Examination of Interpretability of Text Information in Quarterly Financial Statements for Stock Price Fluctuations Using Decision Tree Regressor

Mana MATSUDA* and Hiroshi TSUDA*

(Received June 28, 2023)

Due to the development of natural language processing technologies, the effectiveness of text-based investment strategies for stocks has been demonstrated. Focusing on quarterly financial statements, the polarity of the text is often quantified and used for investment strategies. In this study, we quantify the detailed content of quarterly financial statements by using Sentence-BERT to create features and we examine interpretability of features on stock price fluctuations by using decision tree regression. As a result, we find features which have higher interpretability on stock price fluctuations from quarterly financial statements of many companies compared to regression analysis. In conclusion, we find that decision tree regression can realize the non-linear patterns between the content of quarterly financial statements and stock price fluctuation.

Keywords: natural language processing, stock price fluctuation, decision tree, regression, deep learning

キーワード：自然言語処理，株価変動，決定木，回帰，深層学習

決定木回帰を用いた決算短信テキスト情報の株価変動に対する説明力の検証

松田 眞，津田 博史

1. はじめに

近年，機械学習モデルや自然言語処理技術の発展に伴い，様々なドメインでテキストデータの活用が進んでいる．特に最近では，ChatGPT[1)]が話題である．ChatGPT は様々なドメイン知識を学習していることから，多くの人が日常的に利用する検索ツールとして浸透しつつある．このように GPT[2)]をはじめとする深層学習モデルでは，テキストを精度よく定量化することが可能であり，質疑応答だけでなく，様々な自然言語のタスクに適用することが可能である．金融分野においても，自然言語処理技術を活用しようとする動きは加速しており，関連する研究も増加傾向にある．また，坂地(2023)[3)]は実務家が関わる研究が増えてきていることなどから，国内外の金融・経済において自然言語処理が定着してきていることを示唆した．例えば，河村ら(2021)[4)]は決算短信を対象にテキストマイニングを行い，個人投資家にとって重要な情報である業績要因を含む業績予測文の抽出に取り組んだ．この背景には，企業から大量に発表される金融テキストを人手で目を通し，重要

*Department of Mathematical Science, Doshisha University, Kyoto
Telephone : +81-774-65-6681, E-mail : ctwh0908@mail4.doshisha.ac.jp, htsuda@mail4.doshisha.ac.jp

な情報を見つけ出すことが困難であるということがある．また，投資をするうえで，それぞれの情報がどの程度重要であるかを判断し，投資戦略を作成することも困難である．そこで，自然言語処理技術や機械学習の手法を活用して，有用な投資戦略を作成する取り組みも存在する．このような研究では，金融テキスト以外にニュース記事やSNSの投稿を用いた研究が多く存在するが，ここでは，本研究で対象とする決算短信を用いた先行研究を紹介する．はじめに，決算短信を用いて株式市場を予測し，投資戦略を作成する研究が少ないことに注意されたい．1つ目の先行研究は，白方(2018)[5]による研究である．白方は，決算短信・有価証券報告書・四半期報告書を対象にLSTM[6]を用いたセンチメント分析を行った．特に決算短信については，センチメントを表す極性値を用いた業績予想と株価予測について検証した．その結果，業績予想について，極性値には将来の業績の方向を予測する可能性があることが分かった．一方で，株価予測については，極性値が景気変動の影響を考慮しきれていなかったことから，極性値だけを用いた株価予測は困難であることを示唆した．また，白方・津田(2018)[7]では，日本銀行の金融経済月報に対してRNN†を用いた日経平均株価の株価指数変化率の予測を行った．この研究では，テキストデータを単語(名詞，動詞，形容詞，副詞)の頻度を用いて定量化した．その結果，ベンチマークとした主成分回帰(PCR)‡による予測精度を上回った．また，RNNの作成期間を調整することで，トレンド的中率も改善できた．決算短信に関する2つ目の先行研究は，山本ら(2020)[8]による研究である．山本らは，決算短信と四季報のテキストを対象にネガティブ・ポジティブ単語辞書を用いて投資情報を抽出し，定量化を行った．そして，景気・企業業績関連指数や企業属性との関連性や，株価リターンとの関連性について分析し，投資戦略としての利用可能性を検証した．その結果，決算短信や四季報から抽出した投資情報は，成長特性を示し，さらに成長特性をもつ数値情報を除いた後でも，投資戦略として有効であることが分かった．つまり，既存の投資戦略とは異なる超過収益の源泉を有していることを明らかにした．このように先行研究において，決算短信の特性や，投資戦略としての有効性が示されつつある一方で，決算短信に適したテキストマイニングの探究は課題の1つである．先行研究では，決算短信の極性情報が定量化されることが多かった．そこで松田・津田(2023)[9]の研究では，Sentence-BERT[10]を用いて得た分散表現から直接特徴量を作成した．Sentence-BERTは事前学習済みのBERT[11]を類似度指標によってファインチューニングしたモデルであり，テキストの意味をより精度よく表した分散表現を獲得できる．そのため，決算短信に含まれる詳細な内容を定量化できると考えられ，極性情報より有効な投資情報を抽出することを目的として用いた．松田・津田は1つの決算短信から分散表現の次元数の特徴量を生成し，それぞれの分析対象企業の過去10年の株価変動に対して回帰分析を用いて時系列回帰を行い，特徴量の統計的有意性を検証した．ここで，目的変数を株価変動と表したが，実際には，固有銘柄の四半期収益率を日経平均§の四半期収益率で回帰したときの残差を目的変数としている．以後この目的変数を株価変動と表記する．また特徴量の具体的な作成方法は，決算短信から一文単位(極性単位)でテキストを抽出し，分散表現を獲得したうえで，分散表現の各成分に着目して，ある決算短信に含まれるテキストの成分値を10個の階級に分割し，それぞれの階級の相対度数を重みとして，決算短信に含まれる全テキストで加重平均をとったものを特徴量とした．ここで，極性単位とは，ネガティブ・ポジティブの極性が混在する一文に対して，極性が切り替わる位置で分割してテキストを抽出する方法を意味する．このようにして，有意性の検証を行った結果，分析対象としたほとんどの企業で有意な特徴

