prediction from tweets, macroeconomic factors and historical prices. *2023 ieee international conference on big data (bigdata)* (pp. 1863–1872).