International Journal of Technology, Science and Engineering **ISSN: 2457-1016**

Vol. 8, Issue IV, Oct-Dec 2025 https://bharatpublication.com/ijtse/

News- and Social-Media-Driven Stock Movement Prediction Using NLP Feature Engineering and Supervised Learning

Shourya Gupta

University of Bath, United Kingdom

[1]*Received: 11 September 2025; Accepted: 09 October 2025; Published: 13 October 2025*

Abstract

Financial markets react not only to fundamentals but also to how information is framed, interpreted, and amplified. With the explosion of unstructured text from news outlets and social platforms, Natural Language Processing (NLP) has become a practical tool for extracting sentiment signals that can complement traditional price and volume features. This paper presents a structured exploration of sentiment-based stock prediction pipelines, comparing lexicon approaches with classical machine learning (ML) and modern transformer-based models, and discussing how these sentiment signals can be fused with time-series models to forecast price direction or returns. We synthesize methods across data acquisition, text pre-processing, sentiment modelling, feature engineering, and predictive learning, and we provide a comparative analysis of representative approaches reported in recent literature (2015–2025). Key findings are: (i) domain-specific sentiment models outperform general-purpose ones, particularly on finance-specific language; (ii) social sentiment can be predictive in event windows and high-attention periods but is noisy outside them; (iii) multimodal fusion (text + market data) often improves performance, but gains are sensitive to leakage control, labelling choices, and back testing rigor; and (iv) explain ability and privacy are increasingly central as sentiment models enter real trading and risk workflows.

1. Introduction

Price formation in liquid markets is strongly tied to information flow. Traditionally, this "information" was measured through structured variables: earnings surprises, macro releases, order flow, and accounting ratios. Today, investors digest a continuous stream of unstructured text: breaking news headlines, analyst commentary, company filings, and social media reactions. The key premise of sentiment-driven market prediction is simple: language reflects beliefs and expectations, and aggregated beliefs can influence demand, volatility, and short-horizon returns.

However, deploying sentiment signals is hard for four reasons:

Noise dominates: most posts are irrelevant, repetitive, or reactive rather than informative.

Finance language is specialized: words like "liability," "beat," "miss," "downgrade," or "guidance" carry domain meaning.

Time alignment is fragile: a model can look "amazing" if it accidentally learns from future information or misaligned timestamps.

Markets adapt: once a sentiment pattern is exploited, it often decays.

Despite these issues, a large body of research shows measurable relationships between textual sentiment and market behaviour, especially around attention spikes, announcements, and short event windows. For example, Twitter sentiment has been linked to abnormal returns around volume peaks, and studies in multiple markets have reported

---

[1] *How to cite the article:* Gupta S (2025); News- and Social-Media-Driven Stock Movement Prediction Using NLP Feature Engineering and Supervised Learning; *International Journal of Technology, Science and Engineering*; Vol 8 Issue 4; 5-14

mixed but informative causality patterns depending on context and data quality. On the news side, text-mining pipelines for market prediction have been explored in Forex headlines and beyond.

This paper contributes a practical, end-to-end view of sentiment-based stock prediction with a comparative analysis of methods from 2015–2025, with an emphasis on what tends to work, what breaks in real settings, and how to evaluate without fooling ourselves.

Table 1. Motivation and problem framing

Dimension	What it means in practice	Why it matters
Information sources	News, social microblogs, forums, filings	Different noise, latency, and "alpha half-life"
Sentiment granularity	Document, sentence, aspect, entity-level	Entity sentiment is closer to tradable signals
Horizon	Minutes to days	Short horizons are more sensitive to timing and leakage
Prediction target	Direction, returns, volatility	Impacts choice of labels and evaluation metrics
Market regime	Calm vs crisis vs earnings season	Signal strength can be regime-dependent

