慶應義塾大学学術情報リポジトリ
Keio Associated Repository of Academic resouces

Title	ニューラルネットワークを用いたテキストデータによる株価予測の試み
Sub Title	Stock price prediction from text data using neural network
Author	前多, 康男(Maeda, Yasuo)
Publisher	慶應義塾経済学会
Publication year	2023
Jtitle	三田学会雑誌 (Mita journal of economics). Vol.115, No.4 (2023. 1) ,p.403 (83)- 411 (91)
JaLC DOI	10.14991/001.20230101-0083
Abstract	
Notes	研究ノート
Genre	Journal Article
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AN00234610-20230101-0083

慶應義塾大学学術情報リポジトリ(KOARA)に掲載されているコンテンツの著作権は、それぞれの著作者、学会または出版社/発行者に帰属し、その権利は著作権法によって保護されています。引用にあたっては、著作権法を遵守してご利用ください。

The copyrights of content available on the KeiO Associated Repository of Academic resources (KOARA) belong to the respective authors, academic societies, or publishers/issuers, and these rights are protected by the Japanese Copyright Act. When quoting the content, please follow the Japanese copyright act.

「三田学会雑誌」115 巻 4 号（2023 年 1 月）

研究ノート

ニューラルネットワークを用いたテキストデータによる株価予測の試み

前多康男 *

1　はじめに

テキストデータを用いたデータ分析は様々な分野で活用されているが，ファイナンスの分野でも金融資産の価格や収益率の予測に使用する試みが行われてきている。本研究ノートでは，テキストデータを用いて金融市場の分析を行っている研究の展望をまず行い，その後で，日本企業について有価証券報告書のテキストデータを用いてニューラルネットワークにより株価収益率の予測を行う。

2　テキストデータを活用した金融市場の予測に関する既存研究の展望

■インターネット掲示板のテキストデータを用いた分析　Antweiler and Frank（2004）は，Yahoo Finance と Raging Bull に投稿された 150 万件以上のメッセージを収集し，これらのメッセージが株式市場に与える影響について分析を行っている。これらのメッセージは市場のボラティリティを予測することには有益であったが，株価の収益率への影響は統計的に有意であるもののそれほど大きいものではないと結論づけている。丸山他（2007）は，Antweiler and Frank（2004）の分析を基にして Yahoo 株式掲示板へのメッセージと株価との関連性を調べている。具体的にはソフトバンクとソニーを事例として，投稿数と株式市場との関係を分析している。投稿を場前，場中，場後に分割して投稿数を数え，メッセージ投稿数と株式リターン，ボラティリティ，出来高などとの相関を計算している。これらに関して有意な相関が観測されている。

Schumaker and Chen（2009）では，Yahoo Finance からニューステキストを収集し，それらのテキストに対してテキストマイニングを行い，金融市場の分析を行っている。具体的

＊　慶應義塾大学経済学部

には収集したテキストデータを Bag of Words などの形にし，ニュースのテキストデータと S&P500 に採用されている企業のニュース発表の 20 分後の株価との関連性をサポートベクターマシンを用いて分析している。Schumaker and Chen (2009)では 3 つのモデルを比較している。モデル 1 では予測にテキスト情報のみを用いており，モデル 2 ではテキスト情報と共にテキストが発表された時点の株価を用いている。モデル 3 ではテキスト情報と共に回帰モデルによって予測された 20 分後の株価を用いている。モデル 2 の予測が良好であったことが報告されている。

坪内他(2016)は，Yahoo 株価掲示板のテキスト情報を用いて株価を予測している。まず，金融実務家が株価予想を行うときに重要であると考えた 1,319 単語に対して，金融実務家自身が自分の判断でセンチメントをスコア化し，その平均値を各語のスコアとする極性辞書を作成している。分析に際しては，掲示板に登場する単語のうち極性辞書に含まれる単語を分析対象としている。極性辞書には 1,319 単語が掲載されているが，最終的な分析には合計で 103 単語が使用されている。103 語を word2Vec で 200 次元のベクトルに変換し cos 類似度を計算した後にクラスタリングを行い，103 語を 25 のグループに分けている。このクラスタリングの結果を用いて，1 から 25 までのグループ番号でテキストの特徴量を表している[(1)]。この 25 次元の特徴量ベクトルを説明変数として，その記事に対応する銘柄の株価の変動率（リターン）[(2)]を被説明変数として外挿予測（2 値分類）を行っている。モデルとしてはロジスティック回帰（LR），サポートベクターマシン（SVM），ランダムフォレスト（RF）の 3 つの方法を用いている。適合率，再現率，F 値を用いて結果の比較を行っており，LR，SVM，RF の順で結果が良くなっていることが報告されている。しかし，全体的にチャンスレベルの予測結果となってしまっており，この原因として最終的に使用された単語が 103 語に留まってしまったことを挙げている。

