Vol.2021-MPS-132 No.5
2021/3/1

BERTとLSTMを利用した株価予測

新津 康平[1,a)]　吉浦 紀晃[1]

概要：株式投資家は売買銘柄を選定する際，テクニカル分析とファンダメンタル分析を用いる．テクニカル分析は，過去の株価の平均価格，高値や安値などを用いて分析することで銘柄を選定する．ファンダメンタル分析では決算書からの情報やニュース等の非構造データを分析することで銘柄を選定している．ニュース等のテキスト情報から売買銘柄を選定することをニュース分析と呼ぶ．本研究では，ニュース分析とテクニカル分析を組み合わせた手法により株価予測を行う．株価データとニュースの埋め込み表現を時系列なデータに組み合わせ，それを LSTM に入力することで予測する．ニュースの埋め込み表現は BERT に入力し獲得する．本研究における分析では，実用的な株価の予測を行うことはできなかった．考えられる理由として個別銘柄の株価はニュースより株式市場全体から受ける景況の方が大きい，ニュースが影響を及ぼす期間が様々でより長い期間の時系列分析が必要であるなどが考えられる．

キーワード：LSTM，BERT，株価予測，自然言語処理

Stock prediction with BERT and LSTM

Kohei Niitsu[1,a)]　Noriaki Yoshiura[1]

Abstract: Stock investors use technical analysis and fundamental analysis when selecting stocks to buy or sell. Technical analysis selects stocks by analyzing the average price of past stock prices, highs and lows, and other data. Fundamental analysis selects stocks by analyzing information from financial statements and unstructured data such as news. The selection of trading stocks based on textual information such as news is called news analysis. In this study, we use a combination of news analysis and technical analysis to predict stock prices. We combine stock price data and news embedded expressions into time-series data, and input them into LSTM to make forecasts. The news embeddings are acquired by inputting them into BERT. In our analysis, we were not able to predict stock prices in a practical way. The possible reasons are that the stock prices of individual stocks are affected more by the business conditions of the stock market as a whole than by the news, and that the period of time over which the news affects the stock market varies and requires a longer time series analysis.

Keywords: LSTM,BERT,Stock Prediction,Natural Language Processing

1. はじめに

金融市場の動向を予測することは，投資家にとって最も重要な仕事の一つである．多くの投資家が，テクニカル分析やファンダメンタルズ分析などの手法を使って株式市場の動向を予測しようとしている．テクニカル分析は，過去の株価や売買高を利用し，株の将来の動きを予測する手法である．市場での取引の数値のみを考慮するため，機械学習や統計を用いた分析が容易である．

ファンダメンタルズ分析とは，企業の決算書に記載されている数値やニュースなどを用いて株価を予測する方法である．この方法では，金融ニュースや市場心理，経済的要因などから株価が決定される．投資家は企業の利益を推定し，投資に適しているかどうかを判断する．

株価を予測する方法は長年研究されており，様々な学問

[1] 埼玉大学大学院理工学研究科数理電子情報部門情報領域
Department of Information and Computer Sciences, SaitamaUniversity
[a)] k.niitsu.792@ms.saitama-u.ac.jp

分野でいくつかの手法が提案され実際の市場で応用されている [1]. 近年，株式市場分析の研究に機械学習の手法が用いられている [3]. 大規模なデータからパターンを学習することができる機械学習は，テクニカル分析に基づいた短期的な取引のためのトレンド予測に用いられることが多い. 機械学習の中でも非線形関数の分析に強い深層学習が広く用いられており，従来の機械学習と比較して良好な結果を出している [3].

中長期的な投資を行うためには，企業の財務諸表，経営状況，競争優位性などから市場における価値に着目するファンダメンタル分析も重要である. なぜなら，上場企業の株価は中長期的には市場原理に基づき市場における企業の価値の評価によって決定しているからである. しかし，トレンドを予測するようなテクニカル分析に加えて，株価を予測するファンダメンタル分析を組み合わせた手法は存在していない.

