過去の類似ニュースに基づいたニュース記事からの株価変動予測

坂本　悠輔†　　山本　岳洋†

† 兵庫県立大学 大学院情報科学研究科　〒 651–2197 兵庫県神戸市西区学園西町 8-2-1
E-mail: †ad21t038@u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp

あらまし　本研究では，ある企業に関するニュース記事が与えられた際に，その企業の翌日の株価変動を予測する問題に取り組む．具体的には，その企業の翌日の超過収益率を予測の対象とする．そのために，与えられたニュース記事のみから予測を行うのではなく，その記事に類似する過去のニュース記事を検索し，それらの記事の翌日の株価変動も考慮し，予測を行う手法を提案する．実験では，TOPIX Core30 の企業群を対象に，日経速報ニュースをコーパスに用いて提案手法の有効性を検証した．評価尺度として，実際の超過収益率と予測との RMSE，超過収益率を正か負に 2 値化した際の F_1 値などを用いて性能を評価した．ベースラインとして，与えられたニュース記事の本文のみから株価変動を予測するモデルを実装し，提案手法との比較を行った．ベースラインと提案手法の比較を行った結果，ベースライン手法，提案手法ともに翌日の超過収益率の予測精度の差があまり見られない結果となった．この結果をもとに，原因と考察を行った．

キーワード　情報検索，自然言語処理，BERT，株価変動予測

1　はじめに

企業や社会に関して日々発信されるニュースは，企業の株価の変動を予測する上で重要な情報源の 1 つである．たとえば，REUTERS[1]や REFINITIV EIKON[2]といった金融や経済などのニュースを扱ったサービスが存在している．また，ニュース記事を用いて株価の変動を予測する研究もいくつか提案されている [12] [8] [13]．本研究では，ある企業に関するニュース記事が与えられた際に，その企業の翌日の株価変動を予測する問題に取り組む．

企業によって株価の変動に影響を与えるニュース記事は異なると考えられる．たとえば，トヨタに関するニュース記事として，『プリウス PHV，電気で 60 キロ走行トヨタが新型公開』というニュース記事，ソフトバンクに関するニュース記事として，『ソフトバンク，中国で「ペッパー」販売』という 2 つのニュース記事について考える．この 2 つの企業のニュース記事を比較した際に，トヨタのニュース記事の場合は，新車の発売というイベントが何度も起こっているため，このイベントが起こったことによる株価の変動は，あまり見られないと考えられる．しかし，ソフトバンクのニュース記事では，ロボット販売という滅多に起こらないイベントに関するニュースであるため，イベントが起こった際に，株価の大きな変動が見られる可能性があると考えられる．

このように，企業によってどのような記事が株価に影響を与えるかを知るためには，その企業に関する過去の類似ニュース記事が重要ではないかと考えた．たとえば，先述したトヨタやソフトバンクのニュース記事の例であれば，過去に類似するニュース記事が出た際に，株価がどのように変動したのかを知ることで，対象としている記事がどのように株価に影響を与えるかを知ることができると考えられる．さらに，過去のニュース記事を用いることで特定の企業の株価変動の予測精度が向上するのであれば，そのような過去のニュースを投資家に提示することで，投資の有用な判断材料になると考えられる．

そこで本研究では，与えられたニュース記事のみから予測を行うのではなく，その記事に類似する過去のニュース記事を検索し，それらの記事の翌日の株価変動も考慮し，予測を行う手法を提案する．

評価実験では，日経速報ニュースをニュースの情報源とし，TOPIX Core30 の企業群を対象に，ニュース記事が出た翌日の超過収益率の変動予測精度を検証する．評価尺度としては，超過収益率に対する MSE や RMSE，超過収益率の増減に対する F_1 値を用いる．ベースラインとして，与えられたニュース記事の本文のみから超過収益率を予測するモデルを BERT [5] をファインチューニングすることで構築する．