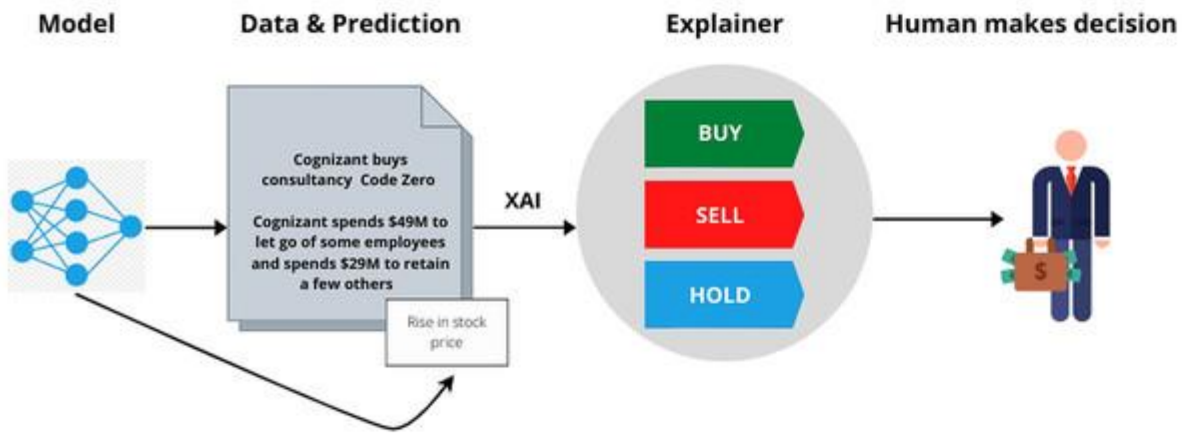

Figure 1: Working of XAI

2. Related Work and Background

Sentiment analysis in finance evolved through three broad phases:

2.1 Lexicon and rule-based finance sentiment

Early sentiment systems relied on dictionaries and heuristics (counts of positive/negative words). In finance, generic dictionaries often fail because common words flip meaning in context. This motivated finance-specific sentiment resources and evaluation frameworks, and later work compared lexicons against learned methods in financial text settings.

2.2 Classical ML with engineered features

A widely adopted pipeline is: tokenize text → compute n-grams / TF–IDF / topic features → train a classifier (SVM, logistic regression) to produce sentiment scores. In Decision Support Systems, Chan & Chong demonstrate a structured approach to financial-text sentiment with classical ML and domain features.

2.3 Deep learning and transformers

Deep neural models reduced dependence on handcrafted features. LSTMs became common for time-series prediction (often combining market features with learned representations. In parallel, transformers and domain-pretrained models improved sentiment extraction. A major milestone is FinBERT, designed for financial communications, which shows strong gains over dictionary methods and several ML baselines in finance text tasks. More recent directions include explainability review and multimodal / privacy-aware pipelines for prediction.

Table 2. Representative literature and what each adds

Study	Data source	Main method	Key takeaway
Ranco et al. (2015)	Twitter (DJIA)	Event-study + sentiment	Sentiment during volume peaks relates to abnormal returns
Nassirtoussi et al. (2015)	News headlines (Forex)	Semantics + sentiment + DR	Headline text can forecast near-term direction
Sul et al. (2017)	Twitter / social	Attention + sentiment	Attention dynamics shape predictive power
Chan & Chong (2017)	Financial text	ML sentiment framework	Domain-aware features improve sentiment quality
Fischer & Krauss (2018)	Price time series	LSTM	Deep sequential models can beat baselines (with caveats)
Jiao et al. (2020)	Social media	Sentiment indices	Social sentiment affects market dynamics under conditions
Mishev et al. (2020)	Multiple datasets	Lexicon→transformer eval	Transformers generally outperform lexicons in finance
Hamraoui & Boubaker (2022)	Twitter + market	Correlation/Granger	Relationship varies; overall effect can be weak in broad samples
Gong et al. (2022)	News (oil)	NLP features + ML	Text can complement ML forecasting features
Huang et al. (2023)	Financial comms	FinBERT	Domain pretraining yields large sentiment gains
Todd et al. (2024)	Literature	Review	Best practices and pitfalls for finance sentiment
Ruan & Jiang (2025)	Text + indicators	FinBERT + SHAP + DP	Trend toward explainable + privacy-aware prediction

3. Data Sources and Problem Formulation

3.1 Text data: news vs social media

News tends to be more curated, with clearer entity references and lower spam. It often has stronger informational content but may be priced in quickly due to high market efficiency.

Social media is faster and more emotional, capturing retail attention and narrative shifts. But it is noisier, vulnerable to bots, and often reflects reaction to price moves rather than causes.

3.2 Market data and alignment

Most pipelines also ingest OHLCV (Open-High-Low-Close-Volume), corporate actions, sector indices, and sometimes volatility proxies. Alignment choices are critical:

Timestamping: publication time vs ingestion time

Trading calendars: market open/close, after-hours news

Windowing: aggregating sentiment in rolling windows (e.g., 15 min, 1 hr, 1 day)

3.3 Prediction targets

Common targets include:

Direction: sign of return over horizon (h) (classification)

Return: continuous return (r_{t,t+h}) (regression)

Abnormal return: market-adjusted return (event studies)

Volatility: realized volatility or GARCH-like proxies

Table 3. Data source characteristics

Property	News	Social media
Latency	Medium (minutes)	Low (seconds)
Noise level	Lower	Higher (spam/bots)
Entity clarity	Higher	Mixed
Emotion/attention	Medium	High
Typical use	Event-driven prediction	Attention + narrative indicators

4. Methodology: End-to-End Pipeline

A practical sentiment-to-price prediction system usually has two layers:

Sentiment extraction model (text → sentiment score)

Market prediction model (sentiment + market features → forecast)

4.1 Text pre-processing

Steps commonly include:

Language filtering, duplicate removal

Entity recognition and ticker linking (e.g., “Apple” → AAPL)

Spam/bot filtering for social posts

Handling sarcasm and emojis (hard but important for social)

4.2 Sentiment modeling approaches

(A) Lexicon-based: score = (positive − negative) / length
Pros: interpretable, fast. Cons: weak context, domain mismatch.