また，機械学習の自然言語処理分野での発展がめざましい. BERT は，Bidirectional Encoder Representations from Transformers の略称であり翻訳，文書分類，質問応答など自然言語処理タスクの 11 部門で最高記録を出している．BERT はより人間的な処理が可能ということで注目されている自然言語処理モデルである. そこで本研究では，既存のテクニカル分析の機械学習の手法も用いるだけでなく市場におけるニュースに着目したファンダメンタル分析を BERT を用いて行い，複合的な株価予測モデルを提案する.

2. 関連研究

2.1 テクニカル分析

株式市場分析に機械学習を用いることを提案している先行研究は数多くある. Kohzadi[1] は，商品価格予測のために非線形な関係を考慮できるニューラルネットワークモデルと線形の時系列解析に用いられる ARIMA モデルを用いて株価を予測し予測結果を比較した. ニューラルネットワークモデルは，ARIMA モデルよりも平均二乗誤差が 27 %，平均二乗誤差が 56 %低くなった. また，ニューラルネットワークモデルでは，絶対平均誤差と平均絶対パーセント誤差も低い結果となった.

また，Kara[2] は，イスタンブール証券取引所 (ISE) の National100 指数の価格を予測するためにニューラルネットワークとサポートベクタマシンを適用した. 10 種類のテクニカル指標を入力として使用し，多項式カーネルを用いたニューラルネットワークとサポートベクタマシン予測の最大値確率はそれぞれ 75.74%と 71.52%だった. インプットは，過去の指標価格と出来高データを利用したテクニカル要因のみを用いていた.

しかし，[1] と [2] の実験手順は，株価の時系列データを一切考慮せずにトレーニングデータとテストデータを使用したため，投資家にとって実用的でなかった. また，時系列データを分析する際には，トレーニングデータとテストデータの間に高い相関関係がある可能性があるため，トレーニングデータはテストデータよりも新しいデータであるべきでない. したがって，実物市場に投資するためには，テストデータセットをモデルからうまく隠すために，時系列データの予測日よりも前の日付を持つトレーニングセットを定義しなければならない.

近年，RNN(Recurrent Neural Network) を用いた時系列分類の報告が増えてきている. RNN は時刻ごとにデータを受け取り，出力層では時刻ごとの結果を出力する. 時刻ごとに入出力が発生するという特徴はあり，入力層と出力層がそれぞれデータ入力，結果出力という機能をもつ点においては，通常のニューラルネットワークと同様の仕組みであり，時系列のデータ分析に強いと結果が出ている.

また，RNN のうちより過去の時系列データの記憶に特化しているモデルが LSTM(Long Short Term Memory) と呼ばれている. LSTM は，シーケンスラベリング [3]，音声認識 [4]，異常検出 [5]，金融時系列予測 [6] などの逐次データタスクに広く利用されている. 多くのタイプの時系列問題では，予測を成功させるために単純な LSTM モデルまたはスタック LSTM モデルが使用されてきた. 株価は，時系列なデータの集合であり，テクニカル分析を行う際は LSTM を用いることにする.

2.2 自然言語処理について

様々な言語タスクに適用可能な埋め込み表現の学習は，非ニューラル [10] およびニューラル [11] の手法を含め，数十年に渡って活発な研究が行われてきた. 言語モデルの事前学習は，多くの自然言語処理タスクの改善に有効であることが示されている [12]. 単語だけでなく文章においても有効であり文の埋め込み表現や段落の埋め込み表現のような粒度でも置き換えられ言語タスクに用いられている. 事前に学習した埋め込み表現を様々な言語タスクに適用するためには，転移学習をする. ELMo[13] のような事前学習のアプローチは，事前に訓練された表現を追加の特徴として含むタスク固有のアーキテクチャを使用する. Generative Pre-trained Transformer (OpenAI GPT) [14] のような転移学習のアプローチは，最小限のタスク固有のパラメータを導入し，すべての事前学習されたパラメータを微調整するだけで，各タスクに特化させる. しかし，これらの事前学習では文章を左から右へ読み進める方法だった. Melamud ら (2016)[15] は，LSTM を用いて左右両方の文脈から単一の単語を予測するタスクを通じて文脈表現を学習することを提案した. また，ラベル付けされていないテキストから事前に訓練された単語埋め込みパラメータのみが用いられている [16]. さらに最近では，文脈トークン表現を生成する文または文書エンコーダーが，ラベル付け

