ニュースおよび高頻度データを用いたディープラーニングによる株式変動の分析—BERT によるニュース評価—

Analysis of Stock Price Fluctuations by Deep Learning Using News and High-Frequency Data: BERT News Evaluation

西良浩 [1*]　菅愛子 [1]　高橋大志 [1]

Yoshihiro Nishi[1], Aiko Suge[1], and Hiroshi Takahashi[1]

[1] 慶應義塾大学大学院経営管理研究科

[1] Graduate School of Business Administration, Keio University

Abstract: ニュースは金融市場の資産価格に大きな影響を与える．ニュースと株価変動の関係性を分析し，ニュースを評価する取り組みはこれまでに多く行われており，ニュースと株価変動の間には関連性があると報告されている．しかしながら，ニュースは非構造化データであり，定型的に精度高く処理し，分析に用いることは難しい．本研究では，多くの自然言語読解タスクで SOTA(State of the Art)を達成している BERT を用いて株式変動を説明するニュース評価モデルを構築し，実証分析を行い，従来の分類モデルよりも高い正答率を得られるか検証を行った．本研究における分析の結果，BERT を用いたニュース評価モデルの精度が最も高かった．

1　はじめに

文書のような非構造化データを対象とし，資産価格の分析に用いる取り組みが模索されている．ニュースが株価に与える影響に関して，これまでに多くの取り組みが報告されており，金融市場の資産価格に大きな影響を与えることが示唆されている[1][3][4]．しかしながら，非構造化データは定型的に扱うことが困難なため，構造化データに比べると扱いが難しい．

近年は情報技術の進展に伴い，マシーンラーニングやディープラーニングを用い，非構造化データを自然言語処理により分析する手法が多く存在している．金融市場において発信されたニュースを用いた株式変動に関する分析をマシーンラーニングやディープラーニングを介して行うことで，より精度の高いモデルが構築できることが期待されている．

2019 年時点で時価総額が最も高い上位 3 社（トヨタ自動車株式会社，日産自動車株式会社，本田技研工業株式会社）を主要な自動車企業とし，分析対象とした．自動車産業は日本における重要な産業であり，経済全体に与える影響は大きい．自動車産業の出荷額は約 52 兆円で，主要製造業の出荷額の 20%を占める．輸出額は約 15 兆円で，日本の総輸出額の 20%を占めている．関連する製造人口は約 550 万人で，これは労働人口の 10%を占めている[6]．

これらを背景とし，本研究ではトヨタ自動車株式会社，日産自動車株式会社や本田技研工業株式会社を分析の対象とし，Bidirectional Encoder Representations from Transformers（BERT）[2]や，Long Short-Term Memory（LSTM）[5]といったディープラーニングを用いたニュース評価モデルの構築を行った．

2　関連研究

金融市場において配信されたニュースが株価の変動に与える影響に関して分析を行った取り組みは数多くある．例えば，ニュース記事のテキストマイニングにより株価変動を分析した研究では，ニュース記事に含まれるファンダメンタルおよびセンチメントに関する情報が株価に反映されている可能性が報告されている．ニューステキストをナイーブ・ベイズ分類器によって分類し，株価との関係について分析した取り組み[4]．ニューステキストを SVM により分析した取り組み[8]，生成したニュース記事を分

* 連絡先: 慶應義塾大学大学院 経営管理研究科 経営管理専攻
〒223 - 8526 神奈川県 横浜市 港北区 日吉 4-1-1
E-mail: nishi_yoshihiro@keio.jp

析用のデータとして追加し，LSTM により分析した取り組み[9]などがこれまでに報告されている．以上の取り組みにより，金融市場において配信されたニュースが株価変動にポジティブもしくはネガティブな影響を与えていると考えられる．

近年では，情報技術の発展に伴い，主として高精度化を目的とし，LSTM のようなディープラーニングを用いて金融市場の分析を行う取り組みが行われている[13].

3 データ

本研究では，分析の対象期間を 2014 年から 2016 年までとし，ニュースデータとマーケットデータを用いて分析を行った．2019 年時点で時価総額が最も高い上位 3 社（トヨタ自動車株式会社，日産自動車株式会社，本田技研工業株式会社）を主要な自動車企業とし，分析対象とした．

3.1 マーケットデータ

マーケットデータとして，トムソン・ロイター社より高頻度データを取得した．マーケットデータには，取引成立価格や取引量などの株式取引に関する情報が含まれており，各行にマイクロ秒単位のタイムスタンプが付されている．2014 年から 2016 年までのトヨタ自動車株式会社，日産自動車株式会社，本田技研工業株式会社に関するマーケットデータ 14 億 990 万 1961 件を取得した．