† 回帰型ニューラルネットワーク(Recurrent Neural Network). 各層の中間層で情報を伝達する構造をもつため，時系列データのモデリングが可能である．

‡ 説明変数の主成分を用いて回帰モデルを構築する手法.

§ 日経平均株価©日本経済新聞社．残差は筆者が独自に算出し，日本経済新聞社は一切関与していない.

量が存在し，過去 10 年の株価変動に対して，説明力を有する特徴量が存在することが分かった．さらに，特徴量と紐づく成分におけるテキストと成分値の関係に着目すると，高い有意性をもつ理由を説明することができた．例えば株価変動に対して有意性をもち，正の相関がある成分の場合，高い成分値をもつ決算短信には，高い成分値をもつテキストが多く含まれ，それらのテキストは業績や財務状態についてポジティブな内容を表していた．また，テキストと成分値の関係を考察するうえで，極性単位のテキストの方が一文単位のテキストと比べて，有意性をもつ理由を説明しやすいことが分かった．理由は，極性単位にすることで，一文の中で極性が相殺される状況を防ぐことができ，極性の線形な分布が強まったからだと考えられる．しかし，先に述べた通り，分散表現は意味を定量化する性質があることから，株価変動に対してネガティブ・ポジティブな情報の成分値を線形なパターンで認識するより，非線形なパターンで認識する方が適していると考える．そこで，本研究では回帰分析の代わりに決定木 [12]を用いた回帰を行った．これによって多くの分析対象企業で，回帰分析と比較して，より高い説明力をもつ特徴量を見出すことができた．本稿では，第 2 章で分析に使用したデータについて，第 3 章で分析手法について，第 4 章で結果と考察について，第 5 章でまとめ，第 6 章で今後の課題について述べる．

2. 使用データ

本研究では，松田・津田(2023)と同様のデータを用いた．具体的には，Table 1 に示す 20 業種のいずれかに属する，106 社を対象にした．対象企業は，東証プライム市場に上場している 3 月決算の企業であり，2012 年度から 2021 年度まで継続して日経平均構成銘柄に選ばれている企業に限定し，さらに，2012 年度から 2021 年度までに発表された全ての決算短信から自動的にテキストを抽出できた企業に限った．固有銘柄の株価は JPX データクラウドから購入したものを使用し，日経平均株価は日経プロファイルのヒストリカルデータから該当日の始値を取得した．また，四半期収益率は，次期の決算短信発表日の翌営業日の始値に対する，当期の決算短信発表日の翌営業日の始値の収益率である．これより，過去 10 年間に発表された約 4200 レポートの決算短信を対象に，企業ごとに，株価変動に対する決算短信テキスト情報の説明力について検証する．

Table 1. Target industry for analysis.

