

# 異常値を含むデータに対する文脈情報統合の有効性に関する研究 — 株価予測を題材とした検証 —

内田陽太・浦崎華瑠

有明工業高等専門学校 創造工学科 情報システムコース

Gauthier Lovic 研究室

2026年1月23日

## 1 はじめに

時系列予測は金融・需要予測・設備保全などで重要である一方、数値特徴量のみでは外部要因に起因する急変動を捉えにくい。本研究は株価日足を題材に、ニュース（文脈情報）を統合したときの汎化性能の変化を、特に急変動日に着目して検証する。

## 2 手法

### 2.1 比較設計（フェーズ1/2）

本研究では二段階で比較を行う。フェーズ1では数値特徴量のみで4モデル（Logistic Regression / LightGBM / LSTM / Transformer）を比較し、フェーズ2では同一の分割・指標のままニュース特徴量を追加して再評価する<sup>5, 6, 7)</sup>。

### 2.2 データとニュース特徴量

対象銘柄はトヨタ自動車（7203.T）で、株価・市場要因はyfinanceから取得する<sup>3)</sup>。予測対象期間（2013–2025年）を一貫した方式でカバーするため、ニュースの取得には全期間にわたりGDELT GKG（日次ファイル）を用いる<sup>1, 2)</sup>。

ニュース特徴量には、GKGの構造化情報からV2Tone（Tone, Positive, Negative, Polarity, ActivityRefDensity, Self-GroupRefDensity, WordCount）を用いる。V2Toneは感情辞書に基づく指標で、Tone = Positive[%] – Negative[%]（平均トーン）、Polarity = 感情語の出現割合[%]（Toneが小さくても高い場合は賛否が拮抗して感情強度が高いことを示唆）、ActivityRefDensity = 能動語[%]、SelfGroupRefDensity = 有名詞[%]、WordCount = 総語数である。

### 2.3 ニュースの取引日アライン（リーケ対策）

公開時刻の不整合を避けるため、非取引日のニュースは次の取引日に繰り上げ、引け後ニュースも翌取引日に繰り上げる（境界時刻はJST 15:30）。さらにGKGは時刻情報が粗くUTC 00:00に丸められたレコードとして取得され得るため、安全側に全件を「引け後扱い（翌営業日に繰り上げ）」としてアラインし、将来にしか分からぬ情報の混入（リーケ）を抑制する。

### 2.4 ラベル定義と評価枠組み

取引日 $t$ 、ホライズン $H=1$ として、将来リターン $r_H(t)=\frac{P_{t+H}}{P_t}-1$ を用いる。当日ボラ $\sigma_{20}(t)$ から閾値 $\tau_t(H)=k_\tau\sigma_{20}(t)\sqrt{H}$  ( $k_\tau=0.3$ )を定義し、 $r_H(t) > \tau_t(H)$ をup、 $r_H(t) < -\tau_t(H)$ をdown、それ以外をneutralとする。さらに $a_t=|r_H(t)|$ の上位 $p=0.10$ を急変動日集合（subset）として評価する。

分割は時系列リーケを抑えるためPWFE分割を用いる<sup>4)</sup>。評価は翌営業日方向性の三値分類とし、 $|r_H|$ 上位10%を急変動日集合（subset）として、通常日（main）と分けてmacro-F1、balanced accuracy、混同行列を報告する。入力窓は $T=60$ で、ラベル日 $t$ 当日の特徴 $x_t$ は入力に含めず直前 $T$ 日から予測する。

## 3 結果

結果はPWFE外側検証の予測を結合したpooled集計で示される。mainとsubsetで傾向が異なり、mainではLogistic Regression/LightGBMが改善した一方、subsetではLightGBMが相対的に改善した。評価集合のクラス分布を表1に示す。

表1：評価集合のクラス分布（pooled）

集合	down	neutral	up
main (pooled, N = 1982)	675 (34.1%)	566 (28.6%)	741 (37.4%)
subset top10% (pooled, N = 196)	92 (46.9%)	0 (0.0%)	104 (53.1%)

main：Logistic RegressionとLightGBMはニュース統合で改善し、macro-F1はそれぞれ0.332172→0.353197、0.329829→0.349798となった。一方、LSTMとTransformerは改善せず（0.334890→0.325760、0.341312→0.333113）であった。この比較結果を表2に示す。

表2：main（通常日）における性能比較（pooled）

モデル	フェーズ1（数値のみ）		フェーズ2（ニュース統合）	
	macro-F1	bal. acc.	macro-F1	bal. acc.
Logistic Regression	0.332172	0.339152	<b>0.353197</b>	<b>0.356976</b>
LightGBM	0.329829	0.330392	<b>0.349798</b>	<b>0.350153</b>
LSTM	<b>0.334890</b>	<b>0.338991</b>	0.325760	0.336064
Transformer	<b>0.341312</b>	<b>0.343847</b>	0.333113	0.333511

subset（急変動日）：LightGBMは改善した一方、他モデルは低下または改善が限定的である。なおsubsetは定義上neutralが欠損し得るため、macro-F1やbalanced accuracyは不安定になり得る点に留意し、混同行列と併せて解釈する必要がある。この比較結果を表3に示す。

表3：subset（急変動日）における性能比較（pooled）

モデル	フェーズ1（数値のみ）		フェーズ2（ニュース統合）	
	macro-F1	bal. acc.	macro-F1	bal. acc.
Logistic Regression	<b>0.312461</b>	<b>0.429139</b>	0.284436	0.357232
LightGBM	0.295167	0.372910	<b>0.318668</b>	<b>0.410117</b>
LSTM	<b>0.256121</b>	<b>0.321279</b>	0.219676	0.259615
Transformer	0.286105	0.378344	<b>0.296148</b>	<b>0.380644</b>

## 4 まとめ

卒業論文の実験設定（PWFEによるリーケ抑制、GKG日次の安全側アライン、main/subsetの二系統評価）に基づき、ニュース統合の効果を比較した。その結果、mainではLogistic RegressionとLightGBMが小幅に改善した一方、LSTM/Transformerでは改善が見られなかった。subsetではLightGBMが相対的に改善し、日次集計型のニュース特徴量は木モデルで活用されやすい可能性が示唆された。一方で、リーケを避けるための安全側アライン（翌営業日へ繰り上げ）は、実際には当日中に織り込まれたニュースも翌日に回すため、信号の即時性を希釈し得る（リーケ耐性と信号強度のトレードオフ）。

## 参考文献

- 1) GDELT Project, *The GDELT Project (official site)*. <https://www.gdeltproject.org/>.
- 2) GDELT Project, *GKG / V2Tone Codebook* (公式コードブック). <https://www.gdeltproject.org/data.html>.
- 3) R. Ran Aroussi, *yfinance: Yahoo! Finance market data downloader (GitHub)*. <https://github.com/ranaroussi/yfinance>.
- 4) M. López de Prado, *Advances in Financial Machine Learning*. Wiley, 2018.
- 5) G. Ke, Q. Meng, T. Finley, et al., “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in *Proc. NeurIPS*, 2017.
- 6) S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- 7) A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention Is All You Need,” in *Proc. NeurIPS*, 2017.