3.2 ニュースデータ

ニュースデータとして，トムソン・ロイター社が配信を行ったニュースを取得した．日本企業に関するニュースは主として英語もしくは日本語により配信されている．配信されたニュースのテキスト情報には，ヘッドラインと本文があり，ヘッドラインは本文内の重要な内容を要約したテキストデータである．ニュースには配信された日時のタイムスタンプが付されている．

本研究ではニュース配信の前後 1 分間に取引があった英語のヘッドラインを用いて分析を行う．2014 年から 2016 年までのトヨタ自動車株式会社，日産自動車株式会社，本田技研工業株式会社に関するニュース 2,259 件を取得した．表 1 は取得したニュースの内訳を各社ごとに示したものである．取得した 2014 年から 2016 年までにトムソン・ロイター社が配信を行ったトヨタ自動車株式会社に関するニュースは 1,065 件，日産自動車株式会社に関するニュースは 587 件，本田技研工業株式会社に関するニュースは 607 件であった．

表 1: 取得したニュースの件数

	件数
トヨタ自動車株式会社	1,065
日産自動車株式会社	587
本田技研工業株式会社	607
合計	2,259

4 分析手法

図 1 に構築した 3 つのニュース評価モデルに用いた分析の主な手順を示す．本研究において構築したニュース評価モデルは Labeling，Vectorization，Classification Layer の 3 つを通じて，ニュース評価を行っている．

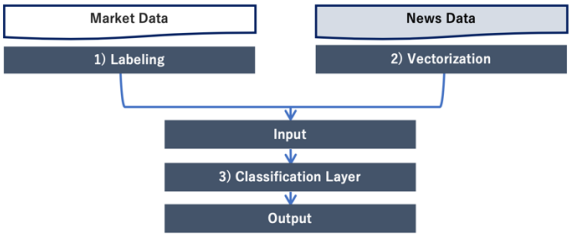

図 1: 分析の主な手順

4.1 株価変動率を元にしたラベル付け

Labeling では，ニュース分類分析を行うために，取得したニュースにラベル付けを行う[11]．ニュース配信前後のマーケットデータを取得し，（1）の定義式により，株価変動率を求め，2014 年から 2016 年までのトヨタ自動車株式会社，日産自動車株式会社，本田技研工業株式会社に関するニュース 2,259 件にラベル付けを行った．ラベルは Positive と Negative の二値とし，α ＞0% の場合は Positive，α ＜0% の場合は Negative とし，ラベル付けを行う．

$$株式変動率(\%) = \frac{(ニュース配信1分後の平均株価) - (ニュース配信1分前の平均株価)}{(ニュース配信1分前の平均株価)} \times 100$$

$$Positive: \alpha > 0\%$$
$$Negative: \alpha < 0\%$$

（1）

4.2 ニュースのベクトル化

Vectorization では, 取得したニュースのベクトル化を行う．ニュースのベクトル化には BERT，Keras の tokenizer，word2vec を用いる．

4.3 ニュース分類分析

Classification Layer では，取得したニュースと付与したラベルを元に，ディープラーニングを用いてニュース分類分析を行う．ニュース分類分析に用いたディープラーニングのモデルは BERT もしくは LSTM である．

5 比較評価方法

本研究では， 3 つのモデルを構築し，クロスバリデーションスコア（正解率）を用いて，比較評価を行った．本章の以降の節において，構築した各モデルについて述べる．

5.1 BERT モデル

BERT を用いてニュースデータをベクトル化し，ニュース分類分析を行ったモデルを BERT モデルとし，比較評価を行う．

BERT とは 2018 年に Google が発表した自然言語処理モデルのことである[2]．自然言語処理の読解のベンチマークタスクにおいて SotA を達成している．BERT はファイン・チューニングを用いたアプローチである．双方向の Transformer を用いており，Attention によるニューラル翻訳を行うモデルである．近年は Attention を用いた自然言語処理モデルが注目されており，多くのタスクにおいて高い性能を示している[12]．図 2 に BERT モデルのアーキテクチャを示す．

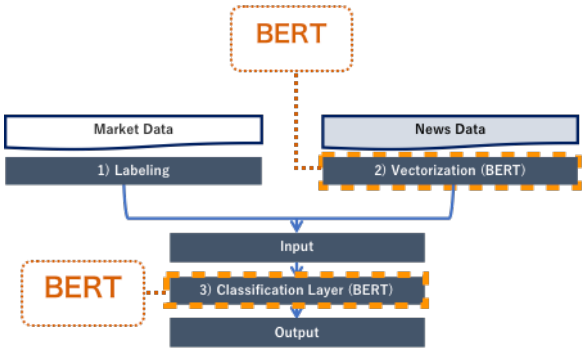

図 2: BERT モデルのアーキテクチャ

[*] https://github.com/keras-team/keras

5.2 Keras の tokenizer + LSTM モデル

Keras[*]の tokenizer を用いてニュースをベクトル化し，LSTM を介してニュース分類を行ったモデルを tokenizer + LSTM モデルとし，比較評価を行う．

