BERT を用いた株価への影響力のあるニュース記事の推定

井口　勝太†　　湯本　高行†

† 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 丁目 2-1
E-mail: †{ad21r047,yumoto}@gsis.u-hyogo.ac.jp

あらまし　株価は突発的な出来事によって大きく変動することがある。このような突発的な出来事はニュース記事として報道される。そこで、本研究ではニュース記事が株価に影響を与えるかを判定する手法を提案する。この手法では、まず株価指数だけを用いた株価予測と株価指数に加えてニュース記事を用いた株価予測を行う。次に、これらの差分からニュース記事の影響力の有無を判定する。差分の精度は株価予測に依存するので、株価予測モデルに BERT を組み込むことで精度の向上を目指す。

キーワード　ニュース, 株価予測,BERT

1　は じ め に

突発的な出来事はニュース記事として報道される場合がある。例えば、コロナウィルスの拡大のニュースであれば、飲食店にはネガティブなニュースである。また、原油価格の高騰のニュースは株価の下落につながることがある。このような株価や経済に影響のあるニュース記事は投資家の株売買や企業の経営戦略に活用することができる。

しかし、芸能やスポーツ、政治といった様々なジャンルの大量のニュース記事の中から経済に影響を与えるニュース記事を見つけることには大きな労力が必要となる。また、その発見が遅くなるとリアルタイム性が損なわれ情報の価値が落ちてしまう。

そこで近年研究が盛んなビッグデータやディープラーニングを用いて株価に影響のあるニュース記事を発見する手法を提案することが本研究の目的である。本論文ではビッグデータである株価指数とニュース記事を用いたモデルを構築し、入力したニュース記事の持つ影響力を推定する手法を提案する。

本書では、2 章で先行研究について紹介するとともに手法の提案について述べる。3 章で、新手法と先行研究の比較を行い、4 章で新手法の評価を行う。最後の 5 章で研究の全体を振り返り、今後の展望を示す。

1.1　関 連 研 究

テキストデータを用いて株価の値動きを推測する研究は参考文献 [3] のようなものがある。[3] では株価分析に特化した極性辞書を作成して、それを使用してロイター通信の日本語版記事をニュース記事の極性を評価し、株価の騰落を予測した。結果として、テキスト情報から株価の推測ができる可能性を示唆している。また、Bollen1 ら [4] は twitter の投稿から、社会全体のムードを Calm, Alert, Sure, Vital, Kind, Happy の 6 つに推定して、Calm の値が 3 日遅れで株価と相関していることを示した。本研究もこういった研究の延長上に存在しており、テキストと株価の関係を分析するものである。

2　ニュース記事の株価への影響力推定

本章ではまず先行研究を紹介し、次に本研究での提案手法について述べる。

2.1　word2vec を使用した影響力推定

先行研究として米田の「ニュース記事の株価への影響力の推定」[2] がある。これはディープラーニングで 2 通りの株価予測をして、その差分を影響力とする手法である。2 通りの予測は株価指標とニュース記事から予測した株価と、株価指標のみから予測した株価である。2 つの差分はニュース記事の有無から発生するためニュース記事の影響力ととらえられる。

入力に使う株価指標は 7 つあり、次の表 1 のようになっている。これらは株の売買や分析に使用される指標であり、機械学習などの複雑な方法は用いずに計算することができる。

表 1　使用する 7 つの指標

Stochastic %K	Stochastic %D	Momentum
Rate of Change	William ' s %R	A/D Oscillator
Disparity5		

表 1 の指標に株価指数を追加した 8 つの数値を使用する。時間的な変動を考慮するため、指標を 7 日分まとめて株価指標としてのモデルへ入力する。

ニュース記事は複数入力すると、どの記事の影響力がどれだけあるのかを算出することが難しくなる。よって、予測対象とする日の前日の分のみを使用する。以上より、入出力の日付関係は次の図 1 のようになる。