Iron & Steel	Pharmaceuticals	Other manufacturing	Chemicals
Machinery	Air transportation	Electrical equipment	Shipping
Automotive	Trading company	Land transportation	Food
Warehousing	Shipbuilding	Precision machinery	Service
Construction	Railway & Bus	Nonferrous metals	Retail

3. 分析手法

本章では，決定木による回帰について述べ，本研究における決定木の利用手法についても述べる．

3.1. 決定木

決定木とは，木構造の決定テストの集合からなり，分割統治法として機能する．葉でないノードには，分割(特徴テスト)が与えられ，ノードに与えられたデータは分割に従って異なる部分集合に分離される [13]．そして，葉ノードに付与されたラベルや予測値が，最終的な葉ノードに属するデータの予測値と

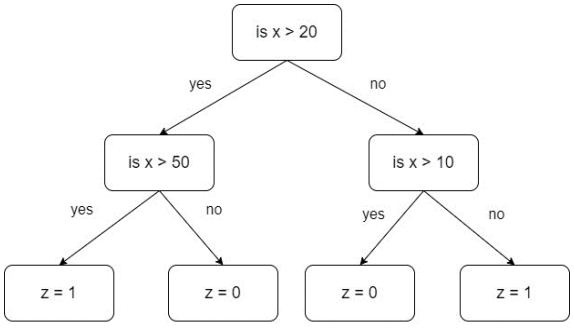

Fig. 1. Sample decision tree.

なる．例として，分類問題に対する決定木を Fig. 1 に示す．ここで，変数xを年齢，変数zをある商品のテレビ CM を見たかどうかを表す変数とすると，この決定木では，年齢によってテレビ CM を見たかどうかを推定することができる．このように決定木のアルゴリズムでは，再帰的に分割によるデータの分離が行われる．そこで，回帰における決定木の分割の選択方法について次項で述べる．

3.2. 決定木による回帰

前項で述べたように，決定木の各ノードでは分割が与えられ，分割によって葉ノードに達するまで再帰的にデータの分離が行われる．そのためアルゴリズムにおいて分割の選択が重要である．回帰では，二乗誤差を利用して分割が作成される．二乗誤差の計算に用いる平均値の推定値には，ノードに分類されたデータの平均値を用いる．ノードに分割が与えられたとき，データを分割が真であるグループと偽であるグループに分類できる．ここで，それぞれのグループをYes，Noと表し，データ数をN_{yes}，N_{no}と表す．さらに，回帰対象の変数をYとする．このとき，それぞれのグループ側の平均値の推定値は，

$$\hat{Y}_{yes} = E[Y_{yes}] = \frac{1}{N_{yes}} \sum_{i=1}^{N_{yes}} Y_{yes}^{i}, \qquad (1)$$

$$\hat{Y}_{no} = E[Y_{no}] = \frac{1}{N_{no}} \sum_{i=1}^{N_{no}} Y_{no}^{i}. \qquad (2)$$

これらの推定値を用いると，2 つのグループの二乗誤差の合計は，

$$\frac{1}{N_{yes}} \sum_{i=1}^{N_{yes}} \left(Y_{yes}^{i} - \hat{Y}_{yes}\right)^2 + \frac{1}{N_{no}} \sum_{i=1}^{N_{no}} \left(Y_{no}^{i} - \hat{Y}_{no}\right)^2 \quad (3)$$

となる．この二乗誤差の合計が最小になるように，分割を作成する．

3.3. 本研究における決定木の利用と評価方法

本研究では，決算短信のテキストから作成した，分散表現の各成分に紐づいた特徴量を説明変数に用いて，各社の時系列データである株価変動に対して時系列回帰を行う．ここで，決定木の構築に用いる説明変数は 1 個であるが，それぞれの企業で分散表現の次元数(384 次元)の特徴量が存在するため，決定木は 384 個作成されることに注意されたい．そのうえで，モデルの予測精度を表す決定係数が最大の決定木を採用して，同様の特徴量を用いて回帰分析したときの決定係数の最大値と比較する．本研究では，1 社当たり過去 10 年分の決算短信(40 レポート)を分析対象とするため，1 社あたりのデータ数は 40 である．そこで，データの 8 割を学習用データ，2 割を検証用データに用いた．決定木の構築において，決定木の深さをハイパーパラメータとして指定できる．本研究では，2 以上の深さを指定して学習用データで決定木を構築し，検証用データに対する作成済みの決定木の決定係数を評価指標として用いる．ここで，先に述べたように，それぞれの企業では 384 個の特徴量及び，決定木が作成されるが，決定係数が最大になるものを採用し，以下の決定木の深さの決定にも用いる．本研究では，業種全体の決定係数の平均値を算出することで，平均値が最大になる深さを採用した．ただし，平均値の算出には，決定係数が負の値をもつ企業は除いた．このように，決定木の深さを決めたことにより，任意の企業で決定係数が最大になる深さを採用していないことに注意されたい．

