

Financial Text と Transformer を用いた転換点検出手法の提案 Proposal for Turning Point Detection Method using Financial Text and Transformer

田口 怜^{1*} 渡邊 光² 坂地 泰紀¹ 和泉 潔¹ 平松 賢士³
Rei Taguchi¹ Hikaru Watanabe² Hiroki Sakaji¹ Kiyoshi Izumi¹ Kenji Hiramatsu³

¹ 東京大学大学院工学系研究科

¹ Graduate School of Engineering, The University of Tokyo

² 東京大学工学部

² Faculty of Engineering, The University of Tokyo

³ 株式会社アイフィスジャパン

³ IFIS Japan Limited.

Abstract: 本研究では、アナリストの個別銘柄に対するセンチメントが、株式市場分析に役立つかを
実証する。これはアナリストレポートに自然言語処理を使用して極性指標を作成することで実現可能と
なる。本研究では、作成した極性指標に対し、異常検知アルゴリズムを用いて異常スコアを算出した。結
果、本研究で提案した手法が極性指標の転換点検出において有効であることが分かった。

1 はじめに

近年、人工知能の爆発的な発展により、オルタナティブデータの活用が世界的に注目されている。特に金融・経済分野での普及が顕著であり、伝統的な財務データに加えて、景気予測や投資戦略など幅広く活用され始めている。日本においては、オルタナティブデータ推進協議会 (JADAA)*¹等の業種間を超えた団体が発足し、オルタナティブデータの積極的な普及を推進している。そのなかでも特にテキストデータは汎用性が高く、筆者の書きぶりや内容の偏向から予測や分類などの抽出が可能であることがわかっている [1, 2]。また、テキストデータの中でもアナリストレポートの情報価値は非常に高いと考えられている [5]。アナリストレポートから作成した極性指標が為替や普通国債等発行額などのマクロ経済指標に対し先行性を持つことがわかっており [12]、市場分析への応用が期待されている。本研究では、アナリストレポートを使用し、個別銘柄の将来に対するセンチメントが市場分析に役立つことを実証する。期待される成果は、本研究で提案した手法が極性指標の転換点検出において

有効であると示すことである。本研究の貢献は、アナリストレポートから作成した極性指標を使用して、新たな市場分析フレームワークにおける基盤技術を開発したことである。

2 先行研究

機械学習による極性指標作成の研究として、次のような研究がある。余野ら [3] では、金融経済月報を Long Short Term Memory(LSTM) で学習し、Latent Dirichlet Allocation(LDA) を用いて個人消費、設備投資、在庫などのトピックごとの指標を作成した。また、各トピックのセンチメントが全体に対してどの程度寄与しているかを求めることで、どのマクロファクターの影響力が高いかを分析している。山本ら [6] では、景気ウォッチャー調査を LSTM で学習し、極性指標を作成し、指数間の相関係数を算出している。

上記の研究では、金融経済月報や景気ウォッチャー調査といった内容が類似したテキストを用いているのに対し、我々は、アナリストレポートを用いて、その個別銘柄のセンチメントを総計して指標とした点で異なる。アナリストレポートを用いたテキストマイニングの研究事例として、次のような研究がある。平松ら [5] では、株価はレポートのセンチメントに強く反応しているのみならず、レポート発行後に株価のドリフトも観測しており、

* 連絡先： 東京大学大学院工学系研究科
〒 113-8656 東京都文京区本郷 7-3-1
E-mail: s5abadiee@g.ecc.u-tokyo.ac.jp

*¹ <https://alternativedata.or.jp>

資産運用実務においてもアナリストレポートのテキスト情報が有益であることが示唆された。Suzuki et al.[6]では、LSTM 等を使用しアナリストレポートから意見文と非意見文を抽出し、純利益と株価の予測を行っている。加えて、小林ら [7] は、独自の深層学習モデルを用いてアナリストレポートからアナリスト予想根拠文の抽出を行っている。