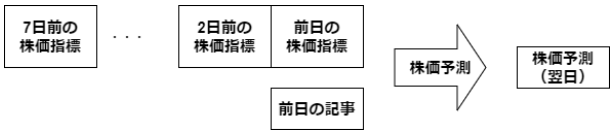

図 1　入力と出力の日付関係

図 1 のように入力は 7 日分の株価指標とニュース記事であ

り、出力は翌日の株価予測である。

ニュース記事は word2vec によって 300 次元のベクトルに変換してからモデルに入力する。この際、学習済み word2vec である朝日単語ベクトル [5] を使用している。

朝日単語ベクトルは朝日新聞の 1984 年 8 月から 2017 年 8 月までの記事のうち、約 800 万記事を学習したものである。朝日単語ベクトルはいくつか種類があるが、Glove で学習したものを Retrofitting で最適化したモデルを使用している。

個別銘柄を使用するとニュース記事が存在しない日が多くなり、モデルを学習することが難しい。そこで、予測する対象は TOPIX17 の企業分類から業界の上位 10 社の株価を標準化したのち純資産比で加重平均をとったものから計算した指数移動平均 EMA とする。金融業界の上位 10 社を次の表 2 に示す。

表 2　金融 10 社

東京海上 HD	オリックス
野村 HD	第一生命 HD
MS&AD インシュアランスグループ HD	SOMPO HD
日本取引所グループ	大和証券グループ本社
T&D HD	SBI IH

上記から計算した業界の指標を予測するためのモデルの概要は次の図 2 のようになる。

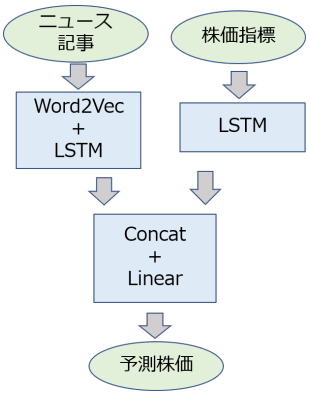

図 2　word2vec を使用したモデル

図 2 のようにニュース記事と株価指標はそれぞれ LSTM [6] によって解析されたのちに Concat 層によって結合され、最終的に Linear 層で全結合される。結果として予測株価という 1 つの値になる。

このモデルを使用して、ニュース記事の株価への影響力を推定する。影響力の計算は次の図 3 のようになる。

図 3 のように株価指標を入力とした株価予測と株価指標とニュース記事を入力とした株価予測の差分を計算する。この差分をニュース記事の株価への影響力とする。

2.2　BERT を用いた影響力推定

上で説明した手法において影響力の推定精度は株価の予測に依存すると考えられる。そこで、モデルの改善のためテキスト処理部を BERT [1] に変更し、さらに予測対象を株価の変動幅とした。本研究では、東北大の事前学習済み BERT を使用し

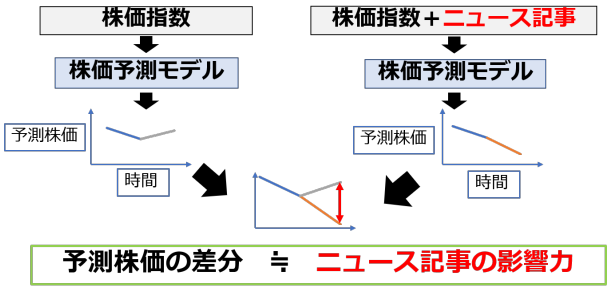

図 3　影響力の計算

た。これは日本語 Wikipedia が学習に用いられている。BERT は transformer という機構を持ち、大量のデータで学習を行うため文脈を考慮することができるとされている。

以上から、モデルは次の図 4 のようになる。

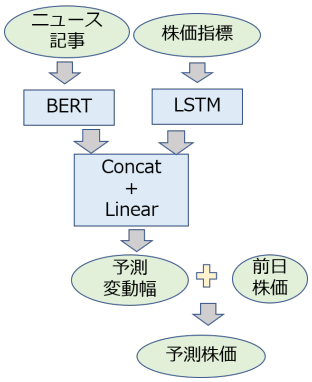

図 4　BERT を使用したモデル

入力に使う株価指標は先行研究と同様に 7 件とした。そこに始値、最高値、最安値、終値の 4 つを加えた 11 の指標を入力とする。11 の指標を先行研究と同様に 7 日分株価指標として入力する。

