DEIM Forum 2022 H23-2

複数ニュース記事の株価に及ぼす影響の分析手法

平野　瑠登† 　　馬　　強††

† 京都大学工学部情報学科　〒 606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †hirano@db.soc.i.kyoto-u.ac.jp, ††qiang@i.kyoto-u.ac.jp

あらまし　ニュースを用いて株価を予測する研究が盛んである．しかしながら，既存手法の多くは複数のニュースによる影響を統合的に考慮していない．そのため，市場がクローズドしている際に発行される複数ニュースの株価に及ぼす影響を分析することが困難である．そこで，本研究では，複数ニュースの市場に及ぼす影響について分析する手法を提案する．提案手法では，まず，市場がオープンしているときのニュースと株価データを用いてニュースと株価のエンコーダーを学習する．得られたエンコーダーを用いて，市場がクローズドしているときの複数ニュースによる株価への影響を予測するネットワークを構築して学習する．実験を行って提案手法の有効性と課題を確認した．

1　はじめに

高齢化は，世界的に最も重要な社会問題の一つである．高齢化は，先進諸国で進行しているが，その中でも日本は特に進行が早く，深刻な問題となっている．高齢化が進むことで，考えなければならない問題は様々であるが，その一つである，「老後 2000 万円問題」と呼ばれる問題が話題になっている．老後 2000 万円問題とは，金融庁の金融審議会「市場ワーキング・グループ」による，「老後 20〜30 年間で約 2,000 万円が不足する」という試算を発端に物議を醸した，いかに老後の資金を形成するか」をめぐる問題のことである．2000 万円という金額は，夫６５歳以上，妻６０歳以上の夫婦のみの無職世帯では月に 5.5 万円不足するという計算で，30 年間で合計 1980 万円不足するという計算に基づいている．近年は，退職金も減少傾向にあり，老後の資金不足はさらに深刻になってゆくと思われる．また，近年は，転職回数が増加傾向にあることや，フリーランスなど，働き方の多様化により，退職金が受け取れないケースも多く，退職金や年金で豊かな老後を過ごすというのは難しくなってきている．このような，確実に進展しつつある高齢化社会において，「貯蓄から投資へ」という考え方は豊かな老後を過ごすために重要なものになってきている．

投資家たちはインターネットなどを通して様々な情報を収集して投資を行っている．しかし，人間の目で全ての情報を処理することは困難である．そこで，高速取引やロボアドバイザなど投資に関する技術やシステムが研究開発されている．アルゴリズム取引ではテクニカル分析がよく利用されるが，市場の分析を，株価の数値データのみを用いて行うことは難しいことが知られている [4]．Huang ら [1] はサポートベクターマシンを，Xu ら [2] はニューラルネットワークの標準的な教師あり学習を用いて，株価の動きの予測モデルを構築した．しかし，株価は非常に多様な要因に依存するので，これらの手法は一般化が難しい [4]．例えば，海外市場の影響で日本の市場が変化する例や，アノマリーと呼ばれる，理論的に説明できない株価の変動なども市場には存在する．そのため，株価の数値データ以外の様々な情報も用いて，投資家の意思決定を支援する研究が行われている [16]．

多くの投資家が投資の判断に用いる，株価に影響を与える数値データ以外の情報は，SNS の発信，企業発表やニュースなどがある．

我々は，このように様々存在する投資を判断する情報の中で，ニュース記事に着目し，発行されるニュース記事の影響の大きさを分析することで，どのニュースが重要なニュースか判断することができ，投資者の意思決定の支援に貢献できると考えた．ニュースを使用した既存研究は多数存在するが，既存のニュースを用いた手法は，ニュースの株価に対する影響の時間や範囲に対する考慮が不十分である [14], [15]．ニュースの重要性を測る要素として，以下の３つが挙げられると考えられる．

(1)　市場に影響を与える時間の長さ

ニュース記事の内容によって，市場に影響を与える時間は異なると考えられる．投資の意思決定に大きく関わる重要なニュースほど，市場に影響を与える長さは大きくなり，重要度の低いニュース記事は，市場に影響を与える時間の長さは小さいと考えられる．

(2)　ニュースが影響を与える企業の数

ニュース記事が言及する内容により，そのニュース記事によって影響を受ける企業の数は異なると考えられる．例えば，ある企業に関するニュース記事が発行されたとき，その企業の子会社や，提携を結んでいる企業などにも影響が及ぶ可能性がある．重要なニュース記事であれば影響を与える企業の数が大きくなり，重要度の低いニュース記事は，影響を与える企業の数が小さいと考えられる．

(3)　ニュースによる出来高変動の幅

ニュース記事の内容によって，出来高が変動する幅は異なると考えられる．多くの投資家の意思決定に影響を与える重要なニュース記事に関しては，出来高は大きく増加するか，大きく減少するというふうに，変動の幅が大きいと考えられる．重要度の低いニュース記事では，出来高はほ

とんど変化しないと考えられる．

松橋 [14] は，１つ目の「市場に影響を与える時間の長さ」に着目した．取引データとニュースを入力して，ニュースの影響時間を推定した．ニュースをクラスタリングし，各クラスタのニュースの影響分布を混合正規分布と仮定し，それぞれのパラメータを推定した．馬場ら [15] は，２つ目の「ニュースが影響を与える企業の数」に着目した．ニュースを用いて，そのニュースが影響を及ぼすと推測される企業のランキングを出力する手法を提案した．