一方で、我々が、BERT を用いてアナリストレポートを学習させている点も他の研究と異なる。BERT を使用した研究事例として、Hiew et al.[8]がある。Hiew et al.[8]では、テキストデータに中国の SNS である weibo を利用し、BERT で極性を付与し、LSTM を用いて株価予測を行っている。

加えて、我々が、フレームワークの一端に Transformer と異常検知アルゴリズムを使用した点も他の研究と異なる。金融分野で Transformer を使用した研究例として、Ding et al.[9]がある。Ding et al.[9]では、Transformer とマルチスケールガウシアン事前分布を導入し、株価予測を行った。また、金融分野で異常検知アルゴリズムを使用した研究例として、Ahmed et al.[10]がある。Ahmed et al.[10]では、オーストラリア証券取引所に上場している株式に対し、Local Outlier Factor などのアルゴリズムを使用し、転換点検出を行った。

3 提案手法

本研究では、以下の 3 ステップにおいて、実証分析を行う。

Step1: BERT を用いてアナリストレポートの文章を学習し、「ポジティブ」「ネガティブ」「ニュートラル」の三値に極性分類を行う。

Step2 : 作成した極性指標（全業種・業種別）に対し、Time2Vec+Transformer を用いて学習を行う。

Step3 : Step2 のアウトプットに対し、ホテリングの T^2 法を用いて異常スコアを作成する。加えて、本実験で得られた結果を集計する。

また、本研究におけるアーキテクチャーは図 1 のように記述できる。

3.1 Bidirectional Encoder Representations from Transformers (BERT) を使用した極性分類手法

本セクションの内容は、セクション 3 の Step1 に対応する。BERT は、Devlin et al. [11] により提案された Transformer ベースの言語モデルである。BERT では、

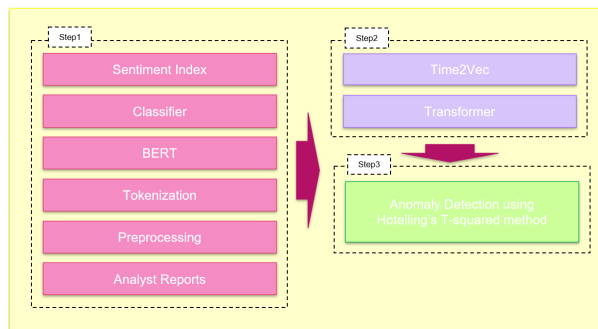


図 1 本研究のアーキテクチャー

入力内のいくつかのトークンをマスキングし、マスキングされた単語を予測することで、うまく双方向の学習を可能にしている。本研究では、日本語 BERT の学習済みモデルとして、東北大学の乾・鈴木研究室が公開しているモデル*2を利用した。本研究では、Taguchi et al.[12]の方法を用いて、BERT を用いて二種類のモデル (UP Model, DOWN Model) 作成する。Taguchi et al.[12]の手法では、レーティングが「買い」に変化したレポートを 1, 変化のないレポートを 0, 「売り」に変化したレポートを -1 とする。最終判定の方法は、表 1 を参照。本手法の詳細は Taguchi et al.[12] を参照。また、本アーキテクチャーは Sakaji et al.[13] のモデルを BERT に置き換えたものである。算出結果についてはセクション 5 を参照。

表 1 極性判定手法

UP Model 判定	DOWN Model 判定	最終判定
0: Neutral	-1: Negative	-1: Negative
1: Positive	0: Neutral	1: Positive
1: Positive	-1: Negative	尤度が高い方
0: Neutral	0: Neutral	0: Neutral

3.2 Transformer+Time2Vec

本セクションの内容は、セクション 3 の Step2 に対応する。本セクションでは、Transformer+Time2Vec を使用した学習手法について述べる。本手法では、セクション 3.1 で作成された極性指標を、Time2Vec でベクトル化し、その Output を Transformer の Input としている。

まず、Transformer に時間の概念を Encode するため、Time2Vec を用いる。Time2Vec は、Kazemi et al.[14]