されていないテキストから事前に訓練され，教師付き各タスクのために転移学習されている [16]．これらのアプローチの利点は，ゼロから学習する必要のあるパラメータが少ないことである．少なくとも部分的には，この利点のために，OpenAI GPT[14] は，GLUE ベンチマークの多くの文レベルタスクで以前に最先端の結果を達成した．これらをもとに，BERT[17] はマスクされた言語モデルを用いて，事前学習された深い双方向性表現を可能にしている．また，文レベルおよびトークンレベルの大規模なタスクの組において最先端の性能を達成し，多くのタスク固有のアーキテクチャーを凌駕する BERT は転移学習ありきのに基づく表現モデルである．事前学習を用いて転移学習を行うことで 11 の自然言語処理のタスクで最先端のベンチマークを記録している．

2.3 ファンダメンタル分析について

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．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 [7]，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．

3. データについて

本研究では，テクニカル分析とニュース分析を同時に行う手法を提案する．その際に使用する，株価データとニュースデータについて以下に記す．マシンスペックの限界からニュースの量は 6,000 件以内，かつ株価操作を受けにくい時価総額の高い企業を選択することにする．上記を踏まえて，分析対象とした企業は任天堂，キーエンスと日本電産である．また十分に会社規模が大きいことから株価操作を受けにくいと判断したため選択した．

3.1 株価データの取得

インベストメントドットコム [20] より企業の株価データを取得した．株価データは毎日の終値，始値，高値，安値と出来高が含まれている．2010 年 1 月 1 日から 2020 年 12 月 31 日までの市場が公開されている日にちの全ての株価データを取得した．任天堂，キーエンス，日本電産の各社株価データを 2,691 件取得した．

3.2 ニュースデータ取得

ニュースデータとして，日経新聞電子版の検索窓に各社名を入力し指定期間内に出力された全てのニュースを取得した．2010 年 1 月 1 日から 2020 年 12 月 31 日までの配信された全てのニュースを取得した．配信されたニュースデータには，タイトルと本文があり，タイトル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もある．任天堂，キーエンス，日本電産の各社のニュースは，それぞれ 5,716 件,1,419 件、5,131 件を取得した．

3.3 ニュースのタイムスタンプについて

日本の株式市場は，土日祝日，などと言った東京証券取引所が営業を行っていない日には株式の取引ができない．また，市場が公開されていたとしても午前 9 時から午後 15 時までが営業時間でありその他の時間には取引できない．そのため，ニュースが配信された日をそのままニュースが配信された日にすることはできない．例えば，土曜日に配信されたニュースは実際その日の最も近い未来の市場公開日から影響を及ぼし始めるためニュース配信日をその日の最も近い過去の市場公開日とする必要がある．また，市場公開があった日だったとしても 15 時以降に配信されたニュースはそれと同様である．そのため，土日，祝日，休場日と市場公開日の 8 時 59 分までに配信されたニュースはタイムスタンプに最も近い過去の日にちに公開されたとする．また，市場公開日かつ 9 時 00 分から 14 時 59 分までに配信されたニュースはそのニュースに記載があるタイムスタンプの日にちに配信されたとする．