そこで本研究では，出来高の変化も価格の予測や投資の意思決定に重要であると考え，３つ目の「ニュースによる出来高変動の幅」に着目した分析を行う．ニュースと，そのニュースで言及されている企業の出来高と株価に関する時系列データを入力し，ニュース発行後に出来高が増加するか，減少するか，変化がないかを予測する．株価を予測する研究は，投資家の利益に直接的につながるため盛んに行われているが，ニュースの影響を分析するために出来高の変化に着目しているものは少ない．しかし，出来高の変化を予測し，どのようなニュースが出来高の変化を引き起こすかを分析することで，投資に対する意思決定において，重要な情報とそうでない情報を見分けることができるようになり，投資の意思決定の支援に貢献できると考える．

本研究では，ニュース記事が，市場が開いている時に発行されたか，閉まっている時に発行されたかによって，２種類のモデルを提案する．これは，市場が開いている時に発行されたニュースは，発行された一つ一つのニュースが，即時出来高に影響を与えるのに対し，市場が閉まっている時に発行されるニュースは，閉まっている間に発行された全てのニュースが，次に市場が開いた時の出来高に影響を与えるという違いがあるため，扱い方を分ける必要があると考えたためである．既存手法では，このように複数のニュース記事を用いたものが少ない．

市場が開いている時に発行されたニュースを用いるモデルは，シングルモデルと呼ぶ．シングルモデルでは，ニュース記事を文に分割し，各文を Sentence BERT に入力してベクトルを生成する．入力は，ある企業のニュース記事のテキストデータとその企業の出来高と株価の時系列データである．出力は，その企業の株の出来高が，ニュース発行から t_{before} 分以内に増えるか減るか変わらないかの予測値である．

市場が閉まっている時に発行されたニュースを用いるモデルは，マルチモデルと呼ぶ．これは，入力するニュースが，次に市場が開くまでに発行された複数のニュース記事であることに由来する．マルチモデルでは，シングルモデルで学習したエンコーダに，各ニュース記事を入力する．入力は，ある銘柄の市場が閉まっている間に発行された複数のニュース記事のテキストデータである．出力は，その企業の株の出来高が，次に市場が開いてから t_{after} 分以内に増えるか減るか変わらないかの予測値である．

本論文の構成は以下の通りである．第２節では関連研究を示し，第３節では本研究が提案するシングルモデルとマルチモデルについて記し．第４節では評価実験の方法と結果を記す．そして，第５節は本研究のまとめである．

2 関 連 研 究

株価の数値データ以外の情報を用いて市場を分析する研究は盛んに行われている．Tetlock [11] は Wall Street Journal のコラムから悲観度を測定し，ダウ工業平均株価と関連していることを示した．また，Bollen ら [12] は，Twitter のツイートから得られる世間の気分状態の測定値と，ダウ工業平均株価との相関性を見出した．これらにより，本研究で扱うニュース記事においても，同様に株価への影響がある可能性が高いと考えられる．

Li ら [13] は，市場が閉まった後のニュースのみを考慮し，前日の終値と翌日の終値の間の動きを予測する手法を示した．LSTM-RGCN という，RGCN の各層の間に LSTM を挟んだモデルを用いて，企業間の関連も考慮に入れている．これにより，ニュース記事が出ていない企業の株価についても，良い精度で予測することに成功した．

これらは海外の株式市場を対象としている．日本国内における，ニュース記事を用いた株式市場の分析の研究として，許ら [3] は，潜在的なトピックを単語埋め込み空間上で表現するトピック埋め込みモデルである TopicVec に回帰機能を付与したトピック埋め込み回帰モデル TopicVec-Reg を提案し，これを用いて金融記事から株価の予測を行った．沖本ら [5] は，日経 QUICK ニュースを対象に，ニュース指標を作成し，その指標に基づく株価市場の予測可能性について研究したが，対象の株価に対して複数のニュースの考慮は行っていない．

松橋ら [14] は，取引データとニュースを入力して，ニュースの影響時間を推定した．ニュースを k-means 法でクラスタリングし，各クラスタのニュースの影響分布を混合正規分布と仮定し，それぞれのパラメータを推定するという手法である．この論文の実験では，クラスタリングを考慮する場合としない場合を比較して実験を行なっているが，実験結果は，ニュース記事のクラスタリングを考慮しない方が誤差が小さくなっていた．これは各記事の分布が各々固有のパラメータを持っているため，より取引数にフィットした形で分布を導出するためであると述べられている．また，定性分析として，マクドナルドとソニーの２社を対象に，ニュースのうち影響時間の長いトピックについて分析した結果，マクドナルドとソニーの両方とも，企業情報に関するニュースが，取引に対して影響を与える時間が長いということがわかった．この研究において，影響の分布に関して，混合正規分布を仮定しているが，影響のピークがニュース発行の後になる場合などを考慮すると，ポアソン分布などの別の分布で分析を行うことによって良い結果が得られる可能性があるという課題点がある．

また，馬場ら [15] は，企業を入力すると，その企業と関連のある企業を，関連が強い順のランキングで出力するというシステムにおいて，株価データだけでなく，ニュース記事を用いることで精度が高まることを示した．株価の変動が，入力された企業の株価変動と似ている企業が関連性が高いとしている．また，株価の変動は，入力された企業に関する直近のニュースが

発行されてからの，実価の前日比の時系列データを用いる．実価とは，株価から業種の影響と市場の影響を差し引いた値のことであり，業種指数と市場指数を仮定し，

$$\text{実価} = \frac{\text{株価}}{\text{業種指数} * \text{市場指数}}$$

という式で算出される．また，出力されたランキングの評価尺度としては，出力された企業に対して，問合せ企業との関連度を，人間が１点から５点で点数をつけ，それを用いて計算される nDCG(Normalized Discounted Cumulative Gain) を用いた．