易・杉本(2017)は，日本語版 Yahoo ファイナンスの Web サイトから日経 225 に採用されている企業に関連するニュースを収集し，ニュースの話題や内容が株価動向へ与える影響について分析を行っている。株価データは，ニュースの発表日において始値＜終値であれば up，始値＞終値であれば down とラベル付けする[(3)]。分析の手始めに，日本語 Wikipedia と毎日新聞の本文データを用いて 200 次元の CBOW モデル（word2Vec）を学習している。ニュースのベクトル化は，ニュースから抽出した単語に対してまずカイ二乗値を用いた選択を行い[(4)]，選ばれた単語に対して word2Vec

(1)　具体的には，各記事テキストに対象となる 103 語が含まれていたら，その単語のグループ番号の特徴量に極性辞書の値を加算している。このような処理で各記事が 25 の特徴量で表現されることになる。
(2)　当該記事が掲載されてから 1，5，10 分後，引けまでの変動率を用いて分析を行っている。
(3)　始値＝終値のニュースは対象外としている。
(4)　カイ二乗値を用いた選択は Hagenau et al. (2012)により提唱された。

によりベクトル化し，それらの平均値をもってニュースのベクトル表現としている。このベクトル値を入力として，サポートベクターマシンにより up と down の 2 値分類を行っている。易・杉本(2017)ではトピックモデル[(5)]を用いてニュースのクラスタリングを行い，クラスタごとに適切なモデルを選択している。各クラスタに対して適切なモデルを構築することでより良い精度が得られること，カイ二乗値による単語選択と単語分散表現の利用により予測精度がやや向上することを結論としている。

■ツイッターのテキストデータを用いた分析

Zhang et al.（2011）は，ツイッターのテキストデータを用いて社会（大衆）の雰囲気（public mood）を分析し，ダウ・ジョーンズ，S&P500，NASDAQ などの株価指数の予測を行っている。まず fear，worry，hope などの感情を表す単語を頼りにツイッターの各テキストデータに感情をタグ付けし，これらのタグを集計することで社会の集団的な感情を計測する。そして各感情を表す単語を含むツイートの数を単純に数えて日次の指標としている。分析の結果，fear や worry を含んでいるテキストに関する指標と株価指数との間の有意な負の相関が観測されている。一方で hope を含んでいるテキストに関する指標と株価指数との間にも有意な負の相関があることが観察されており，この観察事実に対してはより一層の分析が必要であることを指摘している。

Bollen et al.（2011）は，ツイッターのテキストデータに感情分析を行った結果とダウ・ジョーンズ工業株価平均（Dow Jones Industrial Average, DJIA）に相関があるかどうかを調べている。Bollen et al.（2011）では，社会の雰囲気を測るために 2 つのツールを用いている。1 つは OpinionFinder[(6)]でツイートのテキストデータからポジティブ・ネガティブの 2 値の日次データを生成している。もう 1 つは GPOMS（Google-Profile of Mood States）でツイートのテキストデータから日次の 6 次元の感情データ（calm, alert, sure, vital, kind, happy）を生成している。自己組織化ファジーニューラルネットワーク（Self-Organizing Fuzzy Neural Network）を用いた DJIA の変化の予測に際して，感情分析の結果を追加することによって結果が向上することを報告している。

■テキストデータと VAR モデルによる分析

Tetlock（2007）は，ウォール・ストリート・ジャーナル（Wall Street Journal）のコラムの文章を用いて，株式市場とメディアの相互の影響の与え方について分析を行っている。分析の結果，悲観的なニュースが市場価格に対して最初に下向きの力を働かせ，その後ファン

(5)　吉田他(2011)でも，記事の話題に着目して，株式の取引高の増加・減少を予測する分析を行っている。実際には記事のタイトルを取引高を上昇させる確率の高い順番に並べる試みを行っている。その際にトピックモデルを用いている。トピックモデルとは単語の類似性をモデル化したもので，一般的には Latent Dirichlet Allocation（LDA）が用いられる。