4. 提案手法

ニュースを分析し株価予測する手法を提案する．[18] のニュースのみの株価予測では，ニュース配信前の時系列株価は考慮されていない．そのため，ニュースと株価の時系列なデータを作成しニュース配信前の株価情報を考慮した株価予測を行う．また，[17] で示されている通り，ニュースを分析する際は BERT の事前学習済みのモデルを用いて文の埋め込み表現を獲得後，ニュース分析のための転移学習を行う．

4.1 提案手法概要と目的

BERT を用いてニュース分析を行い，その結果を LSTM に入力し株価予測する．モデルの概要を図 1 に示す．

4.2 データセットの構築

データセットを構築する．市場が公開していた日にちを基準に時系列に株価とニュース文章を用いて構築する．時系列なデータセットのうち，前半 8 割をトレーニングデータとして用い後半 2 割をテストデータとして用いる．次に，データセットを 20 日ごとの系列に分解する．この際，20 日毎のデータセットのはじめの日から 21 日後の株価をその系列のラベルとして扱う．

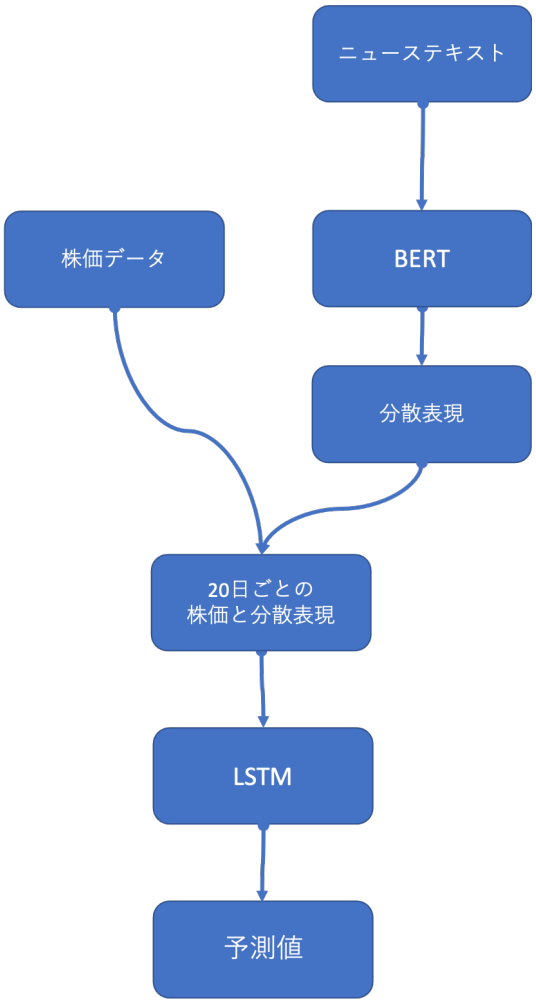

図 1　BERT と LSTM を利用したモデル
Fig. 1　Model with BERT and LSTM

市場 公開日	終値	始値	高値	安値	出来高	ニュース 1	…	ニュースN
2010/1/4								
…	…	…	…	…	…	…	…	…
2020/12/29								

図 2　BERT と LSTM を利用したモデルのデータセットイメージ
Fig. 2　Dataset image of the model with BERT and LSTM

市場 公開日	終値	始値	高値	安値	出来高	ニュース 1	…	ニュースN
date1								
…	…	…	…	…	…	…	…	…
date20								

図 3　BERT と LSTM を利用したモデルのデータセットを系列分割したイメージ
Fig. 3　Image of the data set of a model using BERT and LSTM with series partitioning.

