ヤフーファイナンス掲示板の投稿を用いた株投資リスク低減のための日経VI上昇予測

佐々木 皓大[1]　諏訪 博彦[1,2]　小川 祐樹[3]　梅原 英一[4]　山下 達雄[5]　坪内 孝太[5]　安本 慶一[1]

概要：近年では，資産運用に株式投資を選択する人が増えている．株式投資において，将来市場の動向を予測することは投資リスク低減のために重要である．投資家が将来の日本市場をどう想定しているかを表す指標として，日経 VI がある．この指標の上昇は，投資家が市場の先行きに不安があることを示し，これを予測することで投資リスクの低減に利用できる可能性がある．一方ソーシャルメディアでは，株式投資に関連した投稿も多くなされている．これらには，投稿者の気持ちや意見が含まれている．そこで本研究では，ソーシャルメディアの投稿を集約し，そこで議論されている話題の変化を捉え，日経 VI の上昇を予測する手法を検討する．

キーワード：自然言語処理，日経 VI，機械学習，ソーシャルメディア分析

1. はじめに

資産運用に株式投資を選択する人が増えており，楽天証券が 2020 年 4 月に発表した記事によると，2020 年 3 月の月間新規口座開設数が約 16 万を超え，業界最多を更新したことを伝えている．株式投資において，現在の市場を的確に把握し，今後の動向がどうなるのかを予測していくことは，超過リターンの最大化や投資リスク低減のためにとても重要である．

超過リターンの最大化を達成するための金融商品の値動き予測に関する研究は数多く存在する [1], [2], [3], [4]．一方でリスク低減を目的とした研究では，ポートフォリオ最適化問題に取り組んでいる例 [5], [6] や，企業のイベントに投資家グループがどのように反応するのかを調査した例 [7] がある．これらの研究においては証券会社のトレーディングにおけるリスク分散や上場企業のリスク低減を目的としており，多種多様な銘柄に投資している個人投資家ひとりひとりが利用するのには適していない．このような問題を解決する方法として，市場の価格変動の大きさ，すなわちボラティリティに注目することで市場全体に対する投資リスクを判定することによるリスク低減手法を提案する．

本研究では市場の投資リスクを，日経平均株価市場の今後 1 か月間のボラティリティである日経平均ボラティリティ・インデックス（以下日経 VI と呼ぶ）の上昇と定義し，投資リスクの予測問題として定式化を行う．日経 VI の上昇予測手法について，我々は過去に LDA を使用してソーシャルメディアから得た文書ベクトルを用いた手法の予測可能性を示唆した [8]．しかしこの研究では長期的な期間の評価が行われていない．そこで本研究では，より長期間で精度の評価を行う．さらに精度向上を目的とした，予測手法を提案する．ソーシャルメディアの投稿から，短い文書に強い Doc2Vec や長い文書にも有効な BERT を用いて文書ベクトルを獲得し，獲得した文書ベクトルを日別にまとめることで，日次の話題ベクトルを生成する．話題ベクトルと金融時系列データを特徴量として，ロジスティック回帰，ランダムフォレスト，LigthGBM の機械学習手法を組み合わせることにより，日経 VI の上昇予測モデルを構築し，精度の比較検討をする．さらに金融時系列は営業日のみのデータであるが，ソーシャルメディアは毎日投稿されているため休日の情報も取り入れることができる．そこで休日明けの営業日を予測する場合は，前日の話題ベクトルを特徴量とする休日処理を行う．

2. 関連研究

日経 VI の上昇予測について，ソーシャルメディアを用いた手法の予測可能性を示唆している．諏訪ら [8] は，ヤフーファイナンス掲示板の投稿を LDA を用いてトピック

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[2] 理化学研究所 革新知能統合研究センター
RIKEN Center for Advanced Intelligence Project
[3] 立命館大学 Ritsumeikan University
[4] 東京都市大学 Tokyo City University
[5] Yahoo!JAPAN 研究所 Yahoo!JAPAN Research

を獲得し，日経平均 VI の上昇予測に用いている．

彼らは，投稿を形態素解析で形態素に分割し，LDA を用いて各文書における 100 種のトピックに所属する確率を獲得した．この獲得したベクトルを日別にまとめ機械学習モデルの入力としてる．そして日経平均 VI の 2σ 以上上昇とその他の 2 値分類問題を設定し，ロジスティック回帰およびランダムフォレストでモデルを作成，結果としてロジスティック回帰を用いたモデルにおいて 2014 年 11 月 17 日から 2016 年 6 月 29 日の 395 営業日で評価したところ，Precision，Recall ともに 0.45 の精度を得ている．さらに佐々木ら [9] は，諏訪ら [8] の提案した手法の有効性を評価するため，イントラデイデータの価格情報に基づいた日経平均オプション取引の売買シミュレーションを開発した．その結果，諏訪ら [8] の手法の有効な可能性を示唆した．しかしこの評価期間の日本市場は上昇トレンドであったため，評価期間を延ばして評価する必要性がある．また彼らは予測精度の向上を課題にあげている．そこで本研究では近年発展している自然言語処理の分散表現獲得技術と機械学習手法を組み合わせて比較検討を行う．