ニュース記事は BERT によって 768 次元のベクトルに変換される。株価指標は LSTM によって 32 次元のベクトルに変換される。これらを Concat 層で結合し、Linear 層で全結合する。結果として、変動幅の予測値を得る。変動幅の予測値に前日の株価を足すことで、翌日の株価予測とする。

本研究では BERT のファインチューニングは行わず、株価指標の処理をする LSTM と変動幅を予測する Linear を学習する。

実験は金融、電力・ガス、素材・化学という 3 つの業界に対して行う。それに際してそれぞれの業界に対する株価予測モデルを作成した。

モデルの学習・検証、テストに使用するデータは 2014,2016,2017 年の平日 733 日分を使用している。また、テストデータは 2018 年 1 月 19 日から同年 3 月までの 42 日分のデータを使用している。ニュース記事は、業界に関係ないものも多い。業界に関係が大きいと考えられる記事を抽出するため、表 2〜4 の業界上位の企業の名前が入っている記事を抽出して使用する。

こうして抽出した記事は存在する日と存在しない日がある。抽出済み記事が存在する日の内訳は次の表 3, 表 4, 表 5 のよう

になる。

表 3　ニュース記事ありの日の内訳（金融）

	データ数	抽出済み記事有	割合
学習・検証データ	733	130	17.3%
テストデータ	42	13	31.0%

表 4　ニュース記事ありの日の内訳（電力・ガス）

	データ数	抽出済み記事有	割合
学習・検証データ	733	163	22.2%
テストデータ	42	17	40.4%

表 5　ニュース記事ありの日の内訳（素材・化学）

	データ数	抽出済み記事有	割合
学習・検証データ	733	159	21.7%
テストデータ	42	15	35.7%

表 3, 表 4, 表 5 より、学習データ中のニュース記事が存在する日はおよそ 20% となっている。テストデータ中のニュース記事が存在する日はおよそ 30 から 40%となっている。

表 3 の金融業界では学習・検証データは記事が他業界より 30 件程度少なくなっている。

モデルの学習に際しては、学習・検証データの 733 件のうち 8 割に当たる 581 件を学習に使用する。733 件のデータのうち、2 割に当たる 152 件を検証に使用する。また、学習率は 0.00001 で、エポック数は 100、オプティマイザーは Adam [7] を用いる。

こうして作成したモデルにニュース記事と株価指標を入力した場合と株価指標のみ入力した場合の予測株価の差分をとりニュース記事の影響力を推定する。

3　株価予測モデルの評価

株価予測モデルの評価について示す。

3.1　実験方法

先行研究では、どのようなニュース記事が経済に影響するのかを分析することを目的とするため、業界ごとの分析を行う。本書では先行研究にならって金融業界、電力・ガス業界、素材・化学業界のニュース記事を分析する。この際、業界の上位 10 社の社名を含むこの際使用するニュース記事を使用する。金融業界上位 10 社は表 2 のものを使用する。

電力・ガス業界の上位 10 社は次の表 6 のようになる。

表 6　金融 10 社

東京ガス	大阪ガス
中部電力	関西電力
東邦ガス	東京電力 HD
東北電力	九州電力
中国電力	電源開発

素材・化学業界の上位 10 社は次の表 7 のようになる。

表 7　素材・化学 10 社

信越化学工業	花王
資生堂	富士フィルム HD
ユニ・チャーム	日本ペイント HD
旭化成	日東電工
東レ	SBI HD

本研究で作成したモデル（以下新モデル）と先行研究のモデル（以下旧モデル）に、2018 年の 1 月 19 日から 3 月までの平日 48 日分の株価指標とそれに対応するニュース記事を入力して株価予測を行う。ニュース記事は金融業界が 13 件、電力・ガス業界が 17 件、素材・化学業界が 15 件となっている。翌日の予測に株価指標を 7 日分使用するので予測は 42 件得られる。

以上から、2 つのモデルの予測と実際の値からモデルごとに決定係数と RMSE を計算し比較する。決定係数 R^2 は予測値の実際の値への当てはまり具合を示す指標であり、次の式 (1) のようになる。

