修士論文

金融市場予測のための Web 金融掲示板投稿文書分散表現の獲得手法の提案と評価

上田　健太郎
奈良先端科学技術大学院大学
先端科学技術研究科
情報理工学プログラム

主指導教員: 安本 慶一
ユビキタスコンピューティングシステム 研究室（情報科学領域）

令和 5 年 3 月 17 日提出

本論文は奈良先端科学技術大学院大学先端科学技術研究科に
修士 (工学) 授与の要件として提出した修士論文である。

上田　健太郎

審査委員：

主査　安本 慶一　（情報科学領域 教授）

荒牧　英治　（情報科学領域 教授）

諏訪 博彦　（情報科学領域 准教授）

金融市場予測のための Web 金融掲示板投稿文書分散表現の獲得手法の提案と評価*

上田　健太郎

内容梗概

金融市場の予測は，投資における獲得利益の最大化やリスク低減につながるために，僅かな予測精度向上でさえも非常に重要である．本研究では日本の市場を表す指数の 1 つである日経平均ボラティリティーインデックス（以下日経平均 VI と呼ぶ）の将来の大幅な上昇を高精度に予測することを目指す．日経平均 VI は恐怖指数とも呼ばれ，日経平均 VI が上昇すると日経平均株価が下落することが経験的に知られており，日経平均 VI の上昇を予測することができれば投資におけるリスクを低減させることが可能となる．近年の計算機と自然言語処理技術の急速な発展により，市場予測に過去の金融時系列データだけでなく，テキストデータを用いる研究が多く行われている．テキストデータの中でもソーシャルメディアの投稿文書は毎日多量の文書が記録される上，市場のイベントや投資家心理を含むために，市場予測に有効な可能性が高い．

しかし，市場予測にソーシャルメディアを扱う従来研究の多くは，ソーシャルメディアの感情情報のみに着目しているため，本来テキストに含まれているはずの社会のイベント情報などに影響を受けた投稿者の話題の情報を扱えていない．先行研究の中には感情情報だけでなく話題情報なども同時に扱う試みが存在するが，それらは bag-of-words ベースのトピックモデリング技術を用いているため，文書のコンテキストを一部破棄してしまう問題がある．

本研究では，これらの問題を解決する，ソーシャルメディア投稿文書から金融市場予測に有効な分散表現を獲得する手法 SSCDV を開発した．本手法は文書の

*奈良先端科学技術大学院大学 先端科学技術研究科 修士論文, 令和 5 年 3 月 17 日.

コンテキストを可能な限り保持しつつ，投稿の話題情報と感情情報の両方を共同で利用する．獲得した分散表現と金融時系列データを機械学習モデルに共同で学習させ，市場予測タスクを行うことで，提案アプローチの有効性を評価した．様々な分散表現獲得技術をベースラインとし，予測精度を比較した結果，提案するアプローチを用いたモデルが最も高い，Precision，F-1 Score，Matthews Correlation Coefficient（MCC）を達成した．MCC とは 2 値分類にタスクにおいて 2 クラスが不均衡であっても使用できる評価指標である．提案モデルは VI の大幅な上昇を予測するが，この予測結果はロングストラドル戦略を通じて利益を生み出す可能性がある．そこで，本モデルの予測結果をもとに，投資シミュレーションを行った．その結果，SSCDV による分散表現を学習した VI 予測モデルは，文書を用いずに金融時系列データのみで学習させた VI 予測モデルと比べ，+10,419,430 円の累積利益を獲得した．提案手法により文書から獲得された分散表現は，話題と感情に関する解釈性を有する．そこで，機械学習モデルの解釈手法の一つである SHAP を用い，金融市場に有効なソーシャルメディアの話題と感情について時系列的な解釈行った．その結果，ソーシャルメディアの話題と感情には，どの期間にも一貫して予測に影響を与えるものと，ある期間において予測に影響を与えるものが存在することが確認された．

キーワード

金融市場予測，ソーシャルメディア，自然言語処理

Proposal and Evaluation of a Method to Obtain Distributed Representations of Documents Posted on Web Financial Discussion Boards for Financial Market Forecasting*

Kentaro Ueda

Abstract

Prediction of financial markets is very important to maximize profit and reduce the risk of investment, so even a small improvement in forecasting accuracy is very important. In this study, we aim to predict with high accuracy future large increases in the Nikkei Stock Average Volatility Index (hereinafter referred to as Nikkei 225 VI), one of the indices representing the Japanese market. The Nikkei 225 VI is also called a fear index, and it is known empirically that the Nikkei Stock Average falls when the Nikkei 225 VI rises. With the rapid development of computer and natural language processing technology in recent years, many studies have been conducted using not only historical financial time series data but also textual data for financial market forecasting. Among text data, social media postings have a high potential to be effective for market forecasting because of the large number of documents that are recorded daily and because they include market events and investor sentiment.