3. 問題設定

市場の荒れ具合を知ることは，投資のリスクを減らすうえで重要な要素のひとつである．市場の荒れ具合は，価格変動の変動率 (ボラティリティ) で測ることができる．日本の市場では，日本経済新聞社が日経平均オプションのプレミアムから算出している日経平均ボラティリティ・インデックス (日経 VI) があり，30 日後のボラティリティを表す指数である [10]．日経 VI はリーマンショックが起こった 2008 年 10 月に 92.03，東日本大震災が起こった 2011 年 3 月は 69.88 を記録している．市場が最も不安定な時に最高水準に達するため，日経 VI は恐怖指数とも呼ばれている．そこで本研究では，投資リスク r を日経 VI の大幅な上昇と定義し，定式化を行う．

まず日経 VI の大幅な上昇を定義する．v_t は t 日目の日経 VI とする．ここで $t \in 1,2,\cdots,T$ であり，T は実験期間中の営業日の日数であるとき，日経 VI の 1 日差分 x_t は

$$x_t = v_{t+1} - v_t \tag{1}$$

と表すことができる．このとき日経 VI の 1 日差分の標準偏差は以下である．

$$\sigma = \sqrt{\frac{1}{T-1}\sum_{k=0}^{T-1}(x_k - \bar{x})^2} \tag{2}$$

日経 VI の大幅な上昇と判定するための閾値 α を標準偏差 σ を用いて，

$$\alpha = i\sigma \tag{3}$$

と定義する．制約条件として，i は正の実数とする．さらに本研究では上昇のみを考慮するため，α は正とする．

次に閾値を超えるまでの猶予日数を定義する．これを設定する理由は，急激に大幅な上昇をする場合と，緩やかに大幅な上昇をするときの両方を考慮するためである．ここで定数 d が与えられたときの t 日目の翌日から d 日目の時間窓を $[t+1, t+d]$ とするとき，時間窓内の日経 VI の最大値は以下となる．

$$v_t^d = \max\{v_{[t+1,t+d]}\} \tag{4}$$

制約条件として d は 1 以上の整数とする．例えば $d=1$ のときの時間窓は 1 営業日である．このとき時間窓 d における日経 VI の最大差分 m_t は，

$$m_t = v_t^d - v_t \tag{5}$$

と表せる．よって投資リスク r は閾値 α と，時間窓 d における日経 VI の最大差分 m_t が与えられたとき，以下のように示せる．

$$r = \begin{cases} 1(m_t \geq \alpha) \\ 0(m_t < \alpha) \end{cases} \tag{6}$$

このリスクの解釈は，今後の d 日間に日経 VI の 1 日差分が $i\sigma$ 以上上昇する日である．日経 VI の 1 日差分を図 1 に示す．

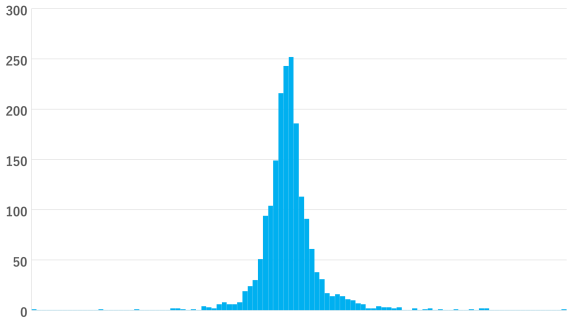


図 1: 日経 VI の 1 日差分

なお本研究では 1 週間が 5 営業日であることから，時間窓の定数 $d=5$ とする．さらに日経 VI の大幅な上昇と判定するための閾値は先行研究 [8] と同値にするため $i=2$ とする．2012 年 11 月 26 日から 2020 年 9 月 30 日までの 1915 日の中でこの定義に当てはまる日数は，287 日であり全体の 14.98%に該当する．

4. 日経 VI 上昇予測手法

4.1 概要

日経 VI 上昇予測モデルについての概要について図 2 に示す．最初にソーシャルメディアの投稿から話題を抽出するために，投稿文書に形態素解析を行い，形態素に分割し，言語モデルを用いて形態素に分割した文書から分散表現を獲得する．この分散表現の各次元を話題とする．次に獲