データセットの概要を図 2 にデータセットを系列に分解したデータ概要を図 3 に示す．

4.3　提案モデル

提案モデルの概要は図 1 に示した通りである．まず，ニューステキストを BERT の事前学習済みモデルに入力することで文の埋め込み表現を獲得する．その埋め込み表現と株価を系列データに従って LSTM に入力し転移学習を行う．学習により獲得したモデルにテストデータを入力することで株価を予測する．

5.　実験と結果

提案手法に基づいて実験を行う．提案手法が有効な分析方法かどうかを検証する．株価データと LSTM のみを用いた分析結果と株価データ，ニュースデータにより構築されたデータと BERT，LSTM を用いて構築したモデルを用いて分析した結果を比較する．比較の際には，予測した株価と実際の株価の平均二乗誤差を用いる．

5.1　実験手順

5.1.1　埋め込み表現の獲得

ニュース本文を形態素解析を行い単語を取得した．その際，形態素解析ツールとして MeCab，辞書は mecab-ipadic-2.7.0-20070801 を使用した．その後，事前学習済みの BERT により文章の埋め込み表現を獲得する．事前学習済みの BERT は東北大学 [19] が作成している事前学習モデルを使用した．事前学習済みの BERT を適用する際，512 文字以上の文章は前半 512 文字のみの単語ををニューステキストとして処理した．

5.1.2　転移学習

獲得した埋め込み表現と株価のデータを LSTM に入力し学習を行う．株価は LSTM に入力する際，標準化を行い入力した．また，LSTM のパラメータとして，隠れ層を 512 次元，LSTM の層を 1 層，ドロップアウトを 0.5，バッチサイズを 16，エポックを 30，学習率を 0.001，損失関数に平均二乗誤差，最適化関数に Adam を利用した．

5.2　結果

学習し終えたモデルにテストデータを入力し株価の予測

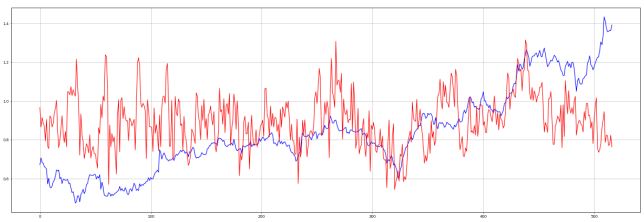

図 4　BERT と LSTM を利用し予測した任天堂の株価
Fig. 4　Nintendo's stock price predicted with BERT and LSTM

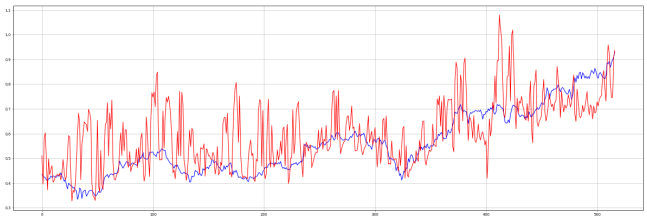

図 5　BERT と LSTM を利用し予測したキーエンスの株価
Fig. 5　Keyence's stock price predicted with BERT and LSTM

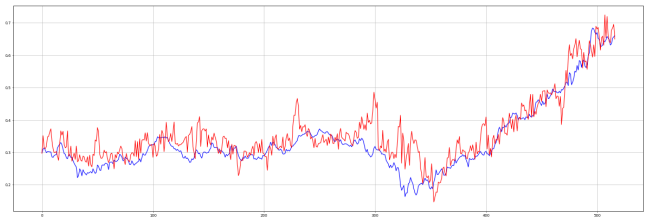

図 6　BERT と LSTM を利用し予測した日本電産の株価
Fig. 6　Nihondensan's stock price predicted with BERT and LSTM