*2 <https://github.com/cl-tohoku/bert-japanese>

によって提案された。Time2Vec は、以下で定義される。

$$\mathbf{t2v}(\tau)[i] = \begin{cases} \omega_i \tau + \varphi_i, & \text{if } i = 0 \\ \mathcal{F}(\omega_i \tau + \varphi_i) & \text{if } 1 \leq i \leq k \end{cases} \quad (1a)$$

$$(1b)$$

時間スカラー τ に対する Time2Vec は、 $\mathbf{t2v}$ と定義される。 $\mathbf{t2v}[i]$ は、 i における $\mathbf{t2v}$ の要素である。 \mathcal{F} は、周期性を持つ活性化関数である。 ω_i と φ_i は学習可能なパラメータである。

次に、学習部分に Transformer を利用する。Transformer は Encoder-Decoder Model をベースとしており、Self-Attention と Feed Forward Neural Networks が組み込まれている。Transformer は、Vaswani et al.[15] によって提案された。また、Transformer の系列データへの応用は Wu et al.[16] で提案されている。

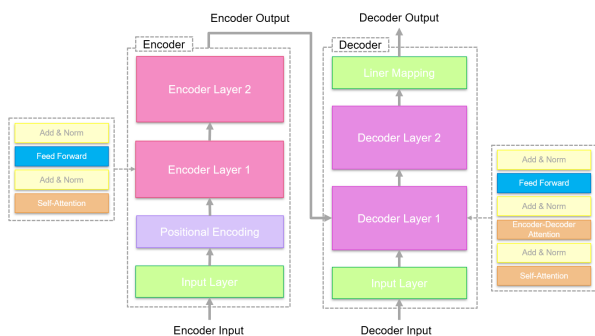


図2 Transformer のアーキテクチャー

時系列を扱う Transformer のアーキテクチャーは図2で表現できる。Transformer 及び比較手法の評価には、MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), MSLE (Mean Squared Logarithmic Error) を用いる。提案手法に Transformer を選択した理由についてはセクション5を参照。各評価指標の定義は付録Aを参照。

3.3 ホテリングの T^2 法

本セクションの内容は、セクション3の Step3 に対応する。異常検知における異常スコアの算出には、ホテリングの T^2 法を用いる。セクション3.2の Output にホテリングの T^2 法を当てはめることで異常スコアを算出する。転換点検出部分に、ホテリング理論を用いた理由は、平均・分散などデータの基本的な分布情報を用いるため、ChangeFinder[17]をはじめとする変化点検知アルゴリズムで発生するパラメータの経験的決定問題を一定程度回避できると考えたからだ。 T^2 法は、Hotelling[18]

により提案された異常検知アルゴリズムである。データが $\mathcal{D} = \{x_1, \dots, x_n\}$ で与えられ、正規分布に従うとする。これよりデータ \mathcal{D} の標本平均 $\hat{\mu}$, 分散共分散行列 $\hat{\Sigma}$ が推定できる。この場合、観測値 x' に対する異常スコア $a(x')$ は以下で定義される。

$$a(x') = (x' - \hat{\mu})^T \hat{\Sigma}^{-1} (x' - \hat{\mu}) \quad (2)$$

本研究における閾値は、検証データの最大値をもとに設定する。加えて異常検知の性能評価には ROC (Receiver Operating Characteristic) Curve を用いた。曲線下の AUC (Area Under the Curve) は0から1の範囲をとり、0.7以上であればモデルの性能が高いとされている。なお、ROC Curve による評価に必要な正解ラベルの付与は著者が手動で行った*3。