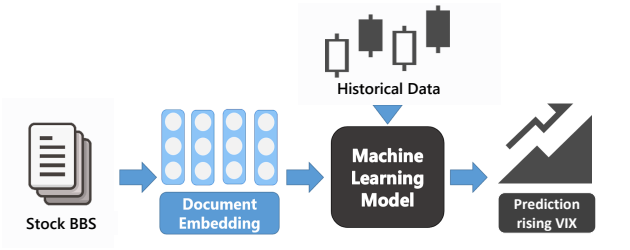


図 2: 日経 VI 上昇予測モデル概要

得した話題の分布を日別にまとめ，日次の話題ベクトルを獲得する．ここに金融時系列データを加えて特徴量を作成し，機械学習モデルの入力とする．最後に機械学習アルゴリズムを用いて日経 VI の上昇を予測する．このフローの詳細を次節から説明する．

4.2 ソーシャルメディアの投稿から話題抽出

最初にソーシャルメディアの投稿から話題を抽出するために，言語モデルを用いて文書の分散表現を獲得する．先行研究 [8] では LDA を用いて投稿を 100 種類のトピックに属する確率を導出している．LDA では単純に単語の出現頻度のみを参照しており，文書間の類似度や語順を考慮することで，より独立した話題の獲得ができる可能性がある．そのため本研究では先行研究 [8] で使用された LDA のほかに，文書間の意味の近さが表現され，短い文書に強い Doc2Vec[11] および文書の語順を考慮でき，長い文書にも有効な BERT[12] を用いる．

ニュース記事を対象とした研究では，ニュース記事の内容ではなく，見出しを用いて言語モデルの入力とすることが多い [13], [14]．これはニュースの内容は語彙数が多く，ノイズが発生してしまうことから避けられている．ソーシャルメディアの投稿においては，内容に荒らしのコメント等がありノイズが多く発生することが予想される．しかし既存の研究 [15] において，荒らしコメントが株価の予測に一定の効果を示すことが報告されているため，本研究では投稿の内容すべてを対象とする．

4.2.1 Doc2Vec による話題抽出

Doc2Vec は，単語から分散表現への変換手法である Word2Vec[16] を文書単位に拡張したものである．Word2Vec は，単語を任意の数の固定長分散表現に変換することで，単語同士の意味の近さや，単語同士の足し引きを可能としている．Doc2Vec は図 3 に示すように，文書自体も単語ベクトルと同じ空間のベクトルで表現されている．したがって，単語と文書の分散表現が同時に学習される．

Doc2Vec のモデルを構築する前の前処理としてソーシャルメディアの投稿から，URL，HTML，改行コードを除去し，半角ひらがなおよびカタカナを全角に変換する．さらに全角英数字を半角に変換し，形態素解析を行う．形態

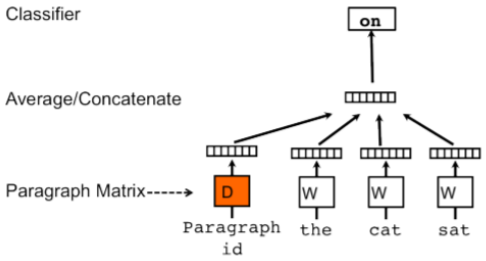


図 3: Doc2Vec のモデル構成図

素解析には Mecab[17] を使用し，辞書には Neologd[18] を使用する．ソーシャルメディアの分析では，流行りや話題の変化を察知することが重要である．よって毎週辞書を更新し，新規語や固有表現に強い Neologd を使用する．なお再現性の確保のため，2020 年 5 月 21 日までの更新で止めている．抽出する形態素は，名詞，動詞，形容詞でかつ Subtype が数値，非自立，代名詞，接尾を除くものとする．さらにストップワードの除去を行う．ストップワードのリストは，京都大学が公開している Slothlib のテキストデータに「ある，する，ちゃう，ない，なる，やる」を追加したものを使用する．形態素に分けられた投稿を用いて Doc2Vec で各投稿の分散表現を得る．Doc2Vec のベクトル数は任意の数を設定できる．この数により文章の分類の精度が変わる可能性がある．そのため複数のベクトル数を比較する．LDA についても Doc2vec と同様の前処理を行い，複数のベクトル数を比較する．