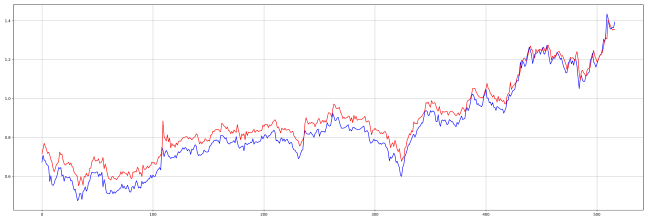

図 7　LSTM のみを利用し予測した任天堂の株価
Fig. 7　Nintendo's stock price predicted with only LSTM

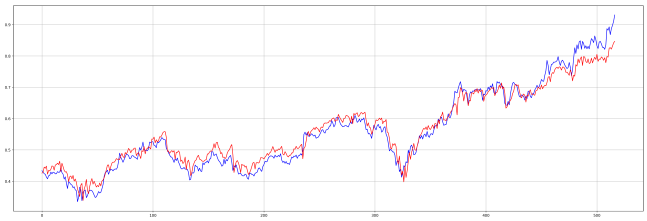

図 8　LSTM のみを利用し予測したキーエンスの株価
Fig. 8　Keyence's stock price predicted with only LSTM

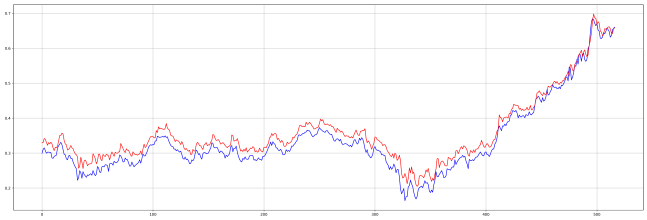

図 9　LSTM のみを利用し予測した日本電産の株価
Fig. 9　Nihondensan's stock price predicted with only LSTM

を行う．株価の予測をした結果が以下の図 4 から図 6 である．グラフの赤線が予測結果，青線が実際の値である．

また，LSTM のみで行った予測結果を図 7 から図 9 に示す．LSTM で学習をする際のパラメータは隠れ層を 512 次元，LSTM の層を 1 層，ドロップアウトを 0.5，バッチサイズを 16，エポックを 30，学習率を 0.001，損失関数に平均二乗誤差，最適化関数に Adam を利用した．株価はそれぞれ標準化された値である．グラフの赤線が予測結果，青線が実際の値である．

表 1　各種モデルの予測株価と実際の株価の MSE の比較
Table 1　Comparison of Mean Squared Error between predicted and actual stock prices for various models

社名	LSTM のみ	BERT と LSTM	MSE 増加率
任天堂	0.00335458	0.05498969	6.1%
キーエンス	0.00076845	0.01461851	5.3%
日本電産	0.00076371	0.00255678	29.9%

それぞれのモデルの精度を比較するために，実際の株価と予測した株価の平均二乗誤差を取得する．精度が高いモデルは実際の株価と差が少なくなるため平均二乗誤差は小さくなる．LSTM のみで予測した株価と実際の株価の平均二乗誤差と，BERT と LSTM を利用し予測した株価と実際の株価の平均二乗誤差を以下の表 1 に示す．表 1 では平均二乗誤差を MSE と表記する．任天堂は 6.1%，キーエンスは 5.3%，日本電産は 29.9%の平均二乗誤差が大きくなってしまった．

6.　考察

結果に示したとおり，LSTM のみで分析した結果の平均二乗誤差より BERT と LSTM を利用した平均二乗誤差の方が大きくなった．そのため，ニュース分析を BERT を用いて行うことはできなかった．考えられる理由として，個別銘柄の株価はニュースより株式市場全体から受ける影響の方が大きい，ニュースが影響を及ぼす期間が様々でより長い期間の時系列分析が必要である，ニュースが配信される前に株価は影響を受けているなどが考えられる．

近年，個別銘柄を複数まとめ一つのパッケージとして市場に公開している ETF(Exchange Traded Fund) と言う金融商品が普及している．ETF を購入すると複数の個別銘柄に分散投資される仕組みになっている．また，ETF は株式指数と相関が高い商品となっている．そのため，個別銘柄の株価はニュースより株式市場全体から受ける影響の方が大きいと推測される．