However, most of the previous studies on social media for market forecasting have focused only on the emotional information in social media, and thus have not dealt with the information on the topics of the posters influenced by social events, which should

*Master's Thesis, Graduate School of Science and Technology, Nara Institute of Science and Technology, March 17, 2022.

be included in the text. Some previous studies have attempted to handle not only sentiment information but also topic information at the same time, but since they use bag-of-words-based topic modeling techniques, they have the problem of discarding part of the context of the document.

In this thesis, we developed SSCDV, a method for acquiring distributed representations effective for financial market forecasting from social media posted documents, which solves these problems. The method preserves the document context as much as possible and jointly utilizes both topical and sentiment information of the postings. We evaluated the effectiveness of the proposed approach by jointly training a machine learning model on the acquired distributed representations and financial time series data to perform a market forecasting task. We compared the forecasting accuracy using various variance representation acquisition techniques as baselines and found that the model using the proposed approach achieved the highest Precision, F-1 Score, and Matthews Correlation Coefficient(MCC). MCC is a metric that can be used even when the two classes are unbalanced. In addition, the proposed model predicts a significant increase in VI, which could be profitable through a long-straddle strategy. Therefore, we conducted an investment simulation based on the forecasting results of this model. As a result, the VI forecasting model that learned the variance representation by SSCDV obtained a cumulative profit of +10,419,430 yen compared to the VI forecasting model trained only on financial time series data. The distributed representation obtained from the documents by the proposed method is interpretable in terms of topic and sentiment. Therefore, we used SHAP, one of the interpretation methods of machine learning models, to interpret time series of social media topics and sentiments effectively for the financial market. As a result, it was confirmed that there are two types of social media topics and sentiments: those that consistently affect forecasts in any period, and those that affect forecasts in a certain period.

Keywords:

Financial market forecasting, Social media, Natural language processing

目 次

図 目 次

表 目 次

1. 序論

投資家は，金融市場の投資リスクを評価し，投資への意思決定を行う．ボラティリティー・インデックス（VI）指標は，金融資産の一定期間の変動率を表し，市場の不安定性を測るための本質的な指標である．VI 指標の正確な予測は，投資におけるリスク低減のために重要である．しかしながら，市場は一般的な時系列予測とは異なり，ノイズが多く，非定常的であり，様々な外的要因を影響を受けて変動する [1]．そのため，市場の予測は，本質的に非常に困難な問題である．これまでの研究では，過去の金融時系列データを用いた予測技術の開発に多くの焦点が当てられてきた．しかし，過去の金融時系列データのみを用いた場合，ニュースや企業の公式発表，投資家のムード情報などが市場に与える影響を捉えられず，その予測には限界がある．そのため，正確な市場予測を行うためには，過去の金融時系列データの使用に加え，異種データソースの活用が非常に重要である．近年の機械学習技術や自然言語処理技術の発展は，学者や実務家に異種データソースを活用した市場予測の有効性を提示しており，特にテキスト資源の活用が注目を浴びている．テキスト資源の活用例として，例えば，財務報告書 [2, 3, 4]，ニュース [5, 6, 7, 8]，ソーシャルメディア [9, 10, 11, 12, 13] を用いた予測などが挙げられる．テキスト文書の中でも興味深い資源は，ソーシャルメディア投稿文書（マイクロブログ，Web 掲示板など）である．ここでは，個人投資家などの金融に関心のある不特定多数の人々が，意見やニュースイベントなどをリアルタイムで共有している．このようなソーシャルメディアの投稿文書には，金融時系列データのみでは補足できない，突発的なイベントや投資家たちの感情や意見に関するシグナルが含まれている可能性が高く，これらを有効活用することで市場予測精度の向上が期待できる．

ソーシャルメディアを金融市場予測に用いた既存の研究の多くは，投稿全体の感情情報のみを用いるアプローチである [9, 10]．ソーシャルメディアから抽出した社会の感情情報は，市場予測において一定の成果を収め，市場予測における感情情報の有用性を示している．しかし，投稿全体の感情情報のみ用いるアプローチの限界は，投稿の話題などのシグナルを逃してしまうことである．ソーシャルメディアの投稿者が，今，何の話題に興味を持って議論しているかを知ること