本論文の構成は以下のとおりである．2 節ではニュース記事を用いた株価変動予測や，その他の情報源を用いた株価変動予測に関する関連研究について述べる．3 節では提案手法である，過去の類似ニュース記事発見による株価変動予測手法について述べる．4 節では本研究で行った評価実験およびベースライン手法について，そして評価実験を行って得られた結果について述べる．5 節では本研究のまとめと今後の課題について述べる．

2　関連研究

2.1　ニュース記事を用いた株価変動予測や分析

ニュース記事を用いた株価変動予測や，関連する分析を行う研究がこれまで提案されている．たとえば，Hu らはニュース記事から株価の変動を予測するためのモデルとして，ニュース記事の系列が株価に与える影響やニュース記事ごとの多様な影

1：https://www.reuters.com/
2：https://www.refinitiv.com/ja/products/eikon-trading-software

響をとらえるための Hybrid Attention Networks（HAN）を提案している [8]．Tang らは，予測したい金融商品に対して明示的に言及しているニュース記事だけでなく，暗黙的にその商品に関連するニュース記事を見つけ，予測に用いる手法を提案している [12]．Du らは，株価のベクトル表現を用いて, ニュース記事の影響力を符号化する方法を提案している [6]．また，米田らはニュース記事が株価に与える影響力の推定を行う手法を提案している．彼らの手法は，株価指数だけから株価指数の変動を予測するモデルと，株価指数とニュース記事の両者を用いて変動を予測するモデルの出力の差をみることで，ニュース記事が株価に与える影響力を推定している [18]．Chen らはニュース記事既存の株価変動予測の研究において，多様なイベントタイプに特定の意味情報が失われていることから，より細かい粒度のイベントを株価予測に取り入れることを提案している [3]．前川らは，すでに公開されているニュース記事から，日経 225 先物を予測するような確率モデルを構築し，有効性を検証している [16]．Liu らは，ニュースのタイトルと内容の中に含まれる補完的な情報を株価変動予測の際に取り込むための Hierarchical Complementary Attention Network （HCAN）を提案している [9]．

また，株価変動予測を行った研究ではないが，関連する研究としては，Fan らの研究がある．彼らは，金融に関するニュースに付与された細粒度のカテゴリを精度よく分類するための手法を提案し，実際に中国の大手金融情報サービスプロバイダーの金融ニュースを対象に性能を評価している [7]．

2.2 その他のテキスト情報を用いた株価変動予測や分析

ニュース記事以外にも，さまざまな情報源を用いて，株価変動を予測する研究がなされている．たとえば, 磯貝らは有価証券報告書の『生産・受注および販売の状況』からサプライヤ・カスタマ関係を抽出し, 互いの企業ニュースが株価へ適切に反映されているかを調査し, カスタマ企業のニュースがサプライヤ企業の株価へ適切に反映されていないことを報告している [19]．指田らは決算短信から構築した因果チェーンに対して SSESTM（Supervised Sentiment Extraction via Screening and Topic Modeling）モデルを適用し，企業間のセンチメントを考慮したリードラグを収益機会とする投資戦略に関する提案及び実証実験を行っている [15]．Chen らは，BERT-based Hierarchical Aggregation モデルを提案し, 多数の金融ニュースを要約することで，EX 市場変動予測を行っている [2]．

Si らは Twitter からのセンチメントをもとにしたトピックを活用して株式市場予測を行う手法を提案している．また，S&P100 の指数に対して，実験を行った結果，既存の最先端の非トピックベースの手法よりも Si らのアプローチが有効で，良いパフォーマンスを示すことが示されている [11]．

Bollen らは，ポジティブな気分とネガティブな気分を測定することができる 2 つの測定機を用いて，毎日の Twitter のテキスト内容を分析している．ここで得られた測定値を利用し，時間の経過と共にダウ平均株価の値と相関があるかどうかを調査している [1]．

Chen らは，企業の株価予測のために，その企業の関連企業の情報取り入れることを提案している．中国本土の株式市場のデータを用いた実験から， Chen らのモデルで学習した表現が企業間の関係を捉え，関連企業の情報を加味した予測モデルが株式市場の予測を正確に行うことができることを示している [4]．