また，系列を 20 日と限定したことが原因で分析が正しく行われていない可能性がある．今回のデータの作成方法では，ニュースが 20 日以上経過したのちに株価へ影響を及ぼすことを考慮していない．また，ニュース分析なのどのファンダメンタル分析では，中長期の投資で用いいられる分析手法である．これらの理由により，20 日以上の系列を用いて分析する必要があると考えられる．

7.　まとめ

本研究では，BERT と LSTM を用いてニュース分析とテ

クニカル分析を同時に行う手法を提案した．ニュースデータと株価データを取得しそれぞれを時系列なデータとして組み合わせた．ニュースデータは BERT により埋め込み表現へ変換し，それを株価データともに 20 日ごとの系列にまとめた．それらを LSTM に入力し学習を行いモデルを作成した．しかし，株価予測は正確には行われなかった．今後の研究では，より長い系列データを作成し検証する必要がある．

参考文献

[1] Nowrouz Kohzadi, Milton S Boyd, Bahman Kermanshahi, and Iebeling Kaastra.: *A comparison of artificial neural network and time series models for forecasting commodity prices*, Neurocomputing, (1996).

[2] Yakup Kara, Melek Acar Boyacioglu, and Omer Kaan Baykan.: *Predicting direction of stock price index movement using artificial neural networks and support vector machines*, The sample of the istanbul stock exchange. Expert systems with Applications,(2011).

[3] Kazuya Kawakami. : *Supervised sequence labelling with recurrent neural networks.* , PhD thesis, Ph. D. thesis,Technical University of Munich,(2008).

[4] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. : *Speech recognition with deep recurrent neural networks.*, In 2013 IEEE international conference on acoustics, speech and signal processing, pages 6645-6649. IEEE,(2013).

[5] Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal.: *Long short term memory networks for anomaly detection in time series.* , In Proceedings, page 89. Presses universitaires de Louvain, (2015).

[6] Wei Bao, Jun Yue, and Yulei Rao.: *A deep learning framework for financial time series using stacked autoencoders and long-short term memory.* , PloS one, 12(7):e0180944, (2017).

[7] Gyozo Gidofalvi：*Using News Articles to Predict Stock Price Movements, Department of Computer Science and Engineering*, Technical Report University of California,(2001)

[8] Wei Bao, Jun Yue, and Yulei Rao.: *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)

[10] Peter F Brown, Peter V Desouza, Robert L Mercer: *Class-based n-gram models of natural language.*, Vincent J Della Pietra, and Jenifer C Lai. (1992).

[11] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge,Thorsten Brants, : *One billion word benchmark for measuring progress in statistical language modeling.*, Phillipp Koehn, and Tony Robinson. (2013).

[12] Andrew M Dai and Quoc V Le.: *Semi-supervised sequence learning. In Advances in neural information processing systems*, (2015)

[13] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. : *Deep contextualized word representations*, In NAACL.(2018)

[14] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever.: *Improving language understanding with unsupervised learning*, Technical report, OpenAI.(2018).

[15] Oren Melamud, Jacob Goldberger, and Ido Dagan.: *context2vec: Learning generic context embedding with bidirectional LSTM*, In CoNLL. (2016)

[16] Ronan Collobert and Jason Weston. *A unified architecture for natural language processing: Deep neural networks with multitask learning*, In Proceedings of the 25th international conference on Machine learning, pages 160-167. ACM.(2018).

[17] Jacob Devlin,Ming-Wei Chang, Kenton Lee, Kristina Toutanova,: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv preprint arXiv:1810.04805, （2018）

[18] Menggang Li, Wenrui Li, Fang Wang, Xiaojun Jia, Guangwei Rui: *Applying BERT to analyze investor sentiment in stock market*, Neural Computing and Applications, (2020).

[19] Kentaro Inui: *Pretrained Japanese BERT models*, https://github.com/cl-tohoku/bert-japanese,(2020-1-20)

[20] インベストメントドットコム社: インベストメントドットコム, https://jp.investing.com/,(2020-1-20)