このように，ニュース記事は株式に大きく影響を及ぼし，分析にとって非常に重要な要素であることがわかった．そこで本研究では，ニュース記事の発行時刻に着目し，市場が開いている時と閉まっている時に分類し，シングルモデルとマルチモデルを構築し，出来高の変化を分析することで，ニュースの重要性が推定できる手法を提案する．本研究では，既存研究では少なかった出来高の分析を行い，また複数のニュースを分析に用いている点が新しい．

3 提案手法

3.1 概　　要

ニュースの株取引に及ぼす影響を調べるため，本研究では，ある銘柄の出来高と株価の時系列データとニュース記事を用いて，その銘柄の出来高の増減を分析する．また，対象とする記事が取引時間内か時間外かによって，シングルモデルとマルチモデルの２つのモデルに分けて分析を行う．

シングルモデルは，市場が開いている時のニュースを用いる際のモデルである．入力は，ある企業のニュース記事のテキストデータとその企業の出来高と株価の時系列データである．出力は，その企業の株の出来高が，ニュース発行から t_{after} 分以内に増えるか減るか変わらないかの予測値である．処理の流れを図 1 に記す．ニュース記事と時系列データをそれぞれニュースエンコーダーと時系列エンコーダーでベクトルに変換したものを Transformer のエンコーダー層と MLP 層からなる識別器に入力して，出来高の変化の予測値を出力する．

マルチモデルは，市場が閉まっている時のニュースを用いる際のモデルである．入力は，ある企業の市場が閉まっている間に発行された複数のニュース記事のテキストデータである．出力は，その企業の株の出来高が，次に市場が開いてから t_{after} 分以内に増えるか減るか変わらないかの予測値である．処理の流れを図 2 に記す．複数のニュース記事と時系列データをそれぞれニュースエンコーダーと時系列エンコーダーでベクトルに変換したものを Transformer のエンコーダー層と MLP 層からなる識別器にに入力して，出来高の変化の予測値を出力する．出力は，増加するか，変わらないか，減少するかの３クラス分類である．

3.1.1 ニュースエンコーダー

ニュースエンコーダーは Sentence BERT [6] を用いて構築する．入力するニュース記事の文をそれぞれ Sentence BERT を用いてベクトルを生成してから，Transformer のエンコーダー層と MLP 層を経てニュース記事のベクトルを生成する．本研究では，Sentence BERT の事前学習済みモデルを，ニュース記事でファインチューニングしてから，ニュース記事の埋め込みベクトルを生成する..

3.1.2 時系列エンコーダー

時系列エンコーダーは TS2Vec [10] を用いて構築する．出来高と株価のそれぞれに対して TS2Vec でベクトルを生成して，Transformer のエンコーダー層と MLP 層を経て，時系列ベクトルを生成する．TS2Vec とは，時系列データを，分類タスク，予測タスク，異常検知タスクに適したベクトル表現に変換するための手法である [2]．モデルの学習には，対照学習と呼ばれる自己教師あり学習が用いられる．

本研究では TS2Vec を，出来高と株価に関する時系列データをベクトルに変換するのに用いる．

3.1.3 教師データ

学習に必要な教師データについて説明する．まず，ラベル付けのルールとして，ニュースが発行される前と後で出来高が変化しているかに基づいてラベル付けを行う．ニュースが発行された企業の，発行時刻の t_{before} 分前から発行時刻までの出来高の合計に対して，ニュース発行時刻から t_{after} 分後の出来高の合計が，増加しているか，減少しているか，変化していないかをラベル付けする．また，変化が ϵ%以内であった場合，変化していないとみなす．

3.2 提案モデルの詳細

3.2.1 シングルモデル: 市場が開いている場合

この節では，市場が開いているときに発行されたニュースを用いた分析手法について説明する．入力は一つのニュース記事，そのニュースで扱われている企業の出来高と株価の時系列データである．まず，ニュース記事の処理の流れは以下の通りである．

a） ニュースエンコーダー

（1） 対象のニュース記事を a_0 とする．

記事 a_0 を，句読点で区切り，記事を文 s_i の集合 $S = \{s_0, \ldots, s_n\}$ とする．

（2） 各 s_i を Sentence BERT に入力し，ベクトルの集合 V_s を生成する．

$$V_S = \{\mathbf{v}_{s_0}, \ldots, \mathbf{v}_{s_n}\}$$

$$\mathbf{v}_{s_i} = SentenceBERT(s_i)$$

（3） 生成されたベクトルの集合 V_S の各要素 $\mathbf{v}_{s_i}$ を Transformer のエンコーダー層に入力し，別のベクトルの集合 $V'_S = \{\mathbf{v}'_{s_0}, \ldots, \mathbf{v}'_{s_n}\}$ に変換する．

（4） 変換してできたベクトルの集合 V'_S の各要素を連結し，一つのベクトル $\mathbf{v}_{a_0} = (\mathbf{v}'_{s_0}, \ldots, \mathbf{v}'_{s_n})$ とする．