Sawhney らは，金融データやソーシャルメディア，株式の関係から得られる時間のシグナルと，グラフニューラルネットワークを用いた手法を提案している [10]．

Xu らは，既存の event-driven 型手法の 2 つの欠点である，銘柄に依存するイベント情報の影響を見落としている点と，関連する他の銘柄のイベント情報の影響を見落としているという点を踏まえ，これらを解決するために Relational Event-driven Stock Trend forecasting (REST) framework を提案している．また，この提案したフレームワークを用いて投資シミュレーションを行った結果，ベースラインよりも高い投資収益率を出すことを示している [13]．

中川らは，日時の株価変動と月初の値との比で表現された月間の株価変動に対し，音声認識の分野などで使用されている Dyanamic Time Warping（DTW）距離を適用した Indexation DTW（IDTW）による手法を提案している [17]．

3 過去の類似ニュース発見に基づく株価変動予測

本節では提案手法について述べる．まず，本研究で取り組む問題について定義する．その後，本研究の提案である，過去の類似ニュース発見に基づいた株価変動予測手法について説明する．

3.1 問 題 定 義

本研究では，ある企業に関するニュース記事が掲載されたとき，その企業の翌日の株価変動を予測することである．具体的には，ある日付 t においてある企業 c のニュース記事 $d_{t,c}$ が与えられたとき, ニュース記事 $d_{t,c}$ を用いて，対象の企業の翌日 $t+1$ における超過収益率を予測する問題に取り組む.

超過収益率について説明するために，まず収益率を説明する．収益率とは, ファイナンス分野の株価の研究において頻繁に使用される概念である. 期の単位は任意に設定することができる. 例えば，期を日と設定すると，日次収益率となる．日以外でしばしば利用される期として週や月がある．本研究では，翌日の収益率を対象とする．ある企業 c の日付 t における終値を $\mathrm{close}_{c,t}$，翌日における終値を $\mathrm{close}_{c,t+1}$ とすると，収益率（日次収益率）$R_{c,t}$ は以下の式で表される．

$$R_{c,t} = \frac{\mathrm{close_{c,t+1}} - \mathrm{close_{c,t}}}{\mathrm{close_{c,t}}} \tag{1}$$

次に超過収益率について説明する．超過収益率 [14] とは，企業の収益率を，日経 225 や TOPIX などを用いて計算された収益率も考慮して，相対的に評価するというものである．$R^b_{c,t}$ を日経 225 や TOPIX などを用いた収益率とすると，超過収益率

$R^a_{c,t}$ の式は以下のように定義できる．

$$R^a_{c,t} = R_{c,t} - R^b_{c,t} \tag{2}$$

本研究では，株価指標として，日経 225 を用いる．日経 225 とは，東証一部上場銘柄 2,000 銘柄以上のうち，代表的な 225 銘柄で構成されたもののことである．上記式から分かるように，日経 225 の収益率に比べて，対象の企業の収益率が高い場合は正の値になり，たとえ対象の企業の収益率が正の値でも，日経 225 の収益率がそれよりも高い場合は，超過収益率は負となる．この超過収益率は，ニュース記事の翌日において，市場から受けたこの企業特有の影響が含まれていると考えられる．ニュース記事および過去のニュース記事を用いてこの超過収益率を予測することが本研究の目的である．また，本研究ではこの超過収益率の変動のことを株価変動と呼ぶ．

3.2 提案手法の流れ

本節では，手法の流れを述べる．提案手法は，ある企業の株価変動予測をする際に，その企業のある日のニュース記事と，その記事に類似しているその企業過去のニュース記事を用いて翌日の超過収益率の予測を行う．本手法の流れは以下のとおりである．

（1） 日付 t に報道された，企業 c のニュース記事 $d_{t,c}$ が与えられる．

（2） 記事 $d_{t,c}$ のみから翌日 $t+1$ の超過収益率を予測する．このモデルについては 3.3 節で述べる．このモデルで得られた超過収益率を $R^a_{c,\mathrm{present}}$ とする．

（3） 記事 $d_{t,c}$ にテキストとして類似する，企業 c に関する t より過去のニュース記事集合 $D_{c,\mathrm{past}}$ を取得する．どのように類似ニュースを取得するかについては，3.4 節で述べる．