Keras の tokenizer とは Keras ライブラリに含まれる文章のベクトル化のためのクラスである. Keras とは Python で記述されたニューラルネットワーク API であり，ディープラーニングライブラリを使用する際に用いられている．Keras の tokenizer に含まれる keras.preprocessing.text は単語のカウントや tf-idf などに基づき，各トークンの係数がバイナリになるベクトルに変換を行う．

LSTM (Long Short-Term Memory) は，時系列データを学習する RNN の一種である．LSTM は RNN を拡張しており，長期的な依存関係の学習を可能としている[5]．分類分析に LSTM モデルを使用し，精度検証にはクロスバリデーションスコア（正解率）を用いた． Compile を loss='binary_crossentropy', optimizer='rmsprop', metrics=['accuracy']とし，分析を行う．図 3 に tokenizer + LSTM モデルのアーキテクチャを示す．

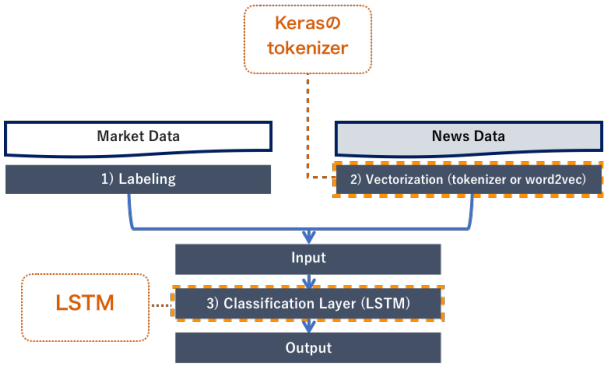

図 3: tokenizer + LSTM モデルのアーキテクチャ

5.3 word2vec + LSTM モデル

word2vec を用いてニュースをベクトル化し，LSTM を介してニュース分類を行ったモデルを word2vec + LSTM モデルとし，比較評価を行う．

ニュースのベクトル化には，最も広く用いられている word2vec の Skip-gram を用いた[7]．文書中の中心の単語から周辺の単語を予測するモデルを Skip-gram という．Skip-gram は，$W_1, W_2, \ldots, W_t$の順で単語が現れる場合に，(2) の定義式を用いて確率変数の

対数の項を最大化するベクトルを学習により探索する．

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(W_{t+j}|W_t) \qquad (2)$$

$p(W_{t+j}|W_t)$は，Hierarchical Softmax を用いて計算を行なっている．Hierarchical Softmax は，頻度の高い単語の順にハフマン木を作成し，階層ごとにロジスティック回帰を用いて全結合ソフトマックスに近似させる手法である[10]．

分類分析には，tokenizer＋LSTM モデルと同様に，LSTM を使用した．Compile を loss='binary_crossentropy', optimizer='rmsprop', metrics=['accuracy']とし，分析を行う．図 4 に word2vec + LSTM モデルのアーキテクチャを示す．

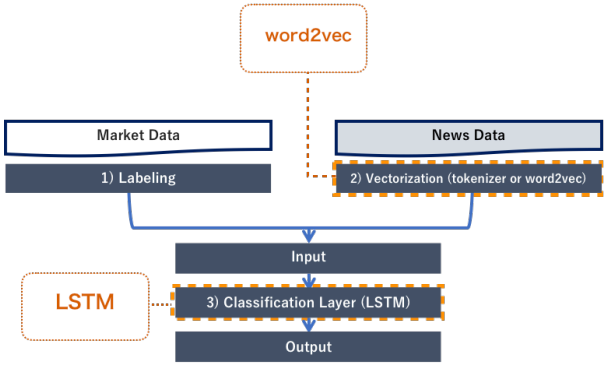

図 4: BERT モデルのアーキテクチャ

6 分析結果

6.1 ニュースのラベル付け

表 2 はラベル付けの結果を示したものである．2014 年から 2016 年までのトヨタ自動車株式会社，日産自動車株式会社，本田技研工業株式会社に関する 2,259 件のニュースは，Positive なニュースが 1,137 件，Negative なニュース 1,122 件となった．