4. 結果・考察

本研究は，決算短信のテキストから分散表現を用いて得られた特徴量を説明変数として，株価変動を目的変数に用いてモデルを構築した．モデルの精度の評価には，決定係数R^2を用いる．決定係数は，

$$R^2 = 1 - \frac{\sum_{i=1}^{8} (y_i - p_i)^2}{\sum_{i=1}^{8} (y_i - \hat{y})^2} \qquad (4)$$

で定義される．ここで，検証用データ数が 8 であることから$i = 1,2,\cdots,8$である．また，y_iは株価変動の観測値で，p_iは予測値である．$\hat{y}$は株価変動の観測値の平均値である．

4.1. 決定係数の比較

3.3 節で述べたように決定係数を基準にして，特徴量及び，決定木を選択した．(各業種で採用した

決定木の深さを Table 2 に示す．ただし，空運に属する企業は 1 社(ANA ホールディングス)しかなく，深さ 5 まで増やしたところ，負の決定係数しか得られなかったため深さを記載しない．)その結果，両手法で決定係数が負の値になる企業を除いた 93 社のうち 88 社(約 95%)において，決定木で回帰を行うことで，決定係数が向上した．また，Table 3 では，決定係数が 0 未満，0 以上 0.5 未満，0.5 以上の範囲に

Table 2. Depth of decision tree for each industry.

Iron & Steel: 7	Pharmaceuticals: 4	Other manufacturing: 2	Chemicals: 2
Machinery: 2	Air transportation: --	Electrical equipment: 7	Shipping: 3
Automotive: 4	Trading company: 4	Land transportation: 3	Food: 2
Warehousing: 3	Shipbuilding: 4	Precision machinery: 2	Service: 3
Construction: 4	Railway & Bus: 3	Nonferrous metals: 2	Retail: 2

Table 3. Coefficient of determination and number of data.

	dicision tree regression	regression analysis
$R^2 < 0$	14	41
$0 \leq R^2 < 0.5$	34	60
$R^2 \geq 0.5$	58	5

含まれる企業数を手法ごとに示した．

Table 3 から分かるように，負の決定係数をもつモデルが全体の約 25%減少し，0.5 以上の決定係数をもつモデルが全体の 50%増加した．これより，決定木を用いて株価変動に対して回帰を行うことで，回帰分析を用いた場合と比較して，より多くの企業で株価変動に対して説明力を有する特徴量を見出すことができた．加えて，特徴量が有する説明力の大きさを向上させることもできた．一方で，回帰分析より精度が下がった企業が 5 社(東ソー，第一三共，鹿島建設，JFE ホールディングス，ニチレイ)存在した．ここでは，東ソーに着目して，回帰分析の精度を下回った理由を考察する．東ソーでは，決定木による回帰と回帰分析で作成したモデルの決定係数がそれぞれ 0.38，0.65 であった．また，それぞれの手法で株価変動に対して最も高い説明力を有する成分は異なった．具体的には，決定木が第 5 成分，回帰分析が第 320 成分であった．Figure 2，Fig. 3 は，それぞれの成分値(横軸)と株価変動(縦軸)の散布図に，決定木，回帰分析による予測を重ねて可視化したものである．どちらのグラフも青いプロットが学習用データ，赤いプロットが検証用データを表している．よって，直感的には，赤いプロットと緑の実線との誤差が小さいほどモデルの精度が高い．(以後同様のグラフの説明は省略する．)また，Fig. 2 の破線は，決定木の葉ノード以外のノードで与えられる分割によって得られる領域である．グラフから，回帰分析の方が検証用データに近い値を予測できている．これは，決定木は領域ごとに定数で予測を行うことが原因だと考えられる．さらに，四分位範囲を用いた株価変動の外れ値検出(四分位範囲の 10 倍を閾値に使用)を行ったところ，決定木による手法で決定係

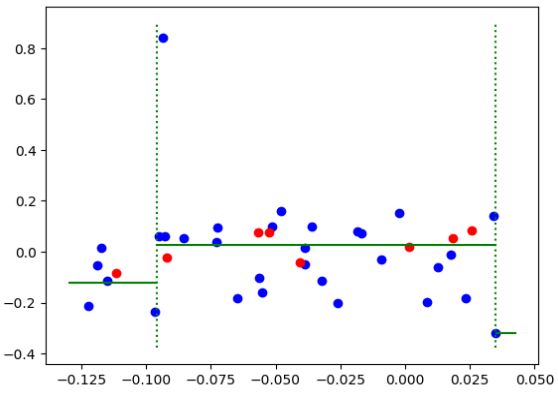

Fig. 2. Prediction by decision tree regression.