（5） v_{a_0} を多層パーセプトロンに入力して，l_{news} をベクトルのサイズとし，

$$\mathbf{v}_{news} = (v_0, \ldots, v_{l_{news}-1})$$

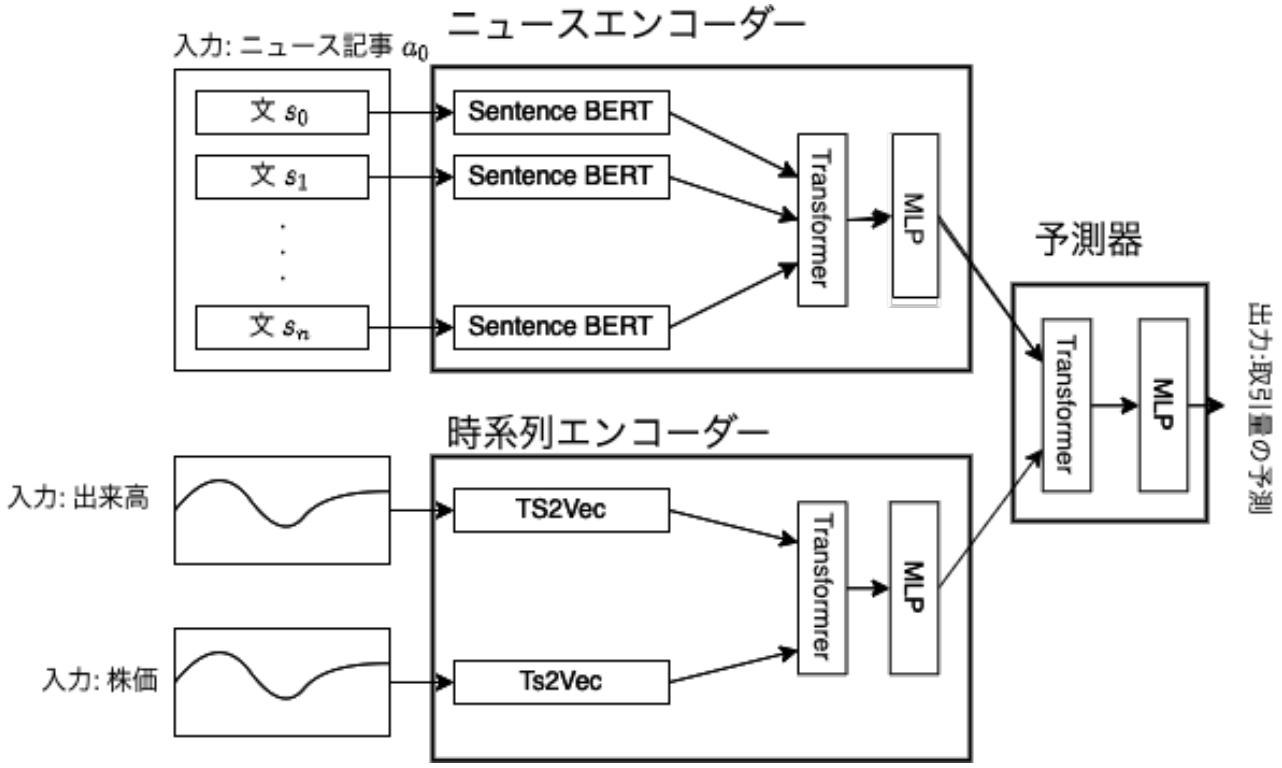

図 1　シングルモデル．a_0 はニュース記事，$s_0, \ldots, s_n$ は記事中の各文を表す．

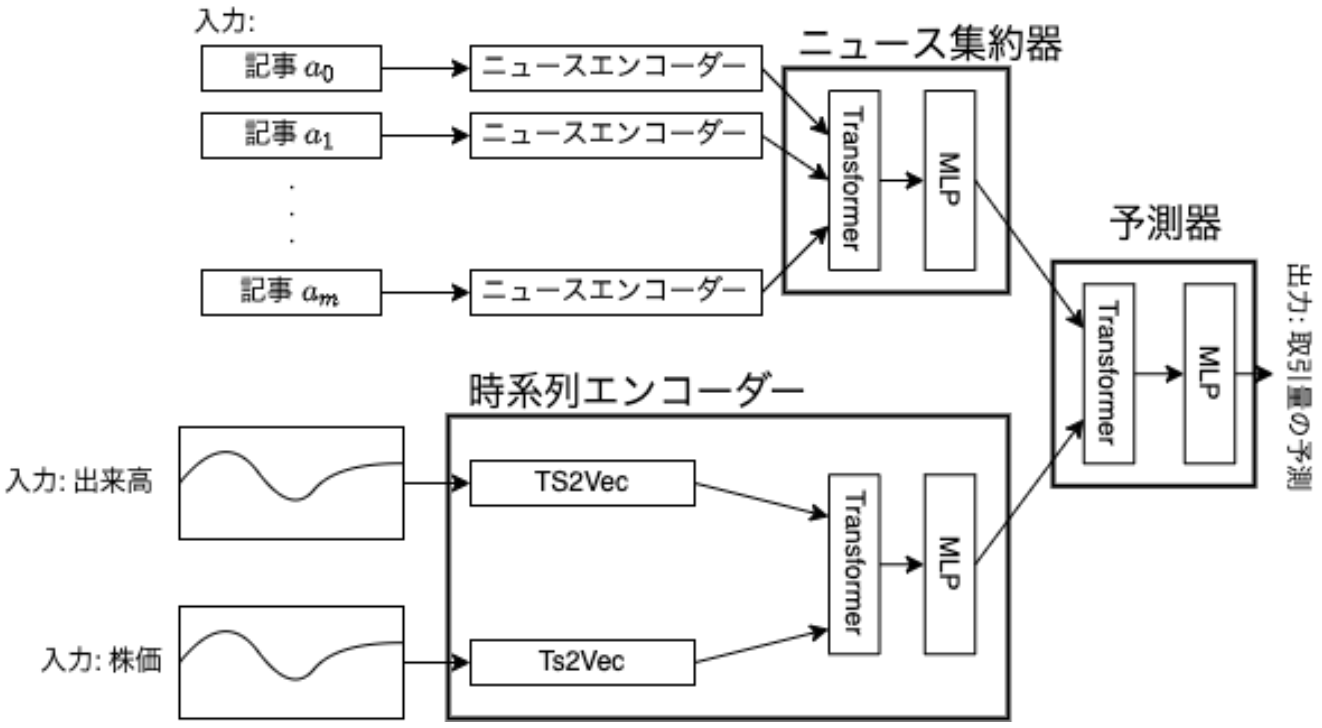

図 2　マルチモデル．ニュースエンコーダーは，シングルモデルで学習したニュースエンコーダーを利用する．

とする．ここで，各 v_i は $\mathbf{v}_{news}$ の要素である．

（6） $\mathbf{v}_{news}$ にシグモイド関数を作用させる．

b） 時系列エンコーダー

次に，時系列データの処理の流れは以下の通りである．

（1） まず，duration は，入力する時系列データの期間とし，前処理を施した出来高データ，株価データをそれぞれ

$$\mathbf{volume} = (volume_0, \ldots, volume_{duration-1})$$

$$\mathbf{price} = (price_0, \ldots, price_{duration-1})$$

とする．ここで，各 $volume_i$ は **volume** の要素を表し，各 $price_i$ は **price** の要素を表す．

（2） 出来高データと株価データそれぞれを，TS2Vec を

用いて変換し，ベクトルの集合

$$V_{volume} = (\mathbf{ts}_0^{volume}, \ldots, \mathbf{ts}_{duration-1}^{volume})$$

$$V_{price} = (\mathbf{ts}_0^{price}, \ldots, \mathbf{ts}_{duration-1}^{price})$$

$$\mathbf{ts}_i^{volume} = TS2Vec(volume_i)$$

$$\mathbf{ts}_i^{price} = TS2Vec(price_i)$$

を生成する．ここで，各 $\mathbf{ts}_i^{volume}$ は V_{volume} の要素を表し，各 $\mathbf{ts}_i^{price}$ は V_{price} の要素を表す．

（3） 生成された V_{volume} と V_{price} を連結して，

$$V_{timeseries} = (\mathbf{ts}_0^{volume}, \ldots \quad, \mathbf{ts}_{duration-1}^{volume}, \mathbf{ts}_0^{price}, \ldots, \mathbf{ts}_{duration-1}^{price})$$

とする．

（4） 生成されたベクトルの集合 $V_{timeseries}$ の各要素 $\mathbf{v}_i$ を Transformer のエンコーダー層に入力し，別のベクトルの集合