は，市場予測において重要なことの一つである．そのため，より予測に有効な分散表現を獲得すべく，Chen らや Nguyen らは，トピックとセンチメントの情報を活用する手法を提案している [14, 15]．Chen らは，収集した文書から感情辞書を用いて感情特徴表現を抽出し，LDA を用いてトピックベクトルを獲得している [14]．彼らの手法は，それらを単純に連結して機械学習の入力としているため，各トピックに対する感情情報が失われている．Nguyen らの提案した TSLDA は，Latent Dirichlet Allocation（LDA）[16] ベースのアプローチであり，文書のトピックとセンチメントを同時に推論する [15]．TSLDA は，トピックのみを記述する LDA と比べ，市場予測において高いパフォーマンスを発揮している．しかし，彼らの提案した TSLDA は，bag-of-words ベースのトピックモデリングにより単語出現頻度の統計的関係のみを記述しているため，文書の一部のコンテキストを破棄しているという問題が残っている．また一般に，文書の分散表現を獲得する手法としては，様々な自然言語処理のタスクにおいて高パフォーマンスを発揮した BERT[17] を用いることも重要な候補の一つとしてあげられる．しかし，このような事前学習モデルは，ソーシャルメディアが含む多量のノイズ（スペルミス，タイプミス，略語，スラング，インターネット専門用語など）のためにパフォーマンスが著しく低下することが知られており [18, 19]，ソーシャルメディア投稿文書に対する分散表現の獲得手法としては不向きな可能性が高い．従って，金融市場予測に有効なソーシャルメディア文書埋め込みでは，コンテキスト情報を可能な限り効果的に扱い，ノイズによるパフォーマンスの低下問題を克服しつつ，トピック情報と感情情報を豊富に含むことが課題である．

この課題を克服するために，本研究では，Mekara らの提案した文書分散表現，Sparse Composite Document Vectors(SCDV)[20] を，感情極性辞書から参照された感情値を用いて拡張した，SSCDV と呼ばれる分散表現獲得手法を提案する．拡張元となる SCDV は静的埋め込みに対して Gaussian Mixture Models(GMM) ソフトクラスタリング [21] を適用することで，文書のトピック情報を有効に扱うことのできる文書分散表現である．LDA のようなトピックモデルで得られる確率分布は通常，単語やトピックや文書に埋め込まれた実際の意味情報よりも、単語出現頻度の統計的関係を記述することを優先するため，一部のコンテキスト情報が破棄

される．一方，SCDV に利用される Skip-Gram with Negative Sampling(SGNS)[22] のアプローチでは，Window-Size に依存した情報を利用して単語の本質的な意味情報を埋め込むことができる．さらに，SCDV アルゴリズムでは単語の与える文章ベクトルへの影響が文脈によって異なるため，ある程度多義語への対応も可能となっている．上記の理由より，SCDV を拡張することで開発される提案埋め込み手法 SSCDV は，既存研究で提案されてきた LDA を拡張して開発する埋め込み手法よりも有効であることが期待される．

提案する SSCDV は，次のように構成される．まず SCDV と同様に GMM ソフトクラスタリングにより，単語ベクトルに対するトピック空間を学習させる（4.2 節）．次に，辞書の極性値を用いることで，文書内の各単語についてポジティブクラスとネガティブクラスへの所属値を算出する（4.3 節）．次に，算出した所属値と単語埋め込みを用いてセンチメントベクトルを獲得し（4.4 節），GMM ソフトクラスタリングにより学習された各クラスターにおける確率分布と逆文書頻度で重み付けすることでセンチメントトピックベクトルを得る（4.5 節）．文書内単語のセンチメントトピックベクトルの和によって文書は表現される．ただし，この時ルールベースの手法により否定されたと判断されたトークンに関してはセンチメントの重みを反転させたセンチメントトピックベクトルを使用する（Algorithm: 1）．最終的な文書ベクトルは，計算コストと複雑さの低減のためにスパース化される（4.6 節）．我々の提案する SSCDV は，SCDV と類似しているが，事前辞書を用いた感情クラスを適用するという重要な変更が加えられており，文書の豊富なセマンティクスをより効果的に埋め込むことを可能とする．