(6)　OpinionFinder の詳細については Wilson et al.（2005）を参照すると良い。

ダメンタル水準まで戻す（reversion）こと，異常に高いまたは異常に低い悲観的雰囲気は高い取引量を引き起こすことを見出した。

テキストの分析には，General Inquirer（GI）[7]を使用している。具体的には，ハーバード社会心理辞書（Harvard psychosocial dictionary）の77のGIカテゴリーに着目し，その登場頻度の日次データを生成する。そして，その77カテゴリーを主成分分析によって1つのメディア要因に圧縮する。このメディア要因は悲観的な単語と強く関連付けられていたので悲観要因（pessimistic factor）と呼んでいる。

Tetlock（2007）における具体的な分析手法としては，ダウ・ジョーンズ工業株価平均の日次データを用いて，VARモデルの推定を行っている。推計式は，

$$Dow_t = \alpha_1 + \sum_{j=1}^{5} \beta_{1j} Dow_{t-j} + \sum_{j=1}^{5} \gamma_{1j} BdNws_{t-j} + \sum_{j=1}^{5} \delta_{1j} Vlm_{t-j}$$

で，Dow_tは時点tのダウ・ジョーンズ工業株価平均の収益率，$BdNws_t$は時点tの悲観要因，Vlm_tは時点tのニューヨーク証券取引所の取引高である[8]。推計の結果，$\gamma_{11} = -8.1, \gamma_{12} = 0.4, \gamma_{13} = 0.5, \gamma_{14} = 4.7, \gamma_{15} = 1.2$を得ている。悲観要因の1標準偏差分の変化は，次の日のダウ・ジョーンズ指数の収益率を8.1ベーシスポイント低下させるが，5日後までにはそのうちの6.8ベーシスポイントが戻ってしまうことになる。したがって，悲観的なニュースの影響は一時的なものであると考えられる。

また，悲観要因がコラムの内容を合理的に表しているとすれば，過去の経済変数が悲観指数に影響を与えているはずである。このことを確認するために，Tetlock（2007）では，

$$BdNws_t = \alpha_2 + \sum_{j=1}^{5} \beta_{2j} Dow_{t-j} + \sum_{j=1}^{5} \gamma_{2j} BdNws_{t-j} + \sum_{j=1}^{5} \delta_{2j} Vlm_{t-j}$$

の形のVARモデルを推計し，ダウ・ジョーンズ工業株価平均の収益率が悲観要因に与える影響を調べている。推計の結果，$\beta_{21} = -5.8, \beta_{22} = 2.3, \beta_{23} = 2.1, \beta_{24} = 2.3, \beta_{25} = 4.2$を得た。ダウ・ジョーンズ指数の収益率が前日に対して1%下落すると悲観指数は1標準偏差の5.8%上昇していることがわかる。

沖本・平澤（2014）は，QUICK端末で配信されている日経ニュースからニュース指標を作成し，その指標が株価に対して本源的な情報を有しているかどうかを検証している。金融工学研究所がニュースに対して付与したタグ情報[9]を用いて，各ニュースにポジティブ・ネガティブ（ポジ・ネガ）のラベルを付けてい

（7） General Inquirerについては，ハーバード大学のサイト（http://www.wjh.harvard.edu/ inquirer/）を参照すると良い。

（8） 実際の推計にはいくつかの外生変数を追加している。

る。ニュースを日次で取得し，ポジティブな記事の数を P，ネガティブな記事の数を N として，

$$NI1 = \frac{P-N}{P+N+1}, \quad NI2 = \frac{-N}{P+N+1}$$

とニュース指標を作成する。Tetlock (2007) に倣って，ニュースが株価に対して本源的な影響を与える理論を情報理論，ニュースが株価に対して本源的な情報を保有しないが，マーケットのセンチメントには影響を与える理論をセンチメント理論，ニュースが株価やセンチメントに全く影響を与えない理論を無情報理論と呼び，それぞれの検証を行っている。情報理論のもとではニュースは株価に対して有意かつ恒久的な影響を与える。センチメント理論のもとではニュースが株価に対して短期的には有意な影響を与えるが，長期的にはその効果は消えることになる。無情報理論のもとでは，ニュースは株価に対して短期的にも長期的にも有意な影響を与えない。Tetlock (2007) と同様に，Tpx を TOPIX の日次対数収益率，NI をニュース指標，Vol を東証一部の日次出来高の対数値として，

$$Tpx_t = \alpha_1 + \sum_{j=1}^{5} \beta_{1j} Tpx_{t-j} + \sum_{j=1}^{5} \gamma_{1j} NI_{t-j} + \sum_{j=1}^{5} \delta_{1j} Vol_{t-j} \qquad (1)$$