4 実験

BERT 極性指標作成には、アナリストレポートデータ (373,050 件) を用いる。データクレンジングには IFIS 社の技術を使用した。データの期間はいずれも 2010/1 ~ 2020/12 とする。実験では、全業種極性指標と業種別極性指標をそれぞれ作成する。極性指標の作成方法についてはセクション3.1を参照。全業種極性指標は、業種別極性指標の出力 (1, 0, -1) をすべて足し合わせたものとする。時系列深層学習モデルの選択には5つのモデルを用いた。1つ目は、Transformer+Time2Vec である。2つ目は、Bi-directional LSTM (BiLSTM) である。3つ目は、LSTM である。4つ目は、GRU である。5つ目は、Bi-directional GRU (BiGRU) である。データの前処理として、全業種・業種別極性指標に対し三か月移動平均を取る。各モデルにおける train:valid:test の割合は、3:1:1 である。また、各モデルのエポック数は200とする。

5 結果

まず、セクション4の通り、全業種極性指標と業種別極性指標をそれぞれ作成する。

全業種極性指標について、時系列順に並べて三か月移動平均を取った結果は、図3のようになる。次に、Transformer+Time2Vec と比較モデルにおける実験の結果を示す。

比較実験の結果は表2の通りである。本研究では、MAE, MSE, RMSE, MSLE が五つの中で一番小さい Transformer+Time2Vec を選択した。

*3 正解を1, 不正解を0, としてラベル付けを行った。

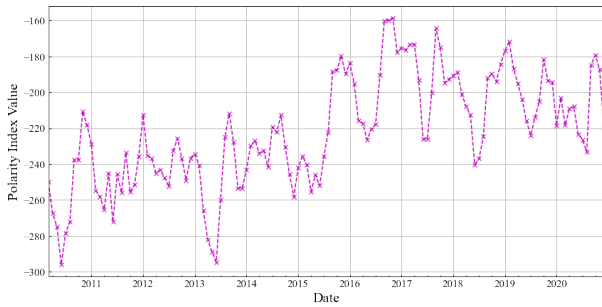


図3 3か月移動平均を取った全業種極性指標

表2 Model Comparison

Model	MAE	MSE	RMSE	MSLE
Transformer+Time2Vec	0.1186	0.0221	0.1485	0.0071
BiLSTM	0.2355	0.0684	0.2616	0.0299
LSTM	0.2243	0.0663	0.2574	0.0296
BiGRU	0.2302	0.0688	0.2622	0.0301
GRU	0.2278	0.0684	0.2616	0.0304

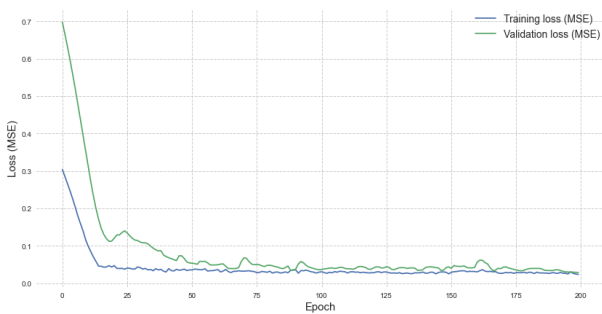


図4 Learning Curve of Transformer+Time2Vec

また、全業種極性指標 Transformer+Time2Vec の学習曲線は図4で示される。図4から過学習等もなく、100エポックあたりから Loss が大幅に変化していないことがわかる。業種別極性指標についても同じ傾向がみられた。