ラベル付けの結果の例として，トヨタ自動車株式会社の Positive なニュースと Negative なニュースを表 2 に示している．米国における新製品の販売に関するニュースである”TOYOTA TO START SELLING NX COMPACT CROSSOVER SUV IN U.S. IN NOV, AIMS TO SELL 42,000 NX SUVS ANNUALLY IN U.S. -EXEC”が株価に Positive な影響を与えているのに対し，トヨタ自動車株式会社と本田技研工業株式会社のタイアップに関するニュースは事実ではないことを伝えたニュースである”TOYOTA MOTOR SAYS NO TRUTH TO REPORT ABOUT TIE-UP TALKS WITH SUZUKI MOTOR”は株価に Negative な影響を与えていることが分かる．

表 2: ラベル付けの結果

	ニュース数	例
Positive	1,137	TOYOTA TO START SELLING NX COMPACT CROSSOVER SUV IN U.S. IN NOV, AIMS TO SELL 42,000 NX SUVS ANNUALLY IN U.S. -EXEC
Negative	1,122	TOYOTA MOTOR SAYS NO TRUTH TO REPORT ABOUT TIE-UP TALKS WITH SUZUKI MOTOR

表 3 は，全てのモデルの教師データ数とテストデータ数を示したものである．ラベル付けを行ったニュースデータ 2,259 件を，教師データ 2,033 件，テストデータ 226 件に分割し，分析用のデータセットに格納した．

表 3: 分析用データセット

	データ数
教師データ	2,033
テストデータ	226
合計	2,259

6.2 ニュース分類分析の精度比較

評価尺度にクロスバリデーションスコア (正解率) を用い，BERT を介したニュース評価モデルと LSTM を介したニュース評価モデルの比較評価を行った．比較評価にあたり，各モデルの分析に用いるデータセットは全て同一のものを使用している．

表 4 は，分類分析の結果を示したものである．BERT モデルの正解率が，構築した 3 つのモデルの中で最も高かった．BERT モデルは，Tokenizer +

LSTM モデルよりも正解率が 10.6 ポイント高く，word2vec + LSTM モデルよりも正解率は 0.9 ポイント高かった．

表 4: ニュース分類分析の結果

	BERT モデル	Tokenizer + LSTM モデル	word2vec + LSTM モデル
正解率	0.624	0.518	0.615

7 まとめと今後の課題

本研究において，自然言語処理読解のベンチマークタスクにおいて SotA を達成した BERT を用いたニュース評価モデルの構築をし，比較評価を行った．評価実験の結果， BERT モデルのクロスバリデーションスコア（正解率）が最も高かった．

テキストマイニングによるラベルごとのニュースの分析や，分類分析の正誤傾向の分析は今後の課題としている．

参考文献

[1] David, M. C., James, M. P., Lawrence, H. S.: What Moves Stock Prices? The Journal of Portfolio Management Spring, 15（3）, pp.4-12, （1989）

[2] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. Bert.: Pretraining of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805, （2018）

[3] Fung G. P. C., Yu J. X., Lam W.: Stock Prediction: Integrating Text Mining Approach using Real-time News, In Proceedings of the IEEE International Conference on Computational Intelligence for Financial Engineering, pp. 395-402, (2003)

[4] Gidófalvi G.: Using News Articles to Predict Stock Price Movements, Department of Computer Science and Engineering, Technical Report University of California, (2001)

[5] Hochreiter S., Schmidhuber J.: Long Short-Term Memory, Neural Computation, Vol. 9, No. 8, pp. 1735-1780, (1997)

[6] Ibuki, H.: *Zidousya sangyou wo meguru kouzou henka to sono taiou ni tui te* [Structural Changes and Responses in the Automobile Industry]. Monthly Report Japan Foreign Trade Council, 745, pp. 7-11, （2016） （in Japanese）

[7] Mikolov T., Sutskever I., Chen K., Corrado G., and Dean J.: Distributed Representations of Words and Phrases and their Compositionality, In Proceedings of the NeurIPS, (2013)

[8] Mittermayer M. A.: Forecasting Intraday Stock Price Trends with Text Mining Techniques, In Proceedings of the 37th Hawaii International Conference on System Sciences, (2004)

[9] Nishi Y., Suge A., Takahashi H.: Text Analysis on the Stock Market thorough "Fake" News Generated by GPT-2, In Proceedings of the INFORMS Annual Meeting, (2019) (to appear)

[10] Rong, X.: word2vec parameter learning explained. arXiv preprint arXiv:1411.2738, （2014）

[11] Takayama, L., Suge, A., Takahashi, H.: LSTM Model for Explaining the Association between News Data and Stock Price Fluctuations. In Proceedings of the 11th JSAI Special Interest Group on Business Informatics, （2018）

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS), （2017）

[13] Wang, J.H., Liu, T.W., Luo, X., Wang, L.: An LSTM Approach to Short Text Sentiment Classification with Word Embeddings. In Proceedings of the 30th Conference on Computational Linguistics and Speech Processing, （2018）