(B) Classical ML: TF–IDF → logistic regression/SVM
Pros: strong baseline, cheap. Cons: brittle across regimes, vocabulary drift.

(C) Transformers / domain pretraining: FinBERT-like models
Pros: context-aware, finance language understanding; strong on benchmarks.

Cons: heavier compute, needs careful fine-tuning and evaluation.

4.3 Feature engineering for market prediction

Common sentiment features:

Mean, median, max sentiment in window

Volume-weighted sentiment (more posts = more “attention”)

Sentiment momentum: (S_t - S_{t-1})

Polarity imbalance: (#pos - #neg)

Event indicators: earnings day, macro release day

4.4 Prediction models

Linear/logistic regression (strong baselines)

Tree ensembles (XGBoost/LightGBM)

Sequential models (LSTM/GRU) for time dependencies

Hybrid fusion: transformer sentiment embeddings + time-series model

Explainability: SHAP on final predictors (growing emphasis)

Table 4. Method choices and tradeoffs

Layer	Option	Strength	Weakness	Best for
Sentiment	Lexicon	Fast, explainable	Context-blind	Quick monitoring
Sentiment	SVM/LogReg	Strong baseline	Drift-sensitive	Small/medium data
Sentiment	FinBERT	Best accuracy in finance text	Compute + tuning	Production-grade sentiment

Layer	Option	Strength	Weakness	Best for
Prediction	Linear	Stable baseline	Limited nonlinearity	Risk-controlled signals
Prediction	Boosted trees	Handles nonlinear mix	Overfit risk	Tabular fusion
Prediction	LSTM	Captures temporal patterns	Harder to debug	Sequential features

5. Experimental Design and Evaluation

5.1 Dataset construction (typical setup)

A realistic dataset often includes:

Text items with timestamps and mapped tickers

Aggregated sentiment features per ticker per time window

Market features at time (t) (and lags)

Labels for (t \to t+h)

5.2 Leakage control (most common failure point)

Three frequent leakage traps:

Using sentiment computed from text posted after the prediction timestamp.

Aggregating features with windows that overlap the label horizon.

Training and testing across overlapping time windows (temporal leakage).

Use strict chronological splits and embargo periods.

5.3 Metrics

Classification: Accuracy, F1, AUC, MCC

Regression: MAE, RMSE, directional accuracy

Trading metrics (if backtesting): Sharpe, max drawdown, turnover, transaction costs

5.4 Comparative study setup

A clean comparison holds constant:

Same time splits

Same labelling rules

Same feature windows

Only swap sentiment model or fusion model

Table 5. Evaluation checklist

Item	Good practice	What goes wrong if ignored
Time split	Walk-forward / rolling	Inflated performance
Embargo	Gap between train/test	Leakage via overlap
Costs	Include realistic slippage	Paper profits only
Stability	Test across regimes	Strategy dies live
Ablations	Remove components	“Black box” improvement claims

6. Results and Comparative Analysis

This section synthesizes what recent work commonly reports, and why results differ across settings.

6.1 News-only vs social-only

News sentiment tends to be more stable but can be priced rapidly.

Social sentiment often shows predictive value during attention spikes and event windows, consistent with event-based findings, but can be weak in broad samples depending on market and filtering.

Studies emphasize attention as a moderator: sentiment matters more when more people are watching.

6.2 Lexicon vs ML vs transformers

Across finance sentiment evaluations, transformer-based approaches generally outperform lexicons and older ML feature pipelines. Domain pretraining is a major driver; FinBERT-style models show strong advantages in financial language tasks.

6.3 Fusion models (text + market data)

Fusing sentiment with technical indicators can improve predictive accuracy, but gains vary. Recent pipelines emphasize explainability and privacy-aware handling of textual data. In commodities, textual features can be complementary to ML predictors.

6.4 Practical interpretation of “predictability”

Even when statistical metrics improve, trading profitability may vanish after costs, especially in highly efficient large-cap equities. This is why rigorous evaluation and realistic assumptions matter.