4.2.2 BERT による話題抽出

BERT モデルの構成図を図 4 に示す．BERT は，多層に重

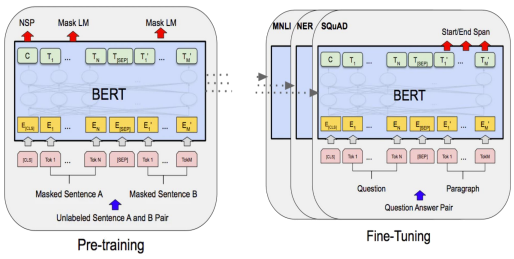


図 4: BERT のモデル構成図

ねた Transformer[19] を用いて文書から分散表現を獲得する双方向 Transformer モデルである．Transformer は，入力のベクトル列の重み付き和を計算する Self-attention 機構と，Self-attention の出力を別のベクトルへ変換する隠れ層からなる．Self-attention 機構が各単語に重みづけすることで，予測精度向上を期待する．BERT の学習は Pre-training(事前学習) と Fine-Tuning に分かれている．Pre-training では，教師なしの文書から Masked Language Model(MLM) と Next Sentence Prediction (NSP) の 2 つのタスクを解き，文書の分

散表現を獲得する．Fine-Tuning では，Pre-training で学習した文書の分散表現を初期値の重みとして，最終層に任意のタスクのラベルが付与された教師あり学習をする．東北大学 乾・鈴木研究室が公開している Pre-training 済みの日本語 BERT モデルを使う．

BERT モデルから分散表現を獲得するために，前節と同じように前処理を行った後，形態素解析を行う．抽出する形態素は語順も対象に入れるため，記号のみを除く．このベクトルサイズは 768 次元である．本研究では，各文書から得た分散表現を日別にまとめるため，Fine-Tuning を行わずに Pre-training 済みモデルから得た出力をそのまま用いる．言語モデルから獲得した分散表現を日経 VI 上昇予測の機械学習の入力に用いるための特徴量計算を次節で述べる．

4.3 機械学習による日経 VI 上昇予測

獲得した分散表現を日経 VI 上昇予測に用いるために，日別にまとめ日次の話題ベクトルを獲得する．日次予測の理由は既存研究において，週次や月次よりも良いパフォーマンスが得られるとされているためである [20], [21]．日次のまとめる方法は，既存研究 [8], [22] において，日別の各投稿の単純平均を用いているため，本研究においてもこれを適応する．さらに投稿数を加えて特徴量を生成する．ソーシャルメディアの情報を用いた特徴量は以下の通りである．

- 投稿数
- 投稿数の 1 日差分，5 日差分
- 投稿数の 1 日比，5 日比
- 投稿数の 2 日間移動平均，5 日間移動平均
- 投稿数と投稿数の 2 日間移動平均の差，5 日間移動平均の差
- 投稿数と投稿数の 2 日間移動平均の比，5 日間移動平均の比
- 話題ベクトル
- 話題ベクトルの 1 日差分，5 日差分
- 話題ベクトルの 1 日比，5 日比
- 話題ベクトルの 2 日間移動平均，5 日間移動平均
- 話題ベクトルと話題ベクトルの 2 日間移動平均の差，5 日間移動平均の差
- 話題ベクトルと話題ベクトルの 2 日間移動平均の比，5 日間移動平均の比

さらに日経平均，日経 VI の金融時系列データも特徴量に加える．金融時系列の情報を用いた特徴量は以下の通りである．

- 日経平均株価の始値
- 日経平均株価の始値の 1 日差分，5 日差分
- 日経平均株価の始値の 1 日比，5 日比
- 日経平均株価の始値の 2 日間移動平均，5 日間移動平均
- 日経平均株価の始値とその 2 日間移動平均の差，5 日間移動平均の差
- 日経平均株価の始値とその 2 日間移動平均の比，5 日間移動平均の比
- 日経 VI の始値
- 日経 VI の始値の 1 日差分，5 日差分
- 日経 VI の始値の 1 日比，5 日比
- 日経 VI の始値の 2 日間移動平均，5 日間移動平均
- 日経 VI の始値と日経 VI の始値の 2 日間移動平均の差，5 日間移動平均の差
- 日経 VI の始値と日経 VI の始値の 2 日間移動平均の比，5 日間移動平均の比

一般に日次ベースの金融商品の予測では，営業日のみを対象としている．これは，金融時系列のデータが営業日のみしかないためである．そのため休日に大きなイベントがあったときにこれを予測することは難しい．ソーシャルメディアやニュースは休日や営業日は関係なく投稿されているにもかかわらず，既存研究 [5], [8] において休日の情報は無視されている．休日の情報をうまく組み込むことができれば，予測精度の向上が期待できる．そこで休日明けの営業日を予測するとき，前日のソーシャルメディア情報と，前営業日の金融時系列の情報を入力として予測する．この他に休日の情報をを平均化する方法も考えられるが，前日の情報を重要視すること，平均化することによる情報の平滑化をしてしまうことを考慮して用いていない．