（4） $D_{c,\mathrm{past}}$ の各記事 $d_{c,\mathrm{past}} \in D_{c,past}$ について，その記事の翌日の超過収益率を $R^a_{d_{c,\mathrm{past}}}$ とするとき，以下の式で，企業 c の超過収益率 $R^a_{c,\mathrm{past}}$ を予測する．

$$R^a_{c,\mathrm{past}} = \frac{1}{|\,D_{c,\mathrm{past}}|} \sum_{d_{c,\mathrm{past}} \in D_{c,\mathrm{past}}} R^a_{d_{c,\mathrm{past}}} \tag{3}$$

ここで，$|\,D_{c,\mathrm{past}}|$ は集合 $D_{c,\mathrm{past}}$ の要素の個数である．

（5） （2）で求めた超過収益率 $R^a_{c,\mathrm{present}}$ と（4）で求めた超過収益率 $R^a_{c,\mathrm{past}}$ を用いて，以下の式に従い最終的な超過収益率 $\hat{R}^a_{c,t}$ を予測する．

$$\hat{R}^a_{c,t} = \lambda R^a_{c,\mathrm{present}} + (1-\lambda) R^a_{c,\mathrm{past}} \tag{4}$$

ただし λ は $\lambda \in [0,1]$ の定数である．

$\lambda = 0$ のとき，過去の類似ニュース記事のみから，翌日の超過収益率予測をすることになる．また，$\lambda = 1$ のとき，与えられたニュース記事のみから翌日の超過収益率予測をすることになる．

3.3 ニュース記事からの超過収益率予測モデル

3.2 節で述べた，ニュース記事単体のみから超過収益率を予測するモデルには自然言語処理の分野で用いられる BERT [5] を用いて行う．この BERT をファインチューニングすることで構築する．ファインチューニングのために用いるモデルを図 1 に示す．図 1 にあるように，ニュース記事の本文を入力として受け取り，[CLS] トークンに対応する BERT の最終層に全結合層の出力層を加え，翌日の超過収益率を出力する．また，損失関数としては MSE（平均二乗誤差）を用いる．4 節で述べる訓練データの各ニュース記事と翌日の超過収益率を学習データとすることで，与えられたニュース記事のみから，その企業の翌日の株価変動を学習するモデルを構築できると考えた．

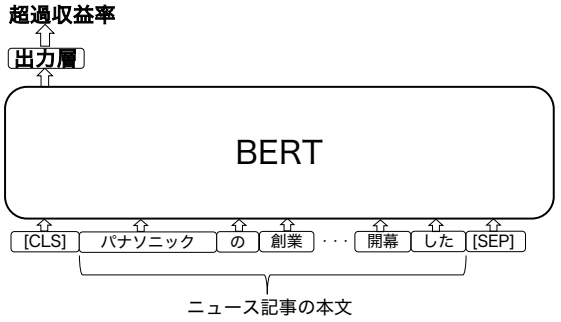

図 1 BERT を用いたニュース記事の本文からの超過収益率予測モデル．

3.4 類似ニュースの検索

3.2 節の（2）で述べたように，ある企業に関して与えられたニュース記事に対して，その企業の過去のニュース記事を取得する．本研究では，BM25 を用いて，与えられたニュース記事の本文が類似するニュース記事を取得する．実装としては，Elasticsearch のランキング関数を用いる．

4 実　　験

本節では本研究で行った実験について述べる．まず，実験に用いるデータについて述べる．次に，提案手法の有効性を検証するためのベースライン手法について述べ，実験で使用する評価尺度と実験設定について説明する．さらに，実験を行って得られた結果について説明し，その結果をもとに考察を行う．

4.1 対象の企業と使用するデータ

ニュース記事が比較的多く存在すると考えられる企業を対象に有効性を検証するため，本研究では TOPIX Core30 の企業群を対象に提案手法の有効性を検証する．TOPIX Core30 とは，TOPIX を構成する銘柄の中でも特に流動性と時価総額が高いとされる 30 の銘柄を算出対象とした株価指数のことである．2018 年 12 月時点の，TOPIX Core30 に含まれる 30 件の企業を予測の対象とする．この企業群には，たとえば，パナソニック，ソニー，トヨタ自動車といった企業が含まれる．