我々は，SSCDV の有効性を評価するため，日本最大級の株式に関するソーシャルメディアである Yahoo!ファイナンス掲示板から SSCDV により，文書分散表現を獲得した．その後，獲得された文書分散表現と過去の金融時系列データを機械学習アルゴリズムにより共同で学習し，幅広いテスト期間の中，VI の大幅な上昇の予測を行なった．さらに，VI の上昇予測を用いた投資シミュレーションを行い，予測結果の実際の投資における有効性の評価を行なった．既存の言語モデルにより獲得した分散表現を用いて同様の予測をベースラインとして実験し，提案手法の精度との比較を行なった．実験の結果，予測精度，投資シミュレーションの両

観点から提案手法の有効性を明らかにした．文書から獲得された分散表現はしばしば解釈性を失い，予測に対するブラックボックス化を招く．しかし，SSCDV により得られる埋め込みはトピックと感情に関する優れた解釈性を提供する．そのメリットを生かし，機械学習の解釈手法として知られる SHAP[23] を用いて，予測モデルの予測と特徴量の関係解釈を行なった．その結果，ソーシャルメディアの話題と感情には，どの期間にも一貫して予測に影響を与えるものと，ある期間において予測に影響を与えるものが存在することが確認された．我々の貢献は以下の通りである．

Algorithm: SCDV を拡張し，トピックと感情を考慮した埋めこみを獲得する SSCDV を提案した．

Accuracy: 機械学習アルゴリズムを用いた VI 予測実験を行なった．ベースラインモデルと比較した結果，提案手法により獲得された埋め込みを特徴量として用いた機械学習モデルが最も高い予測精度となった．

Simulation: 予測結果を活用した投資予測シミュレーションを行なった結果，SSCDV を用いたモデルは価格データのみを用いたモデルと比べ，+10,419,430 円の利益をあげることを示した．

Interpretation of the forecast: SHAP 値による特徴量とモデルの予測に関わる関係の解釈によりソーシャルメディア（web 金融掲示板）のコメントと金融市場の動きに関する重要な知見を獲得した．

以降の章構成は以下の通りである．2 章では，提案手法に関連した既存研究の概要と課題について述べ，本研究の立ち位置を示す．3 章では，本研究において解決したいソーシャルメディアから金融市場予測タスクに有効な分散表現の獲得問題を示し，4 章では 3 章で示した問題を解決するための手法を提案する．5 章では，提案手法の有効性を示すための実験設定を述べる．6 章では実験によって得られた結果を示し．7 章では考察を述べる．最後に 8 章で，本研究についてのまとめを示す．

2. 関連研究

金融市場予測に関する研究は，予測に用いるデータの観点から，大きくの 2 つの方法に分けられる．一つは過去の金融時系列データのみを用いるテクニカル分析型の研究であり，もう一つは金融時系列以外のデータも用いるイベントドリブン型の研究である．本章では，初めにテクニカル分析型の研究について述べ，その後イベントドリブン型の研究でテキストデータを用いた研究について述べる．最後に，本研究において最も関係の深いソーシャルメディアテキストを用いた研究とその課題について述べ，本研究の立ち位置を明確化する．

2.1 金融時系列データを用いた金融市場予測に関する研究

金融市場予測精度向上は獲得利益の最大化につながり，その性能の僅かな向上さえも魅力的である．そのような潜在的な利益のために市場予測は長年，投資家や研究者から関心を集めてきた．深層学習ベースのモデルは市場の非線形的な動きを効果的に捉えることができると考えられており，近年の株価予測に対するアプローチの多くは深層学習ベースの手法である [24, 25, 26]．特に，RNN，LSTM のようなリカレントニューラルネットワークは株価の時系列的な変化パターンを捉えるための重要な機構となっており，市場予測のための自然な選択肢となっている．Bao らは，ウェーブレット変換，スタックドオートエンコーダ (SAE)，LSTM を組み合わせ金融時系列予測のための深層学習フレームワーク WSAEs-LSTM を提案している [25]．入力として過去の株価，テクニカル指標，マクロ経済指標を用いて，翌日の終値を予測するタスクを行い，RNN,WLSTM（ウェーブレット変換+LSTM）,LSTM と比較した結果，予測精度，収益性ともに他の手法よりも優れていると報告している．

リカレントニューラルネットワークの主な欠点は，金融時系列における数ヶ月単位の比較的長期的な依存性を捉える能力が低い点である．そこで，Ding らは，Transformer ベースのモデルを提案し，金融時系列予測を試みた [26]．2 つの実世界の市場データセットを用いた実験の結果，RNN や CNN ベースのアプローチと比較して、提案手法は性能面で大きな優位性を持っていると報告した．その他，