の形の 3 変量 VAR モデルを推計している。γ_{11} が有意に正になり，γ_{12} から γ_{15} は有意とはなっていない。ここから，ニュース指標が株価に対して本源的な情報を有している可能性が示唆されている。Tetlock (2007) の結果とは異なりリバウンドは観察されていない。

五島・高橋 (2016) は，ロイターニュースを指標化し，株式価格との関連性を分析している。回帰型ニューラルネットワークとナイーブベイズ分類器によってニュース記事の文に対してポジ・ネガ分類を行っている。沖本・平澤 (2014) の (1) 式と同様のモデルを推計し，(1) 式の γ_{1j} の符号を見て株式市場に対するニュースの影響を調べている。分析の結果，ディープラーニングを用いて作成したニュース指標において，γ_{11} が有意に 0.070 となり，翌営業日の株式リターンにプラスの影響を与えていることが観測され，γ_{14} は有意に -0.062 となることから，ラグ 4 営業日で株価がリバウンドしていることなどが観測されている。この結果は，Tetlock (2007) と同様であるが，沖本・平澤 (2014) とは異なっている。また，小型株に対するニュース指標の影響が相対的に大きく，長続きしていることも発見している。

3　有価証券報告書のテキストデータを用いた分析

■有価証券報告書を用いた既存文献　有価証券報告書を用いてテキストマイニングを行っている既存文献としては以下のようなものがある。Li (2008) は企業のアニュアルレポートの可読性を言語学で用いられている指標であ

(9)　タグ情報には，対象ニュースの主要企業名，事由付きポジティブ・ネガティブ情報が入っている。金融工学研究所は，ニュースの見出しと本文のテキスト情報から，モデルによってポジティブ・ネガティブ情報の推計を行っている。推計精度は 99%超とのことである。

る Fog Index を用いて表し，可読性が低い企業ほど企業業績が悪いことを明らかにしたが，廣瀬他(2017)は，同様の手法を有価証券報告書のデータに適用し分析を行った。吉田(2018)は，有価証券報告書の「事業等のリスク」に記述されているリスク情報を分析し，企業は直面しているリスクが大きくなるほどリスク情報の記述量も多くなること，収益性の高い企業は多くのリスク情報を開示するほど高い収益性が継続し，反対に収益性の低い企業は多くのリスク情報を開示するほど将来の収益性は小さくなることなどを明らかにしている。加藤・五島(2020)は有価証券報告書の「企業の経営方針・経営戦略や経営者による経営成績の分析」に含まれるテキスト情報を用いて，経営者によって開示された将来見通しが，将来の企業業績に対する予測力を有することを示している。

■分析方法　本稿では，テキストデータとして日本企業の有価証券報告書のデータを用いて，有価証券報告書の公表後の株価が有価証券報告書の「経営方針」の部分のテキストから影響を受けるかどうかを検証することにする。[10] 有価証券報告書の提出年月の次の月から 12 か月分の株価の月次データを平均してその年のデータとし，その年の株価と前年の株価から各年の株価収益率を計算する。そして，各年の有価証券報告書のテキストマイニングを行いベクトル化した特徴量を入力値，その年の株価収益率を目標値として，中間層が 1 層のシーケンシャルモデルを用いたニューラルネットワークで機械学習を行っている。[11]

■データの処理　日経 NEEDS で使用されている 33 業種分類に従って業種ごとに分析を行う。データ数を確保するために，企業数が 100 社以上存在する業種に分析を絞った。それらは，卸売業（313 社），建設業（163 社），サービス業（502 社），機械（229 社），情報・通信業（520 社），食料品（125 社），不動産業（142 社），小売業（340 社），その他製品（111 社），電気機器（244 社），化学（212 社）の 11 業種であった。

株価の月次データは日経 NEEDS からダウンロード[12]を行った。各年の有価証券報告書の「経営方針」の部分にテキストマイニングを行いベクトル化した特徴量を入力値，その年の株価収益率を目標値として機械学習を行う。つまり，各年の有価証券報告書のテキストデータから，その提出月の翌月から 1 年間の平均株価収益率を予測するモデルを構築することになる。