次に，本研究で使用するデータについて説明する．ニュース記事として, 日経速報ニュースアーカイブを使用する．日経速報ニュースは,『「日経電子版」の”速報”で提供している各ジャンルのニュースに、プレスリリース、人事ニュース・人事異動情報を加えたもの』[3]であり，日経速報ニュースアーカイブは日経速報ニュースの過去の記事がアーカイブされたデータである．日経速報ニュースを利用することで，速報性の高い記事を利用することができる．日経速報ニュースでは，記事の掲載時刻，見出し，本文という情報のほかに，ニュース記事に付与された分類コードが含まれている．本研究では,「主要」,「経済」,「政治」,「産業」,「マーケット」,「国際」,「社会」のいずれかの分類に含まれる記事であり，かつ,「人事記事」,「訃報記事」,「表彰記事」,「社説」,「社告・公告」,「インタビュー」,「調査・統計」に含まれない記事, かつこれらのニュース記事の見出しに「株価」,「東証」という言葉が入っているものを除いたニュース記事を対象とした．ニュース記事の見出しから「株価」,「東証」という言葉を除いた理由は，主に 2 点ある．1 点目は「株価」,「東証」という言葉が含まれている企業のニュース記事には，予測対象の企業に関するニュースが含まれていないことが多かったからである．実際に,『東証寄り付き、続落　下げ幅 100 円超える、欧米株安が重荷』のようなニュース記事が見受けられた．対象とするニュースの見出しから，対象とするニュース記事なのかを認識することは困難であると考えられる．2 点目は,「株価」,「東証」という言葉が含まれている企業のニュース記事は，その企業に関する「株価」,「東証」の情報しか含まれていない，もしくは企業の情報すら含まれていないからである．これらの単語を含むニュース記事の中に，企業名を主題としたようなニュース記事はあまり見受けられないと考える．

また，各ニュース記事には株式コードとよばれる，ニュースの主題である企業の銘柄コードが付与されており，これを用いて特定の企業のニュース記事のみを検索することができる．本研究では，与えられた企業のニュース記事として，この株式コードを用いる．利用するニュース記事の期間は，2016 年 4 月から 2019 年 3 月の 3 年間である．

また，株価のデータは Yahoo! Finance を用いて，2016 年 4 月から 2019 年 3 月の期間における TOPIX Core30 の各企業の株価データを取得した．取得した株価データには, 始値, 終値, 高値, 安値, 調整後終値の情報が含まれており，本研究では終値をその日の株価として扱い，収益率や超過収益率を計算した．

4.2　ベースライン手法

提案手法の有効性を検証するため，ベースラインとして 3.3 節で述べた，ニュース記事単体のみから翌日の超過収益率を予測するモデルを用いる．具体的には，各企業について，ニュース記事の見出しを用いて，翌日の超過収益率を予測する．誤差関数として MSE を用い，BERT をファインチューニングする．事前学習モデルとしては，東北大学の事前学習済み日本語 BERT モデル [4]を用いる．この時、最大トークン数を 512，バッチサイズを 8，学習率を $2e^{-5}$で行った．このベースラインと提案手法を比較することで，過去の類似ニュースを考慮することで精度改善に寄与するかどうかを検証できると考えた．なお, 本ベースライン手法を用いて予測される最終的な翌日の超過収益率は, 3. 2節の式（4）の$\lambda = 1$ の場合である．つまり，与えられたニュース記事の情報のみを使用する場合を表している．また，本ベースライン手法のことを，入力記事のみという言葉で定義する．