機械学習アルゴリズムはロジスティック回帰，ランダムフォレスト，LightGBM の 3 つを用いて比較を行う．なお正解ラベルは不均衡データであるため正例のデータが極端に少なく，モデルの学習が偏ってしまう可能性が考えられる．そこで学習データの中で負例を減らすダウンサンプリングを行い学習したモデルとダウンサンプリングを行わないモデルとを比較する．

表 1: 言語モデルと学習アルゴリズムの組み合わせ

	Logistic Regression			Random Forest			Light GBM		
ダウンサンプリング	なし	7:3	5:5	なし	7:3	5:5	なし	7:3	5:5
LDA	32	64	128	32	64	128	32	64	128
Doc2Vec	32	64	128	32	64	128	32	64	128
BERT	768			768			768		

5. 評価実験

5.1 実験設定

一般的な機械学習モデルの評価では，実験期間を 2 分割し，学習と評価の期間に分け検証するホールドアウト法や，分割期間の分布の差を考慮するために実験期間を k 分割し，1 つのサンプルを評価として，残りの (k-1) 個のサンプルを学習としてモデルを作成し，これを k 回繰り返したときの精度の平均で検証する k-fold クロスバリデーション法などが適応される．

本研究の対象は，時系列のデータであり直近のデータが非常に重要な役割を担う可能性がある．上記の評価方法では，学習期間と評価期間が離れてしまう箇所が存在する．そのため，毎回の評価期間は学習期間の直近の 1 回に固定し，これを繰り返すことで評価する．評価方法を図 5 に示す．実験期間 T 日のうち，最初の学習量を担保するため定

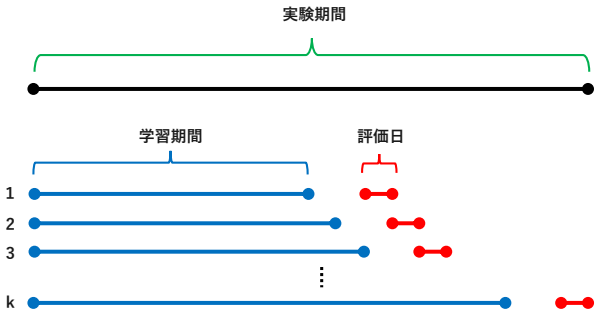


図 5: 評価方法

数 n を決め, 学習期間を 1 日目から n 日目の n 日間として学習する．評価には，n に 3 章で決定した時間窓 d を加えた $n+d$ 日目の 1 日を用いる．これを n を 1 日ずつ増やしてデータを追加し，k 回繰り返す．最後に $n+d+k$ が T になるまで繰り返した後，評価日の合計 k 回で評価する．ダウンサンプリングを行う場合，学習期間の n 日間のうち，日経 VI の上昇と判定されいるラベルが指定した任意の割合になるように，日経 VI の上昇と判定されていない日のデータをランダムに除く．こうしてできた学習データを用いて学習する．そのため，検証時の正解ラベルの比率は変化していないことに注意する．

本研究では，$n = 980$ とする．これは，1 年を 245 営業日としたときの 4 年分である．時間窓は $d = 5$ である．

ソーシャルメディアのデータは，ヤフーファイナンス掲示板の日経平均株価スレッドを使用した．ヤフーファイナンス掲示板は個別銘柄ごとにスレッドが分かれている．しかし本研究の予測対象である日経 VI は日経平均株価に関連した指数であるため，日経平均株価スレッドを対象とする．データ収集期間は，2012 年 11 月 26 日から 2020 年 09 月 30 日 (2,817 日) で，9,463,957 件だった．日経平均株価および日経 VI の時系列データは，JPX データクラウド [23] から収集した．金融時系列のデータ収集期間は 2012 年 11 月 26 日から 2020 年 10 月 7 日 (1920 営業日) まで収集した．ソーシャルメディアのデータと若干の誤差があるのは最後の 5 営業日を正解ラベルの計算に使うためであり，実験期間はさらに最初の 5 営業日を除いた 2012 年 12 月 3 日から 2020 年 9 月 30 日 (1910 営業日) で，そのうち正解とあてはまる日は 288 日あった．検証期間は，2016 年 12 月 14 日から 2020 年 9 月 30 日 (925 営業日) でそのうち正解とあてはまる日は 133 日で，検証期間の 14.37%にあたる．なお日経平均株価および日経 VI の値は始値を用いている．これは予測モデルが，その日の取引が始まる前に日経 VI の上昇を判断することを想定しているためである．