$$V'_{timeseries} = (\mathbf{ts'}_0^{volume}, \ldots \quad, \mathbf{ts'}_{duration-1}^{volume}, \mathbf{ts'}_0^{price}, \ldots, \mathbf{ts'}_{duration-1}^{price})$$

に変換する．

（5） 変換してできたベクトルの集合 $V'_{timeseries}$ の各要素を連結し，一つのベクトル $\mathbf{v}_{timeseries}$ とする．

（6） $\mathbf{v}_{timeseries}$ を多層パーセプトロンに入力して，l_{ts} をベクトルのサイズとし，

$$\mathbf{v}_{ts} = (v_0, \ldots, v_{l_{ts}-1})$$

とする．

（7） $\mathbf{v}_{ts}$ にシグモイド関数を作用させる．

c） 予 測 器

生成されたこれらのベクトルを用いて，予測値を出力する処理の流れは以下の通りである．

（1） まず，$\mathbf{v}_{news}$ と $\mathbf{v}_{ts}$ を連結したベクトルの組を

$$V_{nt} = \{\mathbf{v}_{news}, \mathbf{v}_{ts}\}$$

とする．

（2） 生成されたベクトルの集合 V_{nt} の各要素を Transformer のエンコーダー層に入力し，別のベクトルの集合

$$V'_{nt} = \{\mathbf{v}'_{news}, \mathbf{v}'_{ts}\}$$

に変換する．

（3） 変換してできたベクトルの集合 V'_{nt} の各要素を連結し，一つのベクトル $\mathbf{v}$ とする．

（4） $\mathbf{v}$ を多層パーセプトロンに入力して，サイズ 3 のベクトル

$$\mathbf{v}_{predict} = (decrease, flat, increase)$$

に変換する．ここで，$decrease, flat, increase$ は $\mathbf{v}_{predict}$ の要素を表す．それぞれ減少，変化なし，増加に関する尤度を表している．

（5） $\mathbf{v}_{predict}$ にソフトマックス関数を作用させ，

$$\mathbf{v}_{normalizedpredict} = (\mathrm{ndecrease}, \mathrm{nflat}, \mathrm{nincrease})$$

$$0 \leqq \mathrm{ndecrease}, \mathrm{nflat}, \mathrm{nincrease} \leqq 1$$

$$\mathrm{ndecrease} + \mathrm{nflat} + \mathrm{nincrease} = 1$$

を生成する．ここで，$ndecrease, nflat, nincrease$ は $\mathbf{v}_{normalizedpredict}$ の要素を表す．それぞれ減少，変化なし，増加に関する確率を表している．