（10）前多(2022)では，日本の繊維産業について，有価証券報告書の「経営方針」の部分のテキストにテキストマイニングを行い，イノベーションや社会的責任に関連する単語の頻度と株価収益率との関連を回帰分析を用いて分析を行っている。

（11）Bryan（1997)は，株式リターンを従属変数，企業が公表する MD&A（management's discussion and analysis）に関する情報を独立変数とする回帰分析を行い有意に正の関係を持つことを示した。

（12）日経 NEEDS から株価を月次でダウンロードすると，日次のデータを月ごとに平均したデータが出力される。

■テキストマイニングの方法　ここでは，具体的なテキストマイニングの方法について説明する。各業種ごとに以下の 2 つの分析を行った。

（1）（i）業種に属する各会社の 2017 年から 2021 年までの有価証券報告書の文書データの「経営方針」の部分に対してテキストマイニングを行い，会社・年ごとの文書に対する TF−IDF 分析を行い，TF−IDF 値が上位 4,000 位までの 4,000 語を抽出する。抽出した 4,000 語に対する TF−IDF 値を並べた 4,000 次元のベクトルが会社別年別に得られる。

（ii）（i）で求めた会社別年別の 4,000 次元のベクトル値を入力値として，株価の上昇時に $[1, 0]$，下降時にを $[0, 1]$ となる 2 次元ベクトルを目標とする機械学習を行う。

（2）（i）（1）で求めた TF−IDF 値が上位 4,000 位までの 4,000 語に対して Word2Vec により 300 次元のベクトル化を行い[13]，そのベクトル値を足し合わせることで会社別年別のテキストのベクトル表現を得る。

（ii）（i）で求めた会社別年別の 300 次元のベクトル値を入力値として，株価の上昇時に $[1, 0]$，下降時にを $[0, 1]$ となる 2 次元ベクトルを目標とする機械学習を行う。

■ニューラルネットワーク　機械学習では，業種ごとのデータをプーリングして分析を行っている。各企業のデータは時系列データとしては長さが短いため，RNN や LSTM はその特徴を活かすことができないと考え，ニューラルネットワークとしては中間層が 1 層のシーケンシャルモデルを用いている。エポック数と中間層の次元は，精度が高くなるように業種ごとに調整を行った。活性化関数はシグモイド関数を用いている。

機械学習の実行に際しては，全データ数を 2 で割って，教師データ数とテストデータ数を求め，全データから目標値が $[1, 0]$ であるものと $[0, 1]$ であるものが半分ずつになるように教師データを抽出し，残りのデータをテストデータとしている。

結果は表 1 にまとめてある。入力データに Word2Vec とあるものはニューラルネットワークに Word2Vec の結果の 300 次元のデータを入力している分析の結果で，TF-IDF とあるものは TF-IDF 分析の結果の 4,000 次元のデータを入力している分析の結果である。表には業種ごとのテストデータのデータ数，正解率，F 値，訓練データのデータ数，正解率，F 値を表示している。正解率はテストデータ，訓練データともに全ての業種で 50％を超えており比較的良好な結果を示している。F 値については業種によって偏りがあり，不動産業，小売業などで低い値となっている。これらの業種については値の改善を今後図っていきたい。

（13）Word2Vec には，GitHub で公開されている Wikipedia のテキストを使った事前学習済モデルを使用した。

表 1　機械学習の結果

業種	入力データ	エポック数	中間層次元	テストデータ			訓練データ		
				数	正解率	F 値	数	正解率	F 値
卸売業	Word2Vec	15	8	616	0.599	0.738	616	0.511	0.661
	TF-IDF	6	10	616	0.528	0.524	616	0.523	0.510
建設業	Word2Vec	10	6	288	0.642	0.778	286	0.521	0.668
	TF-IDF	5	5	288	0.653	0.789	286	0.507	0.668
サービス業	Word2Vec	10	2	620	0.544	0.688	620	0.513	0.648
	TF-IDF	10	5	620	0.563	0.685	620	0.515	0.636
機械	Word2Vec	10	2	442	0.514	0.676	440	0.502	0.665
	TF-IDF	10	5	442	0.523	0.552	440	0.534	0.553
情報・通信業	Word2Vec	10	2	585	0.706	0.826	584	0.514	0.665
	TF-IDF	10	5	585	0.569	0.683	584	0.505	0.568
食料品	Word2Vec	10	5	237	0.540	0.680	234	0.513	0.663
	TF-IDF	10	5	237	0.515	0.674	234	0.517	0.672
不動産業	Word2Vec	6	2	202	0.530	0.128	200	0.505	0.108
	TF-IDF	10	5	202	0.515	0.234	200	0.520	0.342
小売業	Word2Vec	10	6	558	0.566	0.160	556	0.529	0.288
	TF-IDF	10	5	558	0.543	0.215	556	0.514	0.262
電気機器	Word2Vec	12	2	496	0.597	0.746	496	0.506	0.668
	TF-IDF	10	10	496	0.514	0.546	496	0.518	0.443
化学	Word2Vec	10	6	320	0.525	0.643	318	0.509	0.649
	TF-IDF	8	10	443	0.512	0.455	442	0.536	0.468