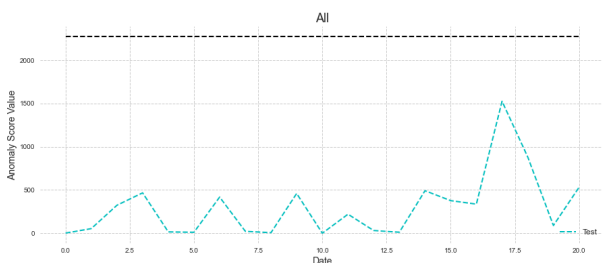


図5 Turning Point Detection for All-Industry Polarity Index

図5は全業種極性指標の異常スコアである。業種別の結果については付録Bを参照。

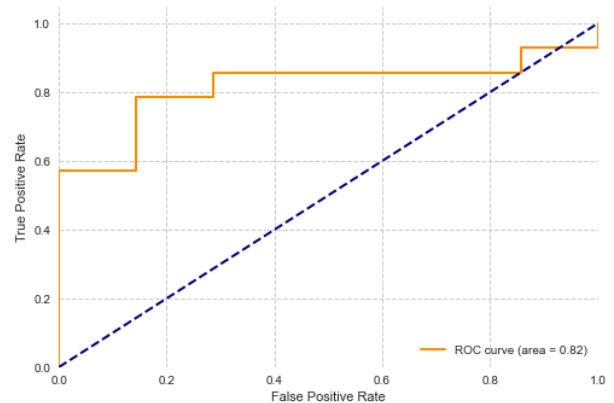


図6 ROC Curve for All-Industry Polarity Index

図6はセクション3の提案手法における ROC Curve である。AUC の値は 0.82 となっている。

6 考察

本研究では、BERT を用いて作成した極性指標を作成し、Transformer + Time2Vec 及びホテリングの T^2 理論で転換点検出を実施することで、市場分析のための新しい基盤技術を作成した。第一に、表2については、5つのうち Transformer + Time2Vec が一番高い結果となった。本研究では、input の特徴量を 1（株式と対応する業種別極性指標のみ）としたが、特徴量を 2 以上に増やした場合にも各評価指標の精度改善考えられる。今後は input の特徴量を増やした場合の検証もいたい。第二に、図6より、AUC の値が 0.82 であるため、本研究で提案した転換点検出手法が極性指標に対し、ある程度有効であることが分かった。今後は、本手法が市場分析に対してどの程度有効であるのか、実際の株価データを用いて検証を行いたい。第三に、閾値の設定について、検証データを用いる方法の他に、分位点やカイ二乗分布を用いる方法もあるので、比較検証なども行いたい。

7 まとめ

本研究では、BERT を用いて作成した極性指標を作成し、極性指標を Transformer + Time2Vec で学習し、その Output についてホテリングの T^2 理論で転換点検出を行った。結果、本研究で提案した手法が極性指標の転換点検出においてある程度有効であることが分かった。また、本手法をさらに拡張し、多変量解析を利用した市

場分析フレームワークの開発などを行いたい。

謝辞

本研究は、JST 未来社会創造事業 (PMJMI20B1, 代表：和泉潔) の助成を受けたものである。また、株式会社アイフィスジャパンより、本研究にご協力いただいた。厚く御礼を申し上げ、ここに感謝の意を表する。