（6） 最後に，予測値 P を以下の式で計算する．

$$P = argmax(\mathbf{v}_{predict})$$

$P = 0$ なら減少，$P = 1$ なら変化なし，$P = 2$ なら増加と予測しているとする．

3.2.2 マルチモデル: 市場が閉まっている場合

この節では，市場が閉まっているときに発行されたニュースを用いた分析手法について説明する．入力は，ある企業の市場が閉まっている間に発行された複数のニュース記事，出来高と株価の時系列データである．処理の流れは以下の通りである．この時，各ニュース記事をベクトルに変換するエンコーダとして，シングルモデルで学習が行われた，ニュースエンコーダーを利用してそれぞれのニュースベクトルを生成して，Transformer のエンコーダー層と MLP からなる集約器を経て，記事集合のベクトルを生成する．

時系列データに関しては，シングルモデルの場合と同様の処理を行う．生成されたこれらのベクトルを用いて，予測値を出力する処理の流れは以下の通りである．

（1） 複数のニュース記事を集約して生成したベクトルを $\mathbf{v_a}$, 時系列データのベクトルを $\mathbf{v_{ts}}$ とする．これらを合わせて

$$V_{at} = \{\mathbf{v}_a, \mathbf{v}_{ts}\}$$

とする．

（2） 生成されたベクトルの集合 V_{at} の各要素を Transformer のエンコーダー層に入力し，別のベクトルの集合

$$V'_{at} = \{\mathbf{v}'_{news}, \mathbf{v}'_{ts}\}$$

に変換する．

（3） 変換してできたベクトルの集合 V'_{at} の各要素を連結し，一つのベクトル $\mathbf{v}$ とする．

（4） $\mathbf{v}$ を多層パーセプトロンに入力して，サイズ 3 のベクトル

$$\mathbf{v}_{predict} = (decrease, flat, increase)$$

に変換する．ここで，$decrease, flat, increase$ は $\mathbf{v}_{predict}$ の要素を表す．それぞれ減少，変化なし，増加に関する尤度を表している．

（5） $\mathbf{v}_{predict}$ にソフトマックス関数を作用させ，

$$\mathbf{v}_{normalizedpredict} = (\mathrm{ndecrease}, \mathrm{nflat}, \mathrm{nincrease})$$

$$0 \leqq \mathrm{ndecrease}, \mathrm{nflat}, \mathrm{nincrease} \leqq 1$$

$$\mathrm{ndecrease} + \mathrm{nflat} + \mathrm{nincrease} = 1$$

を生成する．ここで，$ndecrease, nflat, nincrease$ は $\mathbf{v}_{normalizedpredict}$ の要素を表す．それぞれ減少，変化なし，増加に関する確率を表している．

（6） 最後に，予測値 P を

$$P = argmax(\mathbf{v}_{predict})$$

とし，$P = 0$ なら減少，$P = 1$ なら変化なし，$P = 2$ なら増加と予測しているとする．

3.3 提案モデルの学習

モデルの学習に用いるデータ項目及び例は表 1 の通りである．本研究では，ニュースデータ，株価データ，そして出来高データの 3 つのデータを用いる．株価データと出来高データは，ティックデータから抽出する．これは，馬場 [15] の研究で使用されていたデータを利用する．取引の履歴データには歩み値データを用いる．取引が成立した時点での株価，株数，成立した時刻のデータが企業ごとに格納されているので，銘柄コードで抽出することにより，各企業の株価と出来高に関する時系列データを取得することができる．

ニュース記事のデータは日経 QUICK ニュース社 [1] の記事を用いる．このニュースデータは，国内外の市場に関する情報を集約したものであり，日本国内の株式市場に影響を与える情報が多いと考えられる．東証場況や個別銘柄解説，株式市場や債券・為替など金融市場の場況の解説記事が主となっている．また，株式に関する速報に強みがあり，国内の市況報道に大きな影響力を持つ．これは，馬場 [15] の研究で使用されていたものを利用する．

3.3.1 対象企業

本研究の学習に用いる企業として，日経平均株価を構成する 225 銘柄を用いた．理由としては，企業の業種として，医薬品，電気機器，銀行，食品，サービス，商社，機械，不動産，鉄道・バス，電力など多様なものが存在しており，これにより学習の精度が向上すると考えたためである．

3.3.2 ニュース記事

まず，ニュース記事のデータから，シングルモデルの場合市場が開いている時間帯，マルチモデルの場合市場が閉まっている時間帯に発行された企業ごとのニュースを抽出する．本研究では，ニュース記事の本文にその企業の名前が記載されているものを，その企業のニュースとした．

3.3.3 時系列データの抽出

企業ごとの，株価と出来高に関する時系列データを抽出する．一つの企業に対して，l 分ごとの値の平均をとり，$duration$ 分のデータを一つの時系列データとした．

表 1 データセットの項目と例

	項目	例
ニュース記事	タイムスタンプ	2015-02-05 15:42:38
	本文	エプソンの前期、連結最終益 1125 億円...
取引履歴	株価	1769.0, 1765.0, 1770.0, ...
	出来高	100, 1100, 100, ...