言語モデルと予測モデルの組み合わせを表 1 に示す．

LDA と Doc2Vec のベクトルサイズは，32，64，128 次元で，BERT は，Pre-training の隠れ層 768 次元で，それぞれモデルを作成する．LDA と Doc2Vec は gensim を用いて，BERT は huggingface の transformers を用いて実装する．ロジスティック回帰，ランダムフォレスト，LightGBM のパラメータ設定としてグリッドサーチを用いる．学習データを 2 分割しパラメータを決定した後，学習データ全体で再度学習を行う．なおダウンサンプリングをする場合，正例と負例の比率を 3:7 にして学習するモデルと 5:5 にして学習するモデルを作成する．

予測モデルの評価指標としては，Accuracy，Precision, Recall, F1-measure を用い，これらの値を比較することで行う．なお Precision, Recall, F1-measure に関して，正例に対する数値を比較する．

5.2 結果

ロジスティック回帰の結果を表 2 に示す. ロジスティック回帰において，LDA の 64 次元で学習時の正解ラベルの

正例負例の比率を 3:7 の割合にダウンサンプリングしたモデルが Precision と F1-measure が一番高く，Precision が 0.19，Recall が 0.44，F1-measure が 0.26 となった．BERT の 5:5 にダウンサンプリングしたモデルは Recall が一番高く，Precision が 0.15，Recall が 0.59，F1-measure が 0.24 を得た．Doc2Vec の 64 次元で 3:7 にダウンサンプリングしたモデルは，F1-measure が LDA の 64 次元で 7:3 にダウンサンプリングしたモデルと同様に一番高い結果となり，Precision が 0.17，Recall が 0.48，F1-measure が 0.26 を得た．

表 2: ロジスティック回帰の結果

特徴量	次元数	ダウンサンプリング	Precision	Recall	F1-measure
LDA	32	なし	0.16	0.43	0.23
		7:3	0.17	0.49	0.25
		5:5	0.15	0.51	0.23
	64	なし	0.15	0.25	0.19
		7:3	0.19	0.44	0.26
		5:5	0.16	0.53	0.25
	128	なし	0.16	0.35	0.22
		7:3	0.16	0.44	0.23
		5:5	0.15	0.58	0.24
Doc2Vec	32	なし	0.17	0.26	0.20
		7:3	0.16	0.36	0.22
		5:5	0.14	0.41	0.20
	64	なし	0.17	0.38	0.23
		7:3	0.17	0.48	0.26
		5:5	0.15	0.54	0.24
	128	なし	0.14	0.29	0.19
		7:3	0.16	0.37	0.22
		5:5	0.14	0.46	0.22
BERT	786	なし	0.13	0.29	0.17
		7:3	0.16	0.46	0.23
		5:5	0.15	0.59	0.24

ランダムフォレストの結果を表 3 に示す．ランダムフォレストにおいて，LDA の 64 次元で 3:7 にダウンサンプリングしたモデルは Precision が一番高く，Precision が 0.67，Recall が 0.02，F1-measure が 0.03 を得た．LDA の 128 次元で 5:5 にダウンサンプリングしたモデルは Recall が一番高く，Precision が 0.15，Recall が 0.60，F1-measure が 0.24 を得た．BERT で 5:5 にダウンサンプリングしたモデルは F1-measure が一番高く，Precision が 0.16，Recall が 0.58，F1-measure が 0.25 を得た．

LightGBM の結果を表 4 に示す．LightGBM において，Doc2Vec の 64 次元で 3:7 にダウンサンプリングしたモデルは，Precision が一番高く，Precision が 0.24，Recall が 0.28，F1-measure が 0.26 を得た．LDA の 128 次元で 5:5 にダウンサンプリングしたモデルは Recall が一番高く，Precision が 0.17，Recall が 0.55，F1-measure が 0.26 を得た．なお F1-measure は，上記の 2 つのモデルが一番高い結果となった．