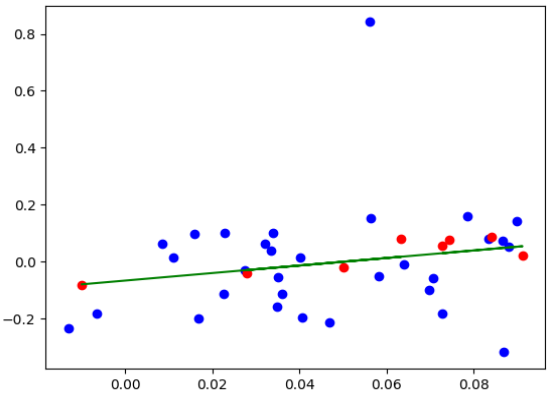

Fig. 3. Prediction by regression analysis.

数が 0 を下回った 14 社のうち 10 社が外れ値を有した．また，これらは検証用データに外れ値を含んでおり，モデルが外れ値に対応できないため，検証用データに対する誤差が大きくなり，決定係数が下がったと考えられる．最後に，同様の手法で検出した外れ値を除いたうえでモデルを構築し，決定係数を算出し直した．その結果，両手法で決定係数が負の値になる企業を除いた 103 社のうち 97 社(約 94%)において，決定木で回帰を行うことで，決定係数が向上した．また，決定係数が 0 未満，0 以上 0.5 未満，0.5 以上の範囲に含まれる企業数は Table 4 のとおりである．この結果から，外れ値が精度の低下の原因であったことが分かった．これ以降，外れ値を除いて作成したモデルを用いて議論する．また，リコーのように，決定木と回帰分析で株価変動に対して最も説明力が高い成分が異なる企業は全体の約 95%であり，同じ成分の企業は 5 社(ヤマハ，武田薬品工業，京王電鉄，神戸製鋼所，リコー)であった．この 5 社については，いずれも決定木を用いることで決定係数が向上した．Figure 4，Fig. 5 では，京王電鉄の株価変動に対して最も高い成分(第 246 成分)と株価変動の関係と，決定木と回帰分析による予測を示す．Figure 4，Fig. 5 から，決定木で第 246 成分値と株価変動の非線形な関係を認識できたことで，検証データと予測の誤差を小さくすることができた．ここで第 246 成分値とテキストの内容の関係を考察する．特に，Fig. 4 において破線で区切られた領域のうち，左から 4 番目(第 4 領域)と 8 番目(第 8 領域)の領域に着目し，成分値がそれらの領域に含まれる決算短信において，成分値の度数が高い階級(上位 3 階級)に属するテキストの内容を考察する．第 4 領域の成分値をもつ決算短信には，ネガティブな内容としては，「百貨店業などの事業の減収・減益」「店舗休業による影響」「固定資産除去費の増加」などが記載されていた．また，ポジティブな内容として，「収益基盤の拡充」「運輸業などの事業における増収」「コスト削減」「ホテル業の改善」「純利益の増加」「負債の減少」「不動産販売業の売上増」などの記載があった．一方で，第 8 領域の成分値をもつ決算短信には，ネガティブな内容として，「減価償却費の増加」「手元資金の減少」「純資産の減少」などの記載があり，ポジティブな内容として，「座席指定料金収入の増加」「ホテル業の増収」「株式併合による配当額の増加」「レジャー・サービス業などの事業で増収・増益」「バス事業の増収」「輸送人員の増加」「販売戸数の増加」「親会社株主に帰属する純資産の増加」などの記載があった．列挙したように，両方の領域にポジティブ・ネガティブなテキストが含まれているが，それぞれ異なる内容が多いことが分かる．このような記載の組み合わせが成分値のそれぞれの領域の特徴であると考えられる．特に，第 8 領域に成分値が属する 2 つの決算短信のテキストは似たような成分値の分布をもち，株価変動も近い値であった．ここで，2018_2(第 8 領域，青いプロッ