4.3　実験設定と評価尺度

実験では，日経速報ニュースアーカイブのデータセットに含まれる TOPIX Core 30 に関するデータセットを訓練データ，検証データ，テストデータに分割する．具体的には，それぞれ 2016 年度を 10326 件，2017 年度を 11468 件，2018 年度を 11437 件に分割する．なお，実験において予測される最終的な超過収益率は 3. 2 節の（4）の式の $\lambda = 0.5$ と λ =0 の場合である．つまり，与えられたニュース記事と，そのニュース記事に類似する過去のニュース記事を使用して翌日の超過収益率の予測を行う場合と，与えられたニュース記事に類似する過去のニュース記事のみを用いて翌日の超過収益率の予測を行う場合を表している．なお，過去の類似ニュース記事のみを用いて，翌日の超過収益率の予測行う提案手法を過去記事のみという言葉で定義する．同様に，与えられたニュース記事と，そのニュース記事に類似する過去のニュース記事を使用して翌日の超過収益率の予測を行う提案手法を，提案手法という言葉で定義する．本研究では検証データをベースライン手法で用いている．検証データは，ベースライン手法で用いる BERT のファインチューニング時に，訓練データ学習のためのエポック数を early stopping により決定するために用いた．ここで，使用するテキストとして類似する過去のニュース記事の件数は 5 件とする．株価変動予測の性能を評価するため，次の評価尺度を用いて予測精度を評価する．まず，真の超過収益率と予測された超過収益率の誤差を検証するため，MSE（平均二乗誤差）と RMSE（二乗平均平方根誤差）を用いる．それぞれの式は以下の式で表すことができる．

$$\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N}(R_n^a - \hat{R_n^a})^2 \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(R_n^a - \hat{R_n^a})^2} \tag{6}$$

上記の式における N はテストデータの個数，R_n^a は n 番目のテストデータに対する真の超過収益率，$\hat{R_n^a}$ は n 番目のテストデータに対してモデルが予測した超過収益率である．

また，回帰としての評価だけでなく，予測を二値分類としてとらえ，性能を評価する．すなわち，実際に翌日の超過収益率が正であった日に正，負であった日に負と予測できたかを評価

3：https://t21help.nikkei.co.jp/reference/cat397/post-27.html （2022 年 2 月 4 日最終閲覧）

4：https://github.com/cl-tohoku/bert-japanese

する．具体的には，正解率（Accuracy），適合率（Precision），再現率（Recall），F_1 値を用いて評価する．翌日の超過収益率が正であるときにモデルの予測が正（Positive）であるときを真陽性 TP とすると，同様に偽陽性 FP，真陰性 TN，偽陰性 FN を定義できる．これらを用いて，正解率（Accuracy），適合率（Precision），再現率（Recall），F_1 値は以下の式で求められる．

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$F_1 = \frac{2\text{PR}}{\text{P} + \text{R}} \tag{10}$$

これらの評価尺度を用いて，手法の性能を評価した．

4.4 実験結果

ベースライン手法と提案手法を用いて，翌日の超過収益率を予測し，その評価尺度の結果を示したものが表 1 である．表 1 の結果のように，提案手法である与えられたニュース記事とテキストとして類似するニュース記事を用いる手法（提案手法），テキストとして類似する過去のニュース記事のみを使用する手法（過去記事のみ），そしてベースライン手法入力記事のみ（入力記事のみ）において，予測性能の違いはあまり見られないことがわかる．また，表 1 の 2 値分類の評価 (Accuracy, Precision, Recall, F_1) の正解率（Accracy）に着目すると，提案手法では，0.500，過去記事のみでは，0.503，入力記事のみでは，0.494 となっている．このことから提案手法，ベースライン手法ともに翌日の超過収益率予測がうまくできていないことがわかる．しかし，表 1 の提案手法の MSE と RMSE に着目すると，MSE が 3.979×10^{-4}，RMSE が 1.995×10^{-2} となっており，1 番数値が小さくなっていることがわかる．