Table 6. Comparative analysis summary (typical patterns in literature, 2015–2025)

Comparison	Common outcome	Why
News vs social	Social better in attention spikes; news steadier	Attention amplification vs curated info
Lexicon vs ML	ML beats lexicon	Better handling of domain terms and negations

Comparison	Common outcome	Why
ML vs transformers	Transformers usually best	Context + domain pretraining
Text-only vs fused	Fused often better	Text complements market microstructure
Static vs regime-aware	Regime-aware more robust	Market behavior changes over time

7. Discussion: What Works, What Breaks

7.1 Why sentiment sometimes "predicts"

There are realistic mechanisms:

Slow diffusion of information to all market participants

Behavioral biases and herding

Retail-driven narrative cycles

Liquidity and attention constraints

Work linking sentiment to market dynamics supports the idea that sentiment can have measurable effects under certain conditions.

7.2 Why it often fails in production

Non-stationarity: language and platform behavior drift

Adversarial behavior: coordinated posting, pump-and-dump

Selection bias: training on popular tickers only

Overfitting: too many features vs limited effective samples

Timing reality: text ingestion delays, API rate limits

7.3 Explainability and governance

As sentiment models move into decision support, interpretability is becoming a requirement. Reviews highlight best practices and common pitfalls for finance sentiment measurement. Newer frameworks explicitly combine prediction with explainability (e.g., SHAP) and even differential privacy to reduce sensitive leakage risk.

Table 7. Deployment risks and mitigations

Risk	Example	Mitigation
Bots/spam	Artificial sentiment spikes	Bot detection, account trust scoring
Drift	New slang, new narratives	Periodic fine-tuning, monitoring
Latency	Late news ingestion	Timestamp audits, delay-aware features
Overfit	Great backtest, poor live	Walk-forward validation, simpler models

Risk	Example	Mitigation
Leakage	Future text in features	Strict cutoffs + reproducible pipelines

8. Conclusion

Sentiment analysis for financial prediction is no longer a novelty; it is a serious feature engineering and modelling problem where the biggest wins come from (1) domain-aware sentiment extraction, (2) tight time alignment, and (3) honest evaluation. The comparative picture from 2015–2025 shows a clear trajectory: lexicons are useful for transparency and quick monitoring, classical ML remains a strong baseline, and transformers (especially finance-pretrained models like FinBERT) deliver the best sentiment quality and often better downstream prediction. Still, predictability is conditional: it is strongest in event windows, attention spikes, and specific market regimes, and it can degrade quickly when widely exploited.

References

Chan, S. W. K., & Chong, M. W. C. (2017). Sentiment analysis in financial texts. *Decision Support Systems, 94*, 53–64. https://doi.org/10.1016/j.dss.2016.10.006

Fischer, T., & Krauss, C. (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

Gong, X., Guan, K., & Chen, Q. (2022). The role of textual analysis in oil futures price forecasting based on machine learning approach. *Journal of Futures Markets, 42*(10), 1987–2017. https://doi.org/10.1002/fut.22367

Hamraoui, I., & Boubaker, A. (2022). Impact of Twitter sentiment on stock price returns. *Social Network Analysis and Mining, 12*(1), 28. https://doi.org/10.1007/s13278-021-00856-7

Huang, A. H., Wang, H., Yang, Y., & Uy, M. C. S. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research, 40*(2), 806–841. https://doi.org/10.1111/1911-3846.12832

Jiao, P., Veiga, A., & Walther, A. (2020). Social media, news media and the stock market. *Journal of Economic Behavior & Organization, 176*, 1–19. https://doi.org/10.1016/j.jebo.2020.03.002

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access, 8*, 131662–131682. https://doi.org/10.1109/ACCESS.2020.3009626

Nassirtoussi, A. K., Aghabozorgi, S., Teh, Y. W., & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment. *Expert Systems with Applications, 42*(1), 306–324. https://doi.org/10.1016/j.eswa.2014.08.004

Nguyen, T. H., & Shirai, K. (2015). Topic modeling based sentiment analysis on social media for stock market prediction. *Expert Systems with Applications, 42*(24), 9603–9611. https://doi.org/10.1016/j.eswa.2015.07.052

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The effects of Twitter sentiment on stock price returns. *PLOS ONE, 10*(9), e0138441. https://doi.org/10.1371/journal.pone.0138441

Ruan, L., & Jiang, H. (2025). Stock price prediction using FinBERT-enhanced sentiment with SHAP explainability and differential privacy. *Mathematics, 13*(17), 2747. https://doi.org/10.3390/math13172747

Sul, H. K., Dennis, A. R., & Yuan, L. I. (2017). Trading on Twitter: Using social media sentiment to predict stock returns. *Decision Sciences, 48*(3), 454–488. https://doi.org/10.1111/deci.12229

Todd, A., Smales, L. A., & Zhang, B. (2024). Text sentiment analysis in finance: A survey of methods and applications. *International Review of Financial Analysis, 94*, 103284. https://doi.org/10.1016/j.irfa.2024.103284