Table 4. Coefficient of determination and number of data.

	decision tree regression	regression analysis
$R^2 < 0$	4	19
$0 \leq R^2 < 0.5$	26	79
$R^2 \leq 0.5$	76	8

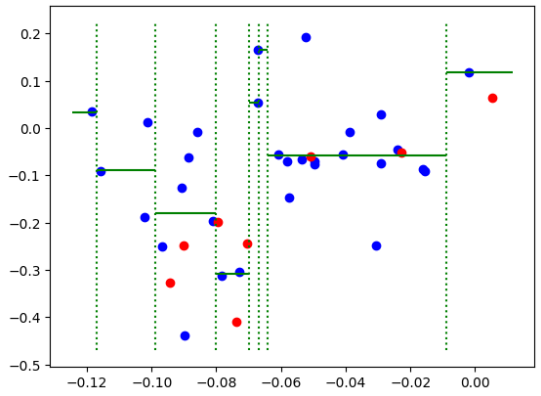

Fig. 4. Prediction by decision tree regression.

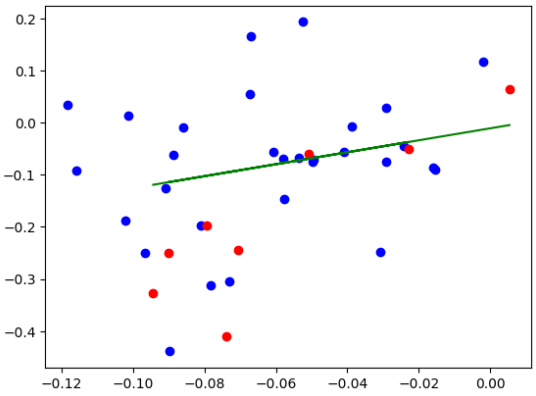

Fig. 5. Prediction by regression analysis.

ト)には 2018_1(第 8 領域，赤いプロット)に含まれない「配当額の増加」を意味する表記があった．この情報が，2018_2 において，2018_1 の株価変動を上回った原因となる情報源の 1 つである可能性があると考える．

4.2. 株価変動に対して説明力を有する成分

本節では，決定木の予測精度をもとに，株価変動に対して高い説明力をもつ特徴量について考察する．ここでは，その一例として日本製鋼所に着目する．日本製鋼所では，第 295 成分の特徴量を用いて作成した決定木(深さ 3)の決定係数が 0.72 で，決定係数の最大値であった．ここで最大値とは，その他の 383 個の特徴量を用いて作成したモデルの決定係数に対して最大であるという意味である．また，先に述べた通り，第n成分の特徴量とは，ある決算短信がm個のテキストで構成されているとき，Sentence-BERT を用いてm個の分散表現を獲得する．さらに，それぞれの分散表現の第n成分値の相対度数分布から，成分値に対して対応する階級の相対度数を重みとした加重平均を算出し，特徴量とした．つまり，それぞれの特徴量は分散表現の成分と紐づいており，本節では，株価変動に対して説明力をもつ特徴量に紐づく成分に着目し，成分値とテキストの内容の関係を考察する．はじめに，日本製鋼所では，回帰分析を用いたときは，決定係数が正の値をもつモデルを作成できなかった．つまり，回帰分析では株価変動に対して説明力をもつ特徴量を決算短信から見出すことができなかった．Figure 6 では，第 295 成分値(横軸)と株価変動(縦軸)の散布図に，破線と実線を用いて決定木による各領域の予測を重ねて可視化したものである．青いプロットは学習用データ，赤いプロットは検証用データを表す．ここで検証用データの中から，成分値の各領域(左から①②③④⑤⑥⑦とする)を代表するデータを適当に選択する．ただし，②⑤には検証用データが属さないため省略する．具体的には，①2017_4，③2018_3，④2020_4，⑥2014_3，⑦2019_3 を選択した．({year}_{n}は year 年度の第 n 四半期の決算短信を意味する．)次に，例として 2020_4 の決算短信に含まれるテキストに対応する成分値のヒストグラムを Fig. 7 に示す．それぞれの決算短信において，特に，度数が高い階級に注目し，テキストの内容を確認する．

2017_4 では，「国内外の景気の緩やかな回復・成長」「財務状況(利益剰余金など)の報告」「売上高減少による営業損失」「今後の見通し・心構え」「現存損失」「営業赤字の継続」「需要低迷の見込み」などの記述があり，今後の目標などが多く記載される中，現状に関してネガティブな記述が存在した．