3.3.4 教師データの作成

シングルモデルの場合，ニュースの発行時刻の t_{before} 分前から発行時刻までの出来高の合計に対して，ニュース発行時刻から t_{after} 分後の出来高の合計が，増加しているか，減少しているか，変化していないかを調べてラベル付けする．また，変化が ϵ%以内であった場合，変化していないとみなす．マルチモデルの場合，市場が閉まる時刻の t_{before} 分前から市場が閉まる時刻までの出来高の合計に対して，次に市場が開く時刻から t_{after} 分後の出来高の合計が，増加しているか，減少しているか，変化していないかをラベル付けする．また，変化が ϵ%以内であった場合，変化していないとみなす．

このようにして作成した学習データを提案モデルに入力し，学習を行う．活性化関数はシグモイド関数，損失関数は交差エントロピー誤差，最適化関数は確率的勾配降下法を用い，ニュースエンコーダ及び時系列エンコーダの MLP は 3 層，予測器の MLP は 2 層とする．

4 評価実験

4.1 データセット

実験に用いるデータセットは，個別銘柄ごとの株取引の履歴およびニュース記事である．シングルモデルは 2015 年 1 月 1 日から 2016 年 10 月 31 日までのデータのうち，12000 個のニュース記事及びそれに対応する株取引データを学習データとして用い，残りの 1557 個のニュースと出来高のペアをテストデータとして用いる．マルチモデルは 2015 年 1 月 1 日から 2016 年 10 月 31 日までのデータのうち，5000 個のニュース記事及び 2208 個の取引データを学習データとして用い，1215 個のニュース記事及びそれに対応する 470 個の取引データをテストデータとして用いる．マルチモデルにおいてニュースデータの数と取引データの数が異なるのは，一つの取引データに対して複数のニュース記事が対応しているためである．実験対象の銘柄のそれぞれの業種の企業数について表 2 に示す．実際の企業名に関しては，日経平均プロフィルの Web ページ [2] に掲載されている．

4.1.1 取引データの前処理

本研究に用いる時系列データである出来高データや株価データは，銘柄によって値が大きく異なることがある．そのため，実際の値をそのまま使用するとうまくいかないと考えられる [15]．そこで，時系列データの前処理として，一つ前の値との比率をとったもの，一つ前の値との差分を取ったものでそれぞれ実験

1：http://www.nqn.co.jp/

2：https://indexes.nikkei.co.jp/nkave/index/component?idx=jpxnkms

表 2　実験対象の銘柄の業種

業種	個数	業種	個数
医薬品	9	電気機器	29
自動車	10	精密機器	5
通信	5	銀行	11
その他金融	2	証券	3
保険	5	水産	2
食品	11	小売業	7
サービス	14	鉱業	1
繊維	4	パルプ・紙	2
化学	17	石油	2
ゴム	2	窯業	8
鉄鋼	4	非鉄・金属	10
商社	7		
建設	9	機械	15
造船	2	その他製造	4
不動産	5	鉄道・バス	8
陸運	2	海運	3
空運	1	倉庫	1
電力	3	ガス	2

を行なった.

一つ前の値との比率を取る場合，時系列データを

$$(x_0, x_1, x_2, \ldots, x_T)$$

とすると，正規化された時系列データは,

$$(1, \frac{x_1}{x_0}, \frac{x_2}{x_1}, \ldots, \frac{x_T}{x_{T-1}})$$

となる.

一つ前の値との差分を取る場合,

$$(x_0, x_1, x_2, \ldots, x_T)$$

とすると，正規化された時系列データは,

$$(0, x_1 - x_0, x_2 - x_1, \ldots, x_T - x_{T-1})$$

となる.

4.1.2　評 価 方 法

モデルの評価尺度としては，テストデータにおける，予測したラベルと正解のラベルを比較して計算される精度を用いる．今回の実験では，$l = 1, duration = 30, t_{\text{before}} = t_{\text{after}} = 5, \epsilon = 5$ とする．ただし，このパラメータが最適かどうかは現段階では不明であり，今後の課題としてパラメータの調整が挙げられる.

4.2　比 較 手 法

データの適切な前処理方法，及び学習の適切な手順を分析するため，実験にあたり複数の手法で学習し，比較・分析を行う．その手法は以下の通りである.

- シングルモデルの学習を行った後，そのエンコーダーを用いてマルチモデルの学習を行う場合

（1） データの前処理を行わない場合

（2） データの前処理として，一つ前の値との比率をとったものを用いる場合

表 3　シングルモデルの学習を行った後，そのエンコーダーを用いてマルチモデルの学習を行う場合

正解率 (%)	前処理の方法		
	前処理なし	比	差分
シングルモデル	42.952	48.128	46.586
マルチモデル (株価のみ)	3.7415	85.884	10.374
マルチモデル (出来高のみ)	85.884	10.374	10.374
マルチモデル (出来高+株価)	85.884	85.884	85.884

表 4　シングルモデルの学習を行わず，直接マルチモデルの学習を行う場合

正解率 (%)	前処理の方法		
	前処理なし	比	差分
マルチモデル	85.884	85.884	85.884

（3） データの前処理として，一つ前の値との差分をとったものを用いる場合

- シングルモデルの学習を行わず，直接マルチモデルの学習を行う場合

（1） データの前処理を行わない場合

（2） データの前処理として，一つ前の値との比率をとったものを用いる場合

（3） データの前処理として，一つ前の値との差分をとったものを用いる場合

4.3　実 験 結 果

実験結果は表 3 と表 4 の通りである．実験の結果，シングルモデルに関しては，前処理を行わないものが最も精度が悪く，前処理として一つ前の値との比率をとったものが最も良い結果となった．マルチモデルに関しては，前処理およびシングルモデルによる事前学習の有無にかかわらず，全て同じ値となった．これは，市場が閉まる直前と開く直後では，出来高が減少する場合が多く，学習データのラベルに偏りが生じてしまったため，このような結果になったと考察できる．今後の課題として，パラメータの調整及びアーキテクチャの見直しが挙げられるため，それらについて研究を進める予定である.