4.5 考　　察

今回，表 1 上の入力記事のみと提案手法の正解率（Accuracy）が，それぞれ 0.494，0.500 という値より，ベースライン，提案手法ともに超過収益率予測が有効に働かなかった結果になったが，その原因について考察する．まず, ベースラインの予測が上手くいっていない理由について書く．ベースラインの予測がうまくいかなかった理由は, 主に 2 点ある．1 点目は，使用したニュース記事の選別が不十分だったことである．ニュース記事の中には，その企業のニュースにもかかわらず，企業名が入っていない記事がある．そのようなニュース記事は取り除く必要性があると考える．具体例として，ソフトバンクのニュース記事がある．ソフトバンクのニュース記事の本文として『5 日のＪＰＸ日経インデックス 400 は続落．終値は前日比 72.34 ポイント安の１万 4505 ････』というものがある．このニュース記事の本文は，ソフトバンクが主題ではないものだと考えられる．本研究では，訓練データやテストデータにもこのようなニュース記事がいくつか含まれていることが考えられる．ゆえに，データセットから企業名が主題ではないようなニュース記事に該当するものは除去する必要があると考える．2 点目として，超過変動率を予測する日時が翌日という設定がふさわしくない可能性が考えられる．ニュース記事に記載されているイベントが，実際に市場に影響を与える日時として，翌日はあまり適切ではないことが予想される．イベントが株価に影響を与えるのは，1 週間後や 1ヶ月後といった長いスパンであることが考えられる．従って，長いスパンでのニュース記事に着目する必要があり，これらの期間での超過収益率予測も行う必要があると考える．

表 1　テストデータにおける各手法の予測性能.

	提案手法	過去記事のみ	入力記事のみ
MSE	3.979×10^{-4}	4.419×10^{-4}	4.028×10^{-4}
RMSE	1.995×10^{-2}	2.102×10^{-2}	2.007×10^{-2}
Accuracy	0.500	0.503	0.494
Precision	0.494	0.494	0.492
Recall	0.791	0.512	0.932
F_1	0.608	0.503	0.644

続いて，過去のニュース記事からの予測が上手くいかなかった原因について述べる．過去記事からの予測が上手くいかなかった理由は，ベースラインの予測が上手くいかなかった原因と同様に，企業名が主題となっていないニュース記事を使用し，そのニュース記事をもとに過去の類似ニュースを取得したことである．具体例として，トヨタ自動車のニュース記事の例を挙げる．トヨタ自動車のニュース記事の本文として,『【NQN ニューヨーク＝戸部実華】７日の米株式市場で日本株の米預託証券（ADR）はほぼ全面安だ ･･･』というものがある．このニュース記事は，トヨタ自動車と関係のないニュース記事であるため, トヨタ自動車が主題ではないニュース記事だと言える．このようなニュース記事から取得した，類似する過去のニュース記事の一つとして,『NQN ニューヨーク＝戸部実華】2 日の米株式市場で日本株の米預託証券（ADR）はほぼ全面安だ ･･･』というものがある．企業名が主題となっていないニュース記事をもとに，過去の類似するニュース記事を取得しても，そのニュース記事にも企業名が含まれていないため，使用するニュース記事としてはふさわしくないと考えられる．したがって，企業名が主題でないニュース記事は，事前に取り除いておく必要があると考える．

一方で，その企業のニュース記事と類似するニュース記事が過去のものから見つかると，株価投資の判断に有用ではないかと考えられる．具体例として，パナソニックのニュース記事を挙げる．パナソニックの 2018 年 10 月 30 日のニュース記事には,『パナソニックの創業 100 周年記念イベントが 30 日、東京国際フォーラムで開幕した。同社の津賀 ･･･』といったものがある．このニュース記事の超過収益率が -2.00×10^{-2} となっている．また，この記事に類似するニュース記事として，2017

年 8 月 24 日の記事が挙げられ,『パナソニックは 24 日、来年 3 月に創業 100 周年を迎えるにあたって発売する 13 種類の家電製品 ･･･』がある．この日時の超過収益率が -5.11×10^{-3} となっている．以上の例により，超過収益率が同じように負の値を示し，かつ近しい値を示していることから，上記のことが考えられる．