参　考　文　献

Antweiler, Werner and Murray Z. Frank (2004) “Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards,” *Journal of Finance*, Vol. 59, pp. 1259–1294.

Bollen, Johan, Huina Mao, and Xiaojun Zeng (2011) “Twitter Mood Predicts the Stock Market,” *Journal of Computational Science*, Vol. 2, pp. 1–8.

Bryan, H. Stephen (1997) “Incremental Information Content of Required Disclosures Contained in Management Discussion and Analysis,” *The Accounting Review*, Vol. 72, pp. 285–301.

Hagenau, Michael, Michael Liebmann, Markus Hedwig, and Dirk Neumann (2012) “Automated News Reading:Stock Price Prediction Based on Financial News Using Context-Specific Features,” 45th Hawaii International Conference on System Sciences.

Li, Feng (2008) “Annual Report Readability, Current Earnings, and Earnings Persistence,” *Journal of Accounting and Economics*, Vol. 45, pp. 221–247.

Schumaker, Robert P. and Hsinchun Chen (2009) “Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin text system,” *ACM Transactions on Information Systems*, Vol. 27, pp. 1–29.

Tetlock, P. C. (2007) “Giving Content to Investor Sentiment: The Role of Media in the Stock Market,” *Journal of Finance*, Vol. 62, pp. 1139–1168.

Wilson, Theresa, Paul Hoffmann, Swapna

Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan (2005) “OpinionFinder: A System for Subjectivity Analysis,” Proceedings of HLT/EMNLP 2005 Interactive Demonstrations.

Zhang, Xue, Hauke Fuehres, and Peter A. Gloor (2011) “Predicting Stock Market Indicators Through Twitter: I hope it is not as bad as I fear,” *Procedia Social and Behavioral Sciences*, Vol. 26, pp. 55–62.

易迪・杉本徹 (2017)「クラスタリングと単語分散表現を用いたニュース記事からの株価動向予測」，第 16 回情報科学技術フォーラム。

沖本竜義・平澤英司 (2014)「ニュース指標による株式市場の予測可能性」，『証券アナリストジャーナル』，第 52 巻，67–75 頁。

加藤大輔・五島圭一 (2020)「有価証券報告書のテキスト分析：経営者による将来見通しの開示と将来業績」，日本銀行ワーキングペーパーシリーズ，2020–J–16。

五島圭一・高橋大志 (2016)「ニュースと株価に関する実証分析：ディープラーニングによるニュース記事の評判分析」，『証券アナリストジャーナル』，第 54 巻，76–86 頁。

坪内孝太・伊藤友貴・山下達雄・和泉潔 (2016)「ファイナンス掲示板情報からの株価予測」，2016 年度人工知能学会全国大会（第 30 回）。

廣瀬喜貴・平井裕久・新井康平 (2017)「MD&A 情報の可読性が将来業績に及ぼす影響：テキストマイニングによる分析」，『経営分析研究』，第 33 巻，87–101 頁。

前多康男 (2022)「テキストマイニングによる繊維産業の企業経営分析」，『三田学会雑誌』，第 115 巻，第 2 号，171–210 頁。

丸山健・梅原英一・諏訪博彦・太田敏澄 (2007)「インターネット掲示板と株式指標の関係に関する研究」，日本社会情報学会全国大会研究発表論文集（第 22 回全国大会）。

吉田政之 (2018)「リスク情報の可読性と将来業績に関する実証分析」，神戸大学大学院経営学研究科大学院生ワーキング・ペーパー，第 06a 巻，1–16 頁。

吉田稔・中川裕志・石田智也・池田翔・中嶋啓浩・松井藤五郎・本多隆虎・和泉潔 (2011)「ニュース記事クラスタリングによる取引高予測の試み」，2011 年度人工知能学会全国大会（第 25 回）。