5　おわりに

5.1　ま と め

本研究では，ニュース記事のテキストデータ，株価と出来高の時系列データを用いて，ニュース発行前と発行後の出来高の変化を予測する手法を提案した．また，ニュースは，取引時間内に発行されたものと取引時間外に発行されたもので分類し，使用するモデルをそれぞれシングルモデル，マルチモデルとした．マルチモデルではシングルモデルでファインチューニングを行なったニュースエンコーダーを利用し，複数のニュース記事を集約することができる．実験結果としてシングルモデルに関しては，時系列データの前処理の方法によって精度が異なる

ことがわかった．前処理を行わないものが最も精度が悪かったため，時系列データを扱う際には前処理を行うことが重要であると考えられる．

マルチモデルに関しては，シングルモデルを用いて事前学習を行う場合と，シングルモデルで事前学習を行わずに直接マルチモデルを学習した場合で同じ結果になった．これは，一般的に市場が閉まる直前よりも，市場が開く直後の方が出来高が少ないことの方が多く，学習データにもそのようなデータが多く含まれていたため，偏った学習結果になってしまったためと考えられる．

5.2 今後の課題

5.2.1 大規模な実験と手法の改善

さらなる実験と分析を行い，課題を明らかにする．パラメータのチューニング，アーキテクチャの再検討などによる手法の改善を行う予定である．

5.2.2 取引データの前処理

本研究では，株価及び出来高に関する時系列データの前処理として，一つ前の値との比率をとったもの，一つ前の値との差分を取ったものでそれぞれ実験を行なった．また，前処理を行なっていない時系列データに関しても実験を行った．しかし，時系列データの前処理にはそのほかにも様々な種類がある．例えば，対数変換を行うものや，比を取った後に，ロジット変換を行うものなどがある．今回実験に用いなかったこれらの前処理を施すことによって，より精度の良い結果が出る可能性がある．

5.2.3 シングルモデルによる事前学習の効率化

本研究における実験では，マルチモデルにおける予測の結果が，シングルモデルを用いて事前学習を行う場合と，シングルモデルで事前学習を行わずに直接マルチモデルを学習した場合で同じになった．この理由として，マルチモデルに関する学習データのラベルに偏りがあったことの他に，シングルモデルの事前学習が十分に行なえていなかった可能性が考えられる．そこで，シングルモデルにおける学習データのデータ数を増やす，時系列データの期間を長くする，または提案モデルのパラメータを調整するなどして，より効果的な事前学習を行うことによって，マルチモデルにおいても，シングルモデルの事前学習が適切に反映されると考えられる．

5.2.4 取引規模の小さい銘柄への対応

本研究の実験で用いた銘柄は，日経平均株価を構成する 225 銘柄であった．これを用いた理由は，様々な業種の銘柄が揃っているためであったが，これらの銘柄は，規模の大きい企業である場合が多く，取引が盛んに行われている銘柄が多い．そのため，1 日あたりの取引があまり多くないような，中小企業などの出来高は，提案手法でうまく予測できるかは現段階で不明である．そこで，評価に用いるデータセットを本研究とは異なるものを用いるなどして，そのような銘柄に対しても安定して精度の良い予測ができるようなモデルを構成することが課題として挙げられる．

謝　　辞

本研究の一部は科研費（19H04116）による．

文　　献

[1] W. Huang, Y. Nakamori, and S. Wang. Forecasting stock market movement direction with support vector machine. COR, 32(10), pages 2513–2522, 2005.

[2] Y. Xu and S. Cohen. Stock movement prediction from tweets and historical prices. ACL, volume 1, pages 1970–1979, 2018.

[3] 許 蔚然, 江口 浩二. トピック埋め込み回帰モデルを用いた株価予測. 第 24 回人工知能学会金融情報学研究会 (SIG-FIN) 予稿集, 2020.

[4] Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. Enhancing stock movement prediction with adversarial training. IJCAI 2019, Macao, pages 5843–5849, 2019.

[5] ニュース指標による株価市場の予測可能性, 証券アナリストジャーナル, Vol. 52, No. 4, pages 67-75, 2014.

[6] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERTnetworks. Empirical Methods in Natural Language Processing(EMNLP), pages 3982–3992, 2019.

[7] Devlin, J. et al.BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805, 2018.

[8] Chicco, Davide. Siamese neural networks: an overview, Artificial Neural Networks, pages 73–94, 2020.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, Advances in Neural Information Processing Systems, pages 6000–6010, 2017.

[10] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong and Bixiong Xu. TS2Vec: Towards Universal Representation of Time Series, arXiv preprint arXiv:2106.10466, 2021.

[11] Tetlock, P. C. Giving content to investor sentiment: The role of media in the stock market, The Journal of Finance, Vol. 62, No. 3, pages 1139–1168, 2007.

[12] Bollen, J., Mao, H. and Zeng, X. Twitter mood predicts the stock market, Journal of computational science, Vol. 2, No. 1, pages 1–8, 2011.

[13] W. Li, R. Bao, K. Harimoto, D. Chen, J. Xu, and Q. Su. Modeling the stock relation with graph network for overnight stock movement prediction IJCAI2020, pages 4541–4547, 2020.

[14] 松橋 志拓. 株取引に対するニュースの影響分析, 京都大学卒業論文, 2019.

[15] 馬場 慧, 馬 強. 株価とニュース報道を用いた上場企業の暗黙関係の発見, 第 8 回データ工学と情報マネジメントに関するフォーラム (DEIM2016), 2016.

[16] 五島圭一, 高橋大志, 寺野隆雄: ニュースのテキスト情報から株価を予測する, 2015 年度人工知能学会全国大会講演集, 2015.