5 まとめと今後の課題

本研究では，ある企業に関するニュース記事が与えられた際に，その企業の翌日以降の株価変動を予測する問題に取り組んだ．まず，4.2 節では，ベースラインとして，ある企業に関するニュース記事が与えられた際に，そのニュース記事のみを用いて，翌日の超過収益率の予測をする手法を提案した．また，3.2 節では，本研究で行う提案手法の流れを説明し，4 節では，実験として，2 種類の提案手法として，与えられたニュース記事に対して，テキストとして類似する過去のニュース記事のみを用いて，翌日の超過収益率を予測する手法と，与えられたニュース記事と，テキストとして類似する過去のニュース記事を用いて，翌日の超過変動率を予測する手法を用いた．本研究の結果として，ベースラインの手法，提案手法のいずれかを行っても，翌日の超過収益率の予測を上手く実施することができなかった．本研究の結果に対する考察は，4.5 節で述べた．今後の課題としては，本研究を実施して浮かび挙がった課題を考慮し，超過収益率の予測精度を向上させる方法を引き続き考える必要があると考えている．

謝辞 本研究は JSPS 科学研究費助成事業 JP21H03774, JP21H03775 による助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, Vol. 2, No. 1, pp. 1–8, 2011.

[2] Deli Chen, Keiko Harimoto, Ruihan Bao, Qi Su, Xu Sun, et al. Group, extract and aggregate: Summarizing a large amount of finance news for forex movement prediction. *arXiv preprint arXiv:1910.05032*, 2019.

[3] Deli Chen, Yanyan Zou, Keiko Harimoto, Ruihan Bao, Xuancheng Ren, and Xu Sun. Incorporating fine-grained events in stock movement prediction. *arXiv preprint arXiv:1910.05078*, 2019.

[4] Yingmei Chen, Zhongyu Wei, and Xuanjing Huang. Incorporating corporation relationship via graph convolutional neural networks for stock price prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1655–1658, 2018.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[6] Xin Du and Kumiko Tanaka-Ishii. Stock embeddings acquired from news articles and price history, and an application to portfolio optimization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 3353–3363, 2020.

[7] Mengzhen Fan, Dawei Cheng, Fangzhou Yang, Siqiang Luo, Yifeng Luo, Weining Qian, and Aoying Zhou. Fusing global domain information and local semantic information to classify financial documents. p. 2413–2420, 2020.

[8] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pp. 261–269, 2018.

[9] Qikai Liu, Xiang Cheng, Sen Su, and Shuguang Zhu. Hierarchical complementary attention network for predicting stock price movements with news. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1603–1606, 2018.

[10] Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 8415–8426, 2020.

[11] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 24–29, 2013.

[12] Tsun-Hsien Tang, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Retrieving implicit information for stock movement prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2010–2014, 2021.

[13] Wentao Xu, Weiqing Liu, Chang Xu, Jiang Bian, Jian Yin, and Tie-Yan Liu. REST: Relational event-driven stock trend forecasting. In *Proceedings of the Web Conference 2021*, pp. 1–10, 2021.

[14] 羽室行信（編集）. データの前処理. 朝倉書店, 2021.

[15] 指田晋吾, 中川慧. 経済因果チェーンと ssestm モデルによる決算情報の伝播を活用した投資戦略. 人工知能学会全国大会論文集第 35 回全国大会, 4J2GS6e05, 2021.

[16] 前川浩基, 中原孝信, 岡田克彦, 羽室行信. 大規模ニュース記事からの極性付き評価表現の抽出と株価収益率の予測. オペレーションズ・リサーチ: 経営の科学, Vol. 58, No. 5, pp. 281–288, 2013.

[17] 中川慧, 今村光良, 吉田健一. 株価変動パターンの類似性を用いた株価予測. 人工知能学会全国大会論文集第 31 回全国大会, 2D11, 2017.

[18] 米田宏生, 湯本高行, 礒川悌次郎, 上浦尚武. ニュース記事の考慮の有無による株価指数の予測結果の差に基づく経済的影響力の推定. 情報処理学会第 83 回全国大会, 6L-05, 2021.

[19] 磯貝明文, 川口宗紀, 小林寛司. サプライヤー・カスタマーのつながりに基づくクロスモメンタムの株価予測可能性. 現代ファイナンス, Vol. 40, pp. 25–48, 2019.