2018_3 では，「欧州では輸出が鈍化，緩やかな成長」「中国の景気の減速」などの記載があり，海外経済が停滞している記述があった．

2020_4 では，「設備投資の抑制」「競争の激化」「プロジェクトの遅れ」「厳しい状況が続く」「投資抑制が進む」「経済が持ち直す動きで推移」「世界の景気持ち直しの動きの継続を期待」「業績の堅調な推移」「事業活動の向上に向けた合併」「需要は緩やかに持ち直した」「生産の持ち直し」などの記載があり，コロナ禍で落ち込んだ景気の低迷と回復傾向にある見通しの記述があった．

2014_3 では，決算短信に含まれるテキストが少ないが，「部品の不具合に起因する損失」や「大口売上の減少(別の製品は増加)」という記述があった．

2019_3 では，「子会社の吸収合併」「子会社化」「収益の柱であった部材の市場縮小・競争激化」「収益力改善の課題」「受注高減少」「売上高減少」などの記載があり，子会社化や合併が与える影響を一概に述べることはできないが，受注高や売上高については減少の傾向の記述があった．

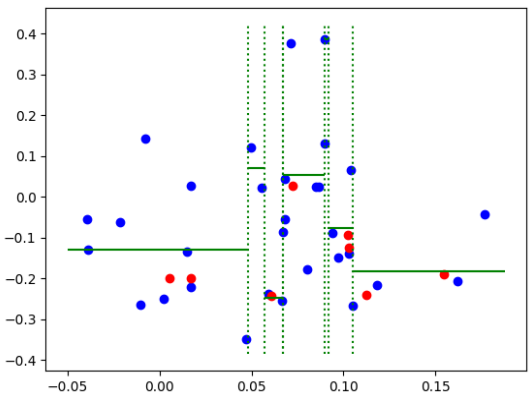

Fig. 6. Prediction by decision tree regression.

ここまで，それぞれの決算短信に含まれるテキストのうち極性を判定しやすいものをピックアップして示した．その結果，決算短信に含まれるテキストの内容と，決定木の各領域に対する予測値の符号に関連性があることが定性的に評価できると考える．具体的には，2020_4 ではコロナ禍からの回復が進み，今後の明るい展望について多く書かれていた．また，2020_4 が属する領域④に対して，決定木は正の値を予測しており，決算短信に記載された明るい展望が株価変動に正の影響を及ぼした可能性があると考えられる．このように成分値の大きさとテキストの極性が線形なパターンでないのは，分散表現がテキストの意味を反映するという性質をもつからである．このことから，領域④における株価変動と第 295 成分値の関係を捉えられたのは，決定木では非線形なパターンを捉えられるからだと考える．実際に，回帰分析で作成可能な線形モデルでは，任意の特徴量と株価変動の関係を捉えることができなかった．したがって，回帰分析の場合と比較して，決定木を用いることで，決算短信の詳細な内容を反映した特徴量を活用して，株価変動を説明できることが分かった．