参考文献

- [1] 坂地泰紀, 和泉潔, 松島裕康, 川瀬和哉 and 林寛, 2019. 接触履歴を用いた地域景況インデックスの自動生成. 知能と情報, 31(2), pp.626-635.
- [2] Sakaji, H., Kuramoto, R., Matsushima, H., Izumi, K., Shimada, T. and Sunakawa, K., 2019, August. Financial text data analytics framework for business confidence indices and inter-industry relations. In Proceedings of the First Workshop on Financial Technology and Natural Language Processing (pp. 40-46).
- [3] Yono, K., Sakaji, H., Matsushima, H., Shimada, T. and Izumi, K., 2020. Construction of macroeconomic uncertainty indices for financial market analysis using a supervised topic model. Journal of Risk and Financial Management, 13(4), p.79.
- [4] 山本裕樹 and 松尾豊, 2016. 景気ウォッチャー調査を学習データに用いた金融レポートの指数化. In 人工知能学会全国大会論文集 第 30 回 (2016) (pp. 3L3OS16a2-3L3OS16a2). 一般社団法人 人工知能学会.
- [5] 平松賢士, 三輪宏太郎, 酒井浩之 and 坂地泰紀, 2017. アナリストレポートのトーンの情報価値. 行動経済学会第 12 回大会予稿集.
- [6] Suzuki, M., Sakaji, H., Izumi, K., Matsushima, H. and Ishikawa, Y., 2020. Forecasting Net Income Estimate and Stock Price Using Text Mining from Economic Reports. Information, 11(6), p.292.
- [7] 小林和正, 酒井浩之, 坂地泰紀, 平松賢士, 2017. アナリストレポートからのアナリスト予想根拠情報の抽出と極性付与, 第 19 回金融情報学研究会, pp.65-70.
- [8] Hiew, J.Z.G., Huang, X., Mou, H., Li, D., Wu, Q. and Xu, Y., 2019. BERT-based financial sentiment index and LSTM-based stock return predictability. arXiv preprint arXiv:1906.09024.
- [9] Ding, Q., Wu, S., Sun, H., Guo, J. and Guo, J., 2020, January. Hierarchical Multi-Scale Gaussian Transformer for Stock Movement Prediction. In IJCAI (pp. 4640-4646).
- [10] Ahmed, M., Choudhury, N. and Uddin, S., 2017, July. Anomaly detection on big data in financial markets. In 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 998-1001). IEEE.
- [11] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- [12] Taguchi, R., Watanabe, H., Hirano, M., Suzuki, M., Sakaji, H., Izumi, K. and Hiramatsu, K., 2021, December. Market Trend Analysis Using Polarity Index Generated from Analyst Reports. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 3486-3494). IEEE.
- [13] Sakaji, H., Sakai, H. and Masuyama, S., 2008, May. Automatic extraction of basis expressions that indicate economic trends. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 977-984). Springer, Berlin, Heidelberg.
- [14] Kazemi, S.M., Goel, R., Eghbali, S., Ramanan, J., Sahota, J., Thakur, S., Wu, S., Smyth, C., Poupart, P. and Brubaker, M., 2019. Time2vec: Learning a vector representation of time. arXiv preprint arXiv:1907.05321.
- [15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.
- [16] Wu, N., Green, B., Ben, X. and O'Banion, S., 2020. Deep transformer models for time series forecasting: The influenza prevalence case. arXiv preprint arXiv:2001.08317.
- [17] Yamanishi, K. and Takeuchi, J.I., 2002, July. A unifying framework for detecting outliers and change points from non-stationary time series data. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 676-681).
- [18] Hotelling, H., 1947. Multivariate quality control.

Techniques of statistical analysis.

付録 A 時系列深層学習モデルの評価指標

本セクションでは、比較実験に用いた比較指標について定義を記す。

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (3)$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (4)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (5)$$

$$MSLE = \frac{1}{N} \sum_{i=1}^N (\log(y_i + 1) - \log(\hat{y}_i + 1))^2 \quad (6)$$

\hat{y}_i は i における予測値, y_i は i における実測値を表す。各評価指標は 0 に近いほど良いとされる。

付録 B 業種別極性指標における実証結果

本付録は業種別極性指標の異常スコアである。業種の選択は 33 業種の中からランダムに行われた。

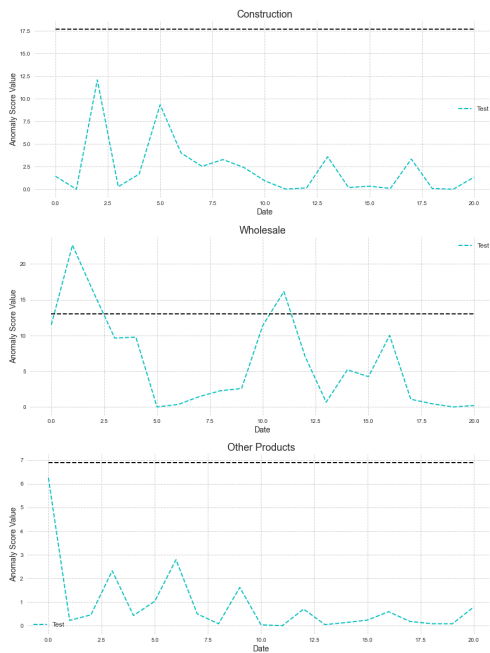


図 7 Turning Point Detection for Industry-Specific Polarity Index