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \bar{y}_i)^2} \tag{1}$$

ただし、y は実際の値、$\hat{y}$ は予測値、$\bar{y}$ は実際の値の平均値、n はデータの件数とする。この決定係数が高いほどモデルの予測精度は高くなる。

RMSE は予測値と実際の値の差を示す指標であり、次の式 (2) のようになる。

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(\hat{y}_i - y_i)^2} \tag{2}$$

この RMSE は低いほどモデルの予測精度は高くなる。決定係数と RMSE によるモデルの評価を次の節で行う。

3.2　実験結果

金融業界に加えて素材・化学業界、電力・ガス業界についても同様の実験を行った。以下では、その結果を示す。

3.2.1　金融業界の実験結果

実験の結果、旧モデルの出力は次の図 5 のようになった。

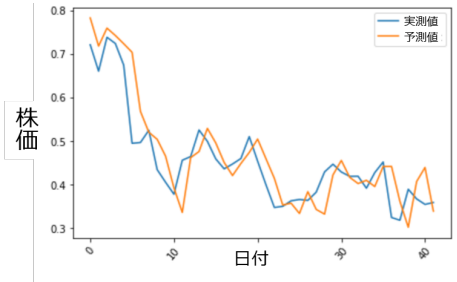

図 5　旧モデルによる予測結果

また、新モデルの出力は次の図 6 のようになった。

図 5 と図 6 を比較すると、新モデルのほうが実際の株価に近い予測をしているように見える。

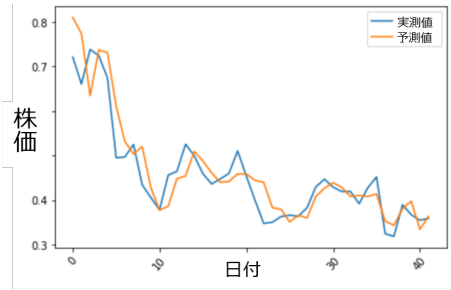

図 6　新モデルによる予測結果

表 8　株価予測モデルの評価

	決定係数（R^2）	RMSE
旧モデル	0.687	0.0592
新モデル	0.806	0.0466

2 つのモデルの決定係数と RMSE は次の表 8 ようになる。

結果、新モデルは RMSE は下がり決定係数は上昇した。このことから、新モデルは旧モデルより株価予測をうまく行うことができるようになった。次に、ニュース記事がある場合とない場合の予測性能を表 9 に示す。

表 9　ニュース記事ありなしの予測性能比較（金融）

	決定係数 (R^2)	RMSE
ニュースあり	0.806	0.0466
ニュースなし	0.775	0.0502

表 9 より、ニュース記事がある場合のほうが決定係数が高く、予測性能が高いことが確認できた。

3.2.2　電力・ガス業界の実験結果

電力・ガス業界における新旧モデルの決定係数と RMSE は次の表 10 ようになる。

表 10　株価予測モデルの評価

	決定係数（R^2）	RMSE
旧モデル	0.278	0.0624
新モデル	0.343	0.0592

電力・ガス業界においては新モデルが旧モデルよりよい結果を出している。次にニュース記事がある場合とない場合の予測性能を表 11 に示す。

表 11　ニュース記事ありなしの予測性能比較（電力・ガス）

	決定係数 (R^2)	RMSE
ニュースあり	0.324	0.0604
ニュースなし	0.480	0.0529

表 11 より、電力・ガス業界ではニュース記事がない場合のほうが予測性能が高いことが分かる。

表 12　株価予測モデルの評価

	決定係数（R^2）	RMSE
旧モデル	-0.084	0.0520
新モデル	0.0530	0.0486

3.2.3　素材・化学業界の実験結果

素材・化学業界における新旧モデルの決定係数と RMSE は次の表 12 ようになる。

素材・化学業界においては旧モデルの決定係数がマイナスであるため、変動の逆の予測をしていると思われる。新モデルでは決定係数が正となった。有意とは言い難いが、わずかに性能が向上したと考えられる。次にニュース記事がある場合とない場合の予測性能を表 13 に示す。