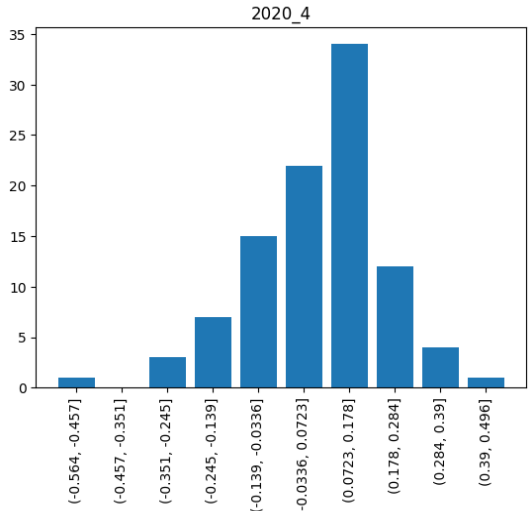

Fig. 7. Histogram of the values of the 295th component

5. まとめ

本研究では，20 業種 106 社を対象に，2012 年度から 2022 年度に発表された決算短信を用いた分析を行った．具体的には，決算短信のテキストを Sentence-BERT の分散表現を用いて定量化し，各成分から複数の特徴量を作成した．そして，決定木と回帰分析を用いて，特徴量がもつ株価変動に対する説明力を検証した．その結果，決定木を用いることで，回帰分析と比較して，約 94%の企業で決定係数を向上させることができた．また，負の決定係数の企業を全体の約 14%減らすことができ，決定木を用いることで，株価変動に対して説明力をもつ特徴量をより多くの企業の決算短信から見出せることが分かった．さらに，説明力が向上した理由を考察した．その結果，決定木では，分散表現の成分値と株価変動の非線形なパターンを捉えられることが分かった．分散表現はテキストの意味を反映するという性質があるため，株価変動への影響という観点で成分値に線形なパターンを認識することは困難だと考える．これより，決算短信の詳細なテキスト情報を定量化した特徴量を活用して株価変動を説明するには，回帰分析より，非線形な決定木が適していると考える．また，本研究で用いた特徴量は決算短信に含まれる内容の組み合わせを数値に反映しており，過去にそのような内容の組み合わせが株価変動に及ぼした影響を評価できたと考える．

6. 今後の課題

本研究で用いた決算短信テキスト情報の定量化手法の信頼性を高める目的で，業種や企業数を増やした分析を行うことが課題である．そのうえで，決算短信のテキストをより反映する特徴量の作成や新たな特徴量を追加することにも取り組みたいと考える．また，決算短信だけでなく，有価証券報告書や四半期報告書に対して同様の手法で，テキストから得られた特徴量の株価変動に対する説明力を検証することで，本研究の手法が決算短信以外の金融テキストでも有効であるかどうかを検証することも課題である．さらに，本研究で見出した特徴量を基準にポートフォリオを作成し，投資シミュレーションを行うことで，株価変動に対する予測能力を検証することも今後の課題である．

参考文献

1) L.Ouyang, J.Wu, X.Jiang, D.Almeida, C.L.Wainwright, P.Mishkin, C.Zhang, S.Agarwal, K.Slama, A.Ray, J. Schulman, J.Hilton, F.Kelton, L.Miller, M.Simens, A.Askell, P.Welinder, P.Christiano, J.Leike, and R.Lowe, "Training Language Models to Follow Instructions with Human Feedback", *arXiv:2203.02155*, (2022).
2) T.B.Brown, B.Mann, N.Ryder, M.Subbiah, J.Kaplan, P.Dhariwal, A.Neelakantan, P.Shyam, G.Sastry, A.Askell, S.Agarwal, A.Herbert-Voss, G.Krueger, T.Henighan, R.Child, A.Ramesh, D.M.Ziegler, J.Wu, C.Winter, C.Hesse, M.Chen, E.Sigler, M.Litwin, S.Gray, B.Chess, J.Clark, C.Berner, S.McCandlish, A.Radford, I.Sutskever, and D.Amodei, "Language Model are Few-Shot Learners", *arXiv:2005.14165*, (2020).
3) 坂地 泰紀, "金融・経済ドメインを対象とした言語処理の新展開", 自然言語処理, **30**[2], 839-843, (2023).
4) 河村 康平, 高野 海斗, 酒井 浩之, "決算短信からの業績要因を含む業績予測文の抽出", 人工知能学会全国大会論文集, (2021).
5) 白方 健司, "自然言語処理と機械学習による株式市場の予測", 同志社大学修士論文, (2018).
6) H.Sepp and S.Jürgen, "Long Short-Term Memory", **9**[8], Neural Comput, 1735–1780, (1997).
7) 白方 健司, 津田 博史, "自然言語処理と深層学習を用いた株式市場の予測", 同志社大学ハリス理化学研究報告, **59**[3], (2018).
8) 山本 零, 川代 尚哉, 栗田 昌孝, "決算短信と四季報テキスト情報の投資戦略への利用可能性検証", ジャフィー・ジャーナル, **18**, 46-62, (2020).
9) 松田 眞, 津田 博史, "Sentence-BERT を用いた決算短信のテキスト情報の株価変動に対する統計的有意性の検証", 同志社大学ハリス理化学研究報告, **64**[2], 37-48, (2023).
10) N.Reimers and I,Gurevych, "Sentence-BERT:Sentence Embeddings using Siamese BERT-Networks", *arXiv:1908.10084*, (2019).
11) J.Devlin, M.Chang, K.Lee, and K.Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv:1810.04805*, (2018).
12) L.Breiman, J.Friedman, R.Olshen, and C.J.Stone, *Classification and Regression Trees*, (Brooks/Cole Publishing, Monterey, 1984).
13) Zhi-Hua zhou, *Ensemble Methods -Foundations and Algorithms-*, 宮岡悦良・下川朝有訳, (近代科学社, 東京, 2017), p5.