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（文脈依存の異常局面）を説明しにくい．本研究は株価日足を題材に，ニュース（文脈情報）を特徴量として統合したとき，汎化性能が（特に急変動日で）どう変化するかを検証する．本質的な難点は「ニュースを使うこと」そのものではなく，(i) ニュースを取引日へ安全に対応付け，将来にしか分からない情報が誤って学習データに混ざらないようにすること，(ii) 急変動局面で効いたかを測れる評価枠組みを用意することにある．

2　手法

二段階比較：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 [5, 6, 7]．

対象データ：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 [3]．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 [1, 2]．

ニュース特徴量：GKG の構造化情報から V2Tone（Tone, Positive, Negative, Polarity, ActivityRefDensity, SelfGroupRefDensity, 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 =代名詞 [%]，WordCount =総語数である．

ニュースの取引日アライン（リーク対策）：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，将来にしか分からない情報の混入（リーク）を抑制する．

ラベル（三値分類）と subset：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．

分割と評価：時系列リークを抑えるため PWFE 分割を用いる [4]．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 $\boldsymbol{x}_t$ 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 が改善する一方，subset では LightGBM が相対的に改善した．

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．

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	0.353197	0.356976
LightGBM	0.329829	0.330392	0.349798	0.350153
LSTM	0.334890	0.338991	0.325760	0.336064
Transformer	0.341312	0.343847	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．

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	0.312461	0.429139	0.284436	0.357232
LightGBM	0.295167	0.372910	0.318668	0.410117
LSTM	0.256121	0.321279	0.219676	0.259615
Transformer	0.286105	0.378344	0.296148	0.380644

4　まとめ

卒業論文の実験設定（PWFE によるリーク抑制，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，リーク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（リーク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.