表 13　ニュース記事ありなしの予測性能比較（素材・化学）

	決定係数 (R^2)	RMSE
ニュースあり	0.053	0.0486
ニュースなし	0.119	0.0469

表 13 よりニュース記事なしの場合のほうが予測性能が高いことが分かる。

3.3　考　　察

新モデルでは、テキスト処理部を BERT に変更したことでテキストの分析力が向上し、予測対象を株価変動幅にしたことで予測タスク自体が易化したため、精度が向上したと考えられる。

また、金融業界の株価予測は決定係数が約 0.8 だったのに対して、電力・ガス業界は決定係数が約 0.3、素材・化学業界は決定係数が約 0.05 となった。このため、金融業界が特別に株価予測しやすい業界であるといえる。

ニュース記事を使用する場合と使用しない場合の比較では、金融業界ではニュース記事有のほうが予測性能が高かった。これは金融業界はニュース記事の影響を受けやすい業界であるためだと考えられる。電力・ガスではニュース記事なしの場合のほうが性能が高かった。これは電力・ガス業界はインフラであり安定しており、ニュース記事の影響を受けにくいためだと考えられる。素材・化学業界でもニュース記事なしの場合のほうが性能が高かった。これは、素材・化学業界のニュース記事は専門用語などが多いため言語モデルでの分析が難しいためだと考えられる。

4　ニュース記事の影響力の分析

4.1　実 験 方 法

どのようなニュース記事が経済に影響するのかを調べるために、提案手法で影響力を計算し、その値をもとにニュース記事を分析する。ここでは、ニュース記事を影響力と照らし合わせて 3 種類に分類する。3 種類は影響力の算出および影響力のあるニュース記事、ないニュース記事、適切に影響力推定ができていないニュース記事とした。

4.2 実験結果

金融、電力・ガス、素材・化学の 3 つの業界に対して実験を行い、その結果を示す。

4.2.1 金融業界の実験結果

提案手法で予測した影響力は 42 件あり、そのうちニュース記事がある日は 13 件である。上記の方法で算出したニュース記事の影響力を図 7 に示す。

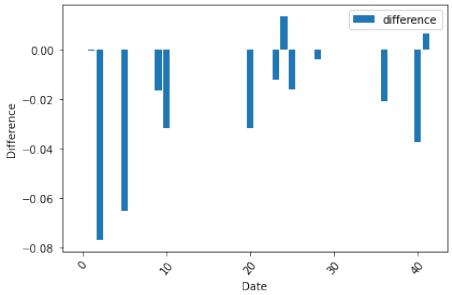

図 7　ニュース記事の影響力

図 7 から影響力が負の場合が多く、正の場合が少ないことが見受けられる。以下ではこの影響力を用いて評価したニュース記事を示していく。

影響力が正の方向に高かったもので、適切でありそうなものを表 14 に示す。簡単のため、以降の表ではニュース記事の冒頭部分のみを示す。

表 14　正方向の影響力があるニュース

日付	ニュース記事	影響力
2018 年 3 月 5 日	次世代金融サービスの創出に向けた新会社「Fintertech 株式会社」の設立について　…	0.01351
2018 年 3 月 29 日	第一生命グループ 2018-20 年度中期経営計画　第一生命ホールディングス株式会社…	0.00632

次世代金融サービスの創出に向けた新会社「Fintertech 株式会社」の設立については、大和証券グループのデジタル化推進を前面に押し出した記事であるため正の方向に影響力があると考えられる。

第一生命グループ 2018-20 年度中期経営計画は影響力は小さいが、正方向のニュースが少ないため例として示す。これは将来を見据えた前向きなメッセージのある記事であるため、わずかに正方向の影響力があると考えられる。

表 15 に影響力が負の方向に大きかったものを示す。

表 15　負方向の影響力があるニュース

日付	ニュース記事	影響力
2018 年 2 月 5 日	中国の平安保険グループ子会社の FinTech ベンチャー企業への出資に関するお知らせ　SBI ホールディングス株式会社…	-0.0652
2018 年 3 月 22 日	Oman ORIX Leasing Company SAOG の株式譲渡に関するお知らせ	-0.0209