表 3: ランダムフォレストの結果

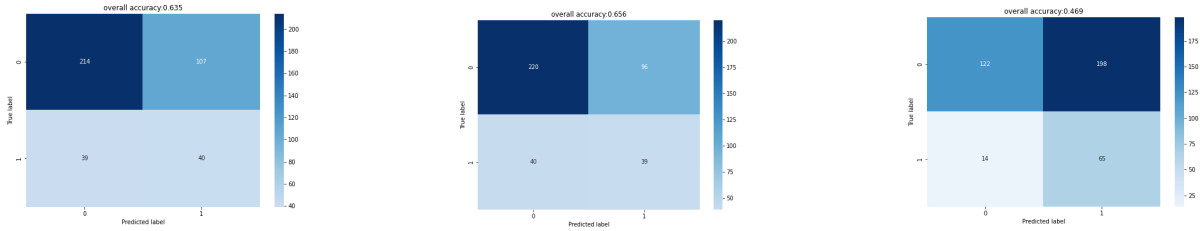
特徴量	次元数	ダウンサンプリング	Precision	Recall	F1-measure
LDA	32	なし	0.29	0.05	0.05
		7:3	0.37	0.05	0.09
		5:5	0.15	0.59	0.24
	64	なし	0.25	0.02	0.04
		7:3	0.67	0.02	0.03
		5:5	0.15	0.51	0.23
	128	なし	0.26	0.04	0.07
		7:3	0.31	0.07	0.11
		5:5	0.15	0.60	0.24
Doc2Vec	32	なし	0.33	0.08	0.12
		7:3	0.27	0.02	0.04
		5:5	0.14	0.44	0.22
	64	なし	0.41	0.05	0.09
		7:3	0.27	0.07	0.11
		5:5	0.16	0.51	0.24
	128	なし	0.30	0.05	0.08
		7:3	0.21	0.02	0.04
		5:5	0.21	0.02	0.04
BERT	786	なし	0.18	0.02	0.03
		7:3	0.36	0.13	0.19
		5:5	0.16	0.58	0.25

表 4: LightGBM の結果

特徴量	次元数	ダウンサンプリング	Precision	Recall	F1-measure
LDA	32	なし	0.17	0.05	0.07
		7:3	0.17	0.22	0.19
		5:5	0.14	0.50	0.22
	64	なし	0.22	0.05	0.07
		7:3	0.20	0.18	0.19
		5:5	0.14	0.47	0.22
	128	なし	0.20	0.12	0.15
		7:3	0.15	0.23	0.18
		5:5	0.17	0.55	0.26
Doc2Vec	32	なし	0.10	0.05	0.06
		7:3	0.16	0.14	0.15
		5:5	0.14	0.40	0.21
	64	なし	0.12	0.06	0.08
		7:3	0.24	0.28	0.26
		5:5	0.16	0.52	0.24
	128	なし	0.12	0.52	0.24
		7:3	0.12	0.12	0.12
		5:5	0.17	0.14	0.15
BERT	786	なし	0.12	0.03	0.05
		7:3	0.20	0.23	0.21
		5:5	0.15	0.50	0.23

5.3 考察

言語モデルの違いで精度の向上は見られなかった．一方で，ダウンサンプリングをすることによりどの言語モデルを使った場合でも精度の向上がみられた．これによりダウンサンプリングが有効な可能性がある．しかし最大の F1-measure が 0.26 と先行研究の検証期間の精度より低い結果となった．これは検証期間においてコロナショックの期間が含まれることが原因の可能性がある．そこで結果のうち一番 F1-measure が高い，ロジスティック回帰の LDA の 64 次元で 3:7 にダウンサンプリングしたモデル (LR-LDA-64-3)，LightGBM の Doc2Vec の 64 次元で 3:7 にダウンサンプリングしたモデル (LG-D2V-64-3)，LightGBM



LR-LDA-64-3　　LG-D2V-64-3　　LG-LDA-128-5

図 6: 先行研究の検証期間での混同行列

の LDA の 128 次元で 5:5 にダウンサンプリングしたモデル (LG-LDA-128-5) の 3 つを用いて検証期間を諏訪ら [8] と合わせ，2014 年 11 月 17 日から 2016 年 6 月 29 日の 395 営業日で評価を行った．この時の結果を表 5 に，混同行列を図 6 に示す．結果としては，F1-measure が最大 0.40 を得て，Recall は先行研究よりも上昇している．このことから下降トレンドにおいての精度改善は必須である．

表 5: 先行研究の検証期間での結果

	Precision	Recall	F1-measure
LR-LDA-64-3	0.28	0.51	0.36
LG-D2V-64-3	0.29	0.49	0.36
LG-LDA-128-5	0.30	0.62	0.40