中国の平安保険グループ子会社の FinTech ベンチャー企業への出資に関するお知らせでは一見ポジティブなニュースのように思われたが、2 月 5 日から 7 日にかけて SBI ホールディングスの株価が低下しているため、負の方向に影響力があると考えられる。

また、Oman ORIX Leasing Company SAOG の株式譲渡に関するお知らせは、オリックス社がオマーン国現地法人を譲渡するという内容で、このような株式譲渡は株価に対してネガティブなニュースとなりやすい。

ここで、株価への影響力の絶対値が小さいニュース記事を表 16 に示す。

表 16　影響力の絶対値が小さいニュース記事

日付	ニュース記事	影響力
2018 年 1 月 30 日	SBI 地域銀行価値創造ファンドの新設と、それに伴う特定子会社の異動に関するお知らせ	-0.00051

この記事は影響力が極めて小さく、株価への影響力がほとんどないと考えられる。

最後に影響力の推定が適切でないものを表 17 に示す。

表 17　影響力の推定が適切でないもの

日付	ニュース記事	影響力
2018 年 3 月 28 日	野村ホールディングス株式会社と LINE 株式会社による 金融事業における業務提携検討開始に関する基本合意書締結について	-0.0372
2018 年 2 月 27 日	量子コンピュータを活用した実証実験の開始について	-0.0317

野村ホールディングスと LINE の提携は野村ホールディングスの 2018 年 3 月 28 日から 2 週間程度の株の値動きを調べると、緩やかな上昇であったため、正方向への影響がある記事であると考えられるが、負方向の影響力が推定されている。

また、量子コンピューターを活用した実証実験については、即座に会社に利益を与えるものではないため、影響は小さいものであると考えられるが、負方向の影響力が推定されている。

これらは、株価への影響力が小さいものである考えられるが、モデルは過大評価しやすいと考えられる。

4.2.2 電力・ガス業界の実験結果

提案手法で予測した影響力は 42 件あり、そのうちニュース記事がある日は 14 件である。提案手法で算出したニュース記事の影響力を図 8 に示す。

図 8 から影響力は正負両方に満遍なく表れていることが分かる。以下ではこの影響力を用いて評価したニュース記事を示していく。

影響力が正の方向に高かったもので、適切でありそうなものを表 18 に示す。

表 18 のように、発電所の建設、運転開始、設備更新のための手続きなどの記事を正の影響力ありと評価することが多かった。

表 19 に影響力が負の方向に大きかったものを示す。

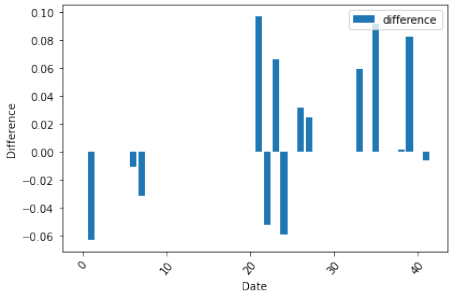

図 8　電力・ガス業界におけるニュース記事の影響力

表 18　正方向の影響力があるニュース

日付	ニュース記事	影響力
2018 年 2 月 28 日	「鬼首地熱発電所設備更新計画 環境影響評価準備書」の届出・送付および縦覧について	0.0969
2018 年 3 月 20 日	「新奥泉水力発電所」営業運転開始について	0.0912

表 19　負方向の影響力があるニュース

日付	ニュース記事	影響力
2018 年 1 月 30 日	経営機構の見直しによるコーポレート・ガバナンスの強化について～「監査等委員会設置会社への移行」および「役付執行役員の新設」～	-0.0635
2018 年 3 月 1 日	ブロックチェーンを使った電気自動車等の充電に係る新サービスの実証実験の実施について	-0.0526

表 19 の 1 つ目の記事は東北電力が事業環境の変化に対応するため、カンパニー制を導入するという内容。2 つ目は中部電力と関連企業が充電履歴を管理する技術の実証実験を行うというもの。

負の方向への影響力があるニュース記事の内容がばらけていたため、共通点を見つけることができなかった。

ここで、株価への影響力の絶対値が小さいニュース記事を表 20 に示す。

表 20　影響力の絶対値が小さいニュース記事

日付	ニュース記事	影響力
2018 年 3 月 26 日	「関西電力グループ中期経営計画達成に向けた重点取組み（2018）」の策定について	0.0014
2018 年 3 月 29 日	デジタルグリッドへの出資について	-0.0059

表 20 の 1 つ目の記事のように、経営計画に関するわずかに正の影響が予測されることが多い。

2 つ目の記事は東京電力によるデジタルグリッドというスタートアップ企業への出資に関するもので、電力・ガス業界においては出資に関する記事は影響力が小さく評価されやすい。

最後に影響力の推定が適切でないものを表 21 に示す。

表 21 の記事は大阪ガスが住ミカタ・プラスというサービスの対象者を広げ、さらに床下、照明、水回りの点検を追加するというもの。この記事が株価に対して負の影響力を持つとは考え難い。

表 21　影響力の推定が適切でないもの

日付	ニュース記事	影響力
2018 年 2 月 7 日	住まいのお困りごとを解決し安心をお届けする「住ミカタ・プラス」のサービスメニュー拡充	-0.0316

4.2.3　素材・化学業界の実験結果

提案手法で予測した影響力は 42 件あり、そのうちニュース記事がある日は 14 件である。提案手法で算出したニュース記事の影響力を図 9 に示す。

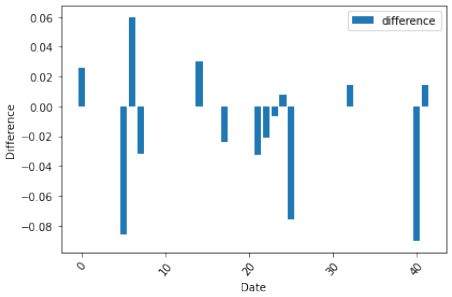

図 9　素材・化学業界におけるニュース記事の影響力

図 9 から影響力は正負両方に満遍なく表れていることが分かる。以下ではこの影響力を用いて評価したニュース記事を示していく。

影響力が正の方向に高かったもので、適切でありそうなものを表 22 に示す。

表 22　正方向の影響力があるニュース

日付	ニュース記事	影響力
2018 年 1 月 29 日	日本化成株式会社の吸収合併について三菱ケミカル株式会社	0.0261
2018 年 3 月 15 日	TenCate Advanced Composites Holding B.V. の株式の取得に関する合意書締結のお知らせ	0.0145

素材・化学業界では表 22 の 1 つ目の記事は、三菱ケミカルによる日本化成株式会社の吸収合併のニュースである。このように、会社の M&A に関するニュースを正の影響力を持つと推定することが多く見られた。

表 23 に影響力が負の方向に大きかったものを示す。

表 23　負方向の影響力があるニュース

日付	ニュース記事	影響力
2018 年 2 月 7 日	食品包装用ラップフィルムの価格改定について	-0.0315
2018 年 3 月 1 日	カラー用カーボンブラックの価格改定について	-0.0211

表 23 の負の影響力を予測された記事として、食品包装フィルム、カラー用カーボンブラックの値上げのような商品の値上げに関するニュースが多く見られた。

次に株価への影響力の絶対値が小さいニュース記事を表 24 に示す。

表 24　影響力の絶対値が小さいニュース記事

日付	ニュース記事	影響力
2018 年 3 月 2 日	3D プリンター用フィラメントメーカー蘭 Dutch Filaments 社の買収について	-0.0062
2018 年 3 月 5 日	資生堂、「新 3 カ年計画」（2018 年～2020 年）を策定	0.0078

24 の 3D プリンター用フィラメントメーカーは、海外企業の買収であり日本との関係が薄いこと、3D プリンタまだ出始めな技術であることから影響力が影響力が小さいと評価されたものだと考えられる。