6. おわりに

本研究では個人投資家の投資リスクを低減するために日経 VI に注目し，この指数の大幅な上昇を予測する手法を提案した．ヤフーファイナンス掲示板の投稿から言語モデルにより分散表現を獲得し，日別にまとめることにより日次の話題ベクトルを生成した．この話題の分布の違いと金融時系列データを機械学習により学習することによって予測を行った．結果として 2016 年 12 月 14 日から 2020 年 9 月 30 日 (925 営業日) の期間で F1-measure が最大で 0.26 を得た．先行研究の期間と合わせた検証期間で比較すると，F1-measure が最大で 0.40 を得ていることや，先行研究と比べ Recall が高くなった．この長期の検証期間の中にはコロナショック時の影響も含まれている．このことから下降トレンドでの予測が低い可能性が考えられる．そのため下降トレンド時の精度向上は今後の課題とする．さらに今後の展望として，下降トレンド時において予測モデルの出力に従って売買行動をとった時の利益やリスク回避を売買シミュレーション [9] において確認することが考えられる．

謝辞

本研究の一部は科研費基盤 C(20K01863) による助成を受けて行われた．

参考文献

[1] Michaël Karpe. An overall view of key problems in algorithmic trading and recent progress. *arXiv preprint arXiv:2006.05515*, 2020.

[2] Zhige Li, Derek Yang, Li Zhao, Jiang Bian, Tao Qin, and Tie-Yan Liu. Individualized indicator for all: Stock-wise technical indicator optimization with stock embedding. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD'19. Association for Computing Machinery, 2019.

[3] Chi Chen, Li Zhao, Jiang Bian, Chunxiao Xing, and Tie-Yan Liu. Investment behaviors can tell what inside: Exploring stock intrinsic properties for stock trend prediction. KDD'19. Association for Computing Machinery, 2019.

[4] 伊藤克哉, 南賢太郎, 今城健太郎, 中川慧. Trader-company 法: メタヒューリスティクスを用いた株価予測金融機関を模したモデルによる時系列予測. 人工知能学会全国大会論文集 第 34 回全国大会. 一般社団法人人工知能学会, 2020.

[5] 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. Association for Computational Linguistics, July 2020.

[6] 今城健太郎, 南健太郎, 伊藤克哉, 中川慧. 株価の残差リターンに注目した深層学習ポートフォリオ最適化. 人工知能学会全国大会論文集, 2020.

[7] Shan Jiang, Hsinchun Chen, Jay F Nunamaker, and David Zimbra. Analyzing firm-specific social media and market: A stakeholder-based event analysis framework. *Decision Support Systems*, No. 67, pp. 30–39, 2014.

[8] H. Suwa, Y. Ogawa, E. Umehara, K. Kakigi, T. Yamashita, and K Tsubouchi. Develop method to predict the increase in the nikkei vi index. In *Proceedings of The 2nd International Workshop on Application of BigData for Computational Social Science in IEEE Bigdata 2017*, 2017.

[9] Kodai Sasaki, Hirohiko Suwa, Yuki Ogawa, Eiichi Umehara, Tatsuo Yamashita, and Kota Tsubouchi. Evaluation of vi index forecasting model by machine learning for yahoo! stock bbs using volatility trading simulation. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

[10] 柴田舞. 日経平均ボラティリティ・インデックスの現物と先物の関係ー期待仮説による実証分析. 先物・オプションレポート 2019 年 2 月号, 2019.

[11] Andrew M. Dai, Christopher Olah, and Quoc V. Le. Document embedding with paragraph vectors. In *NIPS Deep Learning Workshop*, 2014.

[12] 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, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, June 2019.

[13] 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. 人工知能学会全国大会論文集 第 29 回全国大会 (2015). 一般社団法人人工知能学会, 2015.

[14] Yujie Wang, Hui Liu, Qiang Guo, Shenxiang Xie, and Xiaofeng Zhang. Stock volatility prediction by hybrid neural network. *IEEE Access*, Vol. 7, pp. 154524–154534, 2019.

[15] 山下達雄, 坪内孝太. 株価掲示板情報における煽り情報の検出. 人工知能学会全国大会論文集 第 29 回全国大会. 一般社団法人人工知能学会, 2015.

[16] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751, 2013.

[17] Taku Kudo. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. jp*, 2006.

[18] 佐藤敏紀, 橋本泰一, 奥村学. 単語分かち書き辞書 mecab-ipadic-neologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会 (NLP2017). 言語処理学会, 2017.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, Vol. 30, pp. 5998–6008, 2017.

[20] Paul C Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, Vol. 63, No. 3, pp. 1437–1467, 2008.

[21] Boyi Xie, Rebecca J. Passonneau, Leon Wu, and Germán G. Creamer. Semantic frames to predict stock price movement. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 873–883, 2013.

[22] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[23] Jpx data cloud. available from`http://db-ec.jpx.co.jp/item/C430509.html`. Tokyo Stock Exchange, Inc.(2019/4/28).