資生堂、「新 3 カ年計画」（2018 年～2020 年）を策定のような事業計画に関するニュースは、わずかにプラスに評価されることが多い。

最後に影響力の推定が適切でないものを表 25 に示す。

表 25　影響力の推定が適切でないもの

日付	ニュース記事	影響力
2018 年 3 月 28 日	CFRP の新規真空圧成形技術の開発－省エネでの高精度部材成形を実現－	-0.0904
2018 年 3 月 6 日	微量のにおい成分を可視化する“Scent-EYE（セントアイ）”技術（※）を開発	-0.0757
2018 年 2 月 6 日	膜分離活性汚泥法（MBR）の省エネ基礎技術を開発― 散気エネルギーを高効率に洗浄エネルギーへ変換 ―	0.0599

表 25 のように技術開発に関する記事は影響力が大きく評価されることが多く、過大評価されていると思われる。

こういった技術開発の記事は専門性が高く、言語モデルによる分析が難しい可能性がある。

4.3 考　　察

予測モデルは、株価指標とニュース記事を入力として株価を予測する。このため、ニュース記事の影響力の推定は株の値動きを考慮したものとなっている。

現在の手法では複数の企業の株価を合成して使用しているため、記事と株価の関係を直接的に扱うことができていない。個別銘柄の分析も試したいが、個別銘柄を分析するためにはニュース記事がない日が多く、モデルの学習が困難である。よって、個別銘柄を扱う場合はニュース記事以外のテキストデータを使用する必要がある。

金融業界においてはテストデータが下降局面であったため負の影響力が多く現れ、正の影響力の考察が難しい状態だった。

将来性の感じられるニュース記事は正方向、株式譲渡や投資リスクを感じさせるニュース記事は負の方向に影響力があると推定された。

また、ニュース記事は出資や会社の設立などの記事が多く、出資先の企業がどのような企業であるかを考慮することが必要になると考えられる。

電力・ガス業界においては発電所関連の記事が正の影響力と推定されやすい。対して、負の影響力が推定されるものは共通点らしきものが見られなかった。

素材・化学業界では、研究開発に関する記事において影響力の誤推定と思われる事象が見られた。これは、専門性が高いため言語モデルによる分析が難しいと考えられる。

3 業界ともに、経営計画の記事は影響力が小さく推定された。

5 ま と め

本研究ではニュース記事と株価指標を入力として予測した株価と株価のみを入力して予測した株価の差分をとることで、ニュース記事の株価への影響力を算出する。この手法に対して BERT を組み込み、予測対象を前日からの株価指標の変動幅とすることで株価予測の性能を向上させた。

さらに、BERT を組み込みこんだ株価予測モデルを用いてニュース記事の影響力を推定した。金融、素材・化学、電力・ガスの 3 つの業界について影響力を推定した。その結果、金融業界、素材・化学業界では株価の技術の研究・開発の記事の予測が難しいことや、電力・ガス業界では発電所に関する記事が正の影響力となりやすいことなど、業界ごとの特徴が分かった。

今後は影響力の誤評価の解決、さらなる性能の向上のために BERT 部分のファインチューニングや個別銘柄の分析、影響力とニュース記事の関係の定量的な分析、手法の改善をしていく必要がある。

また、影響力の妥当性の検証のために影響力とその同時期の個別銘柄の値動きの比較を全てのテストデータに対して行い集計する必要がある。

謝　　辞

本研究は JSPS 科研費 JP19H04116 の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] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. ,arXiv:1810.04805, 2018.

[2] 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

[3] 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学会, 2013 年 5 月号

[4] Johan Bollen1, Huina Mao1, Xiao-Jun Zeng , Twitter mood predicts the stock market. , arXiv:1010.3003

[5] 田口雄哉, 田森秀明, 人見雄太, 西鳥羽二郎, 菊田洸, 同義語を考慮した日本語単語分散表現の学習, 情報処理学会第 233 回自然言語処理研究会, Vol.2017-NL-233, No.17, pp.1-5, 2017.

[6] Sepp Hochreiter, Jurgen Schmidhuber. Long Short-term Memory.Neural Computation 9(8), pp.1735-1780, 1997.

[7] Diederik P. Kingma, Jimmy Ba, Adam: A Method for Stochastic Optimization , arXiv:1412.6980