敵対的訓練 [27]，カプセルネットワーク [28]，グラフニューラルネットワーク [29] などを用いた様々なアプローチが開発され，過去の金融時系列データの複雑なパターン認識に関する研究が多く行われている．

これらの従来研究では，過去の株価データ，テクニカル指標，マクロ経済指標のような数値データのみを用いて予測を行なっている．これらの手法の限界は予測が過去の数値データのみに依存しているために，突発的な市場のイベントや投資家の感情などの市場に影響を及ぼす他の要因を考慮できない点である．そこで，高精度な市場予測を目的として市場に影響を及ぼす数値以外のデータを用いる研究が行われてきた．特に近年の計算機性能と自然言語処理技術の向上により，社会に記録されたテキストデータを用いるアプローチが着目されている．

2.2 テキストデータを用いた金融市場予測に関する研究

高精度な市場予測を達成するために，過去の価格データのみではなく異種データソースを活用した研究が盛んに行われている．異種データソースの中でも自然言語処理技術の発展により，財務報告書 [30] や金融ニュース [7] のようなテキストデータが，特に着目されている．Rekabsaz らは，財務報告書に対して bag-of-words ベースのセンチメントアプローチによるボラティリティー予測実験を行い，長期的なボラティリティー予測におけるセンチメント分析の有効性を実証している [30]．Duan らは，累積異常リターン予測のための金融ニュースに特化した文書分散表現を提案している [7]．彼らのモデルは累積異常リターン予測タスクにおいていくつかの代替的な最先端の文書表現よりも良い結果を与えることを報告している．このことは，市場予測に特化した文書分散表現技術の有効性を示唆している．

財務報告書や金融ニュースは信頼性が高くノイズが少ない一方で，直接的な投資家の意見や気持ちを反映しているとは言えない．そこで我々は，ソーシャルメディアに着目する．ソーシャルメディアには，ノイズが多いものの，市場に影響を与えるイベントに関する情報や，それを受けた投資家のムード情報も含まれており，市場予測に有効な豊富なシグナルを含んでいる．ソーシャルメディア文書を市場予測のためにより有効に活用するためには，このシグナルを効果的に抽出

する手法の開発が求められる．

2.3 ソーシャルメディアを用いた金融市場予測に関する研究

ソーシャルメディアテキストを用いた多くの研究は投稿全体の感情情報の活用に注力している [9, 10]．彼らは投稿から抽出した感情情報に一定の価値を示した．しかし，文書全体の感情情報を利用したアプローチは，話題のようなセンチメント情報以外の情報を破棄しているという欠点がある．投稿者らがどの話題に対してどの感情を持っているのかという情報は市場予測に価値のある可能性が高い．Si らは Twitter データを活用した市場予測のためにトピックベースのセンチメント時系列アプローチを提案している [31]．実験の結果，彼らはトピックを考慮したセンチメント時系列情報が，トピックを考慮しない単純な意見レキシコンベースの時系列情報よりも株価の上下動予測の性能向上に有効であると結論づけた．我々の知る限り，Nguyen らの研究がトピックとセンチメントを同時に推論するモデルを市場予測に用いた最初の研究である [15]．彼らはトピックとそのセンチメントを同時に推論する LDA ベースのモデル Joint sentiment/topic mdoel（JST）を用いて市場予測を行なった．また，意見語を区別し，トピックとそのセンチメントを同時に推論するモデル TSLDA を提案し，予測に用いた．JST と TSLDA の主な違いは JST ではトピックワードと意見ワードを区別しないが，TSLDA は SentiWordNet を用いたルールベースによりこれらを区別している点である．言い換えると，JST は各単語がトピックとセンチメントの共同分布から生成されると仮定しているが，TSLDA ではそれぞれのトピックに対する異なるセンチメントカテゴリごとに異なる意見語の分布を生成する．5 銘柄に対して株価の上下動の予測を行なった結果，LDA を用いたモデルと比べ，JST では平均して約 0.004 ポイント，TSLDA では約 0.06 ポイント精度が高いことを報告している．これらの研究で用いられたモデルは，いずれも bag-of-words ベースのトピックモデリングにより単語出現頻度の統計的関係のみを記述しているため，文書の一部のコンテキストを破棄しているという問題がある．

この問題の解決のために，本研究では Mekala らの提案した文書分散表現 SCDV[20] を拡張し，トピックと感情極性辞書ベースの感情値を埋め込みに共同で反映させ

るモデルを提案する．SCDV では Skip-Gram with Negative Sampling（SGNS）で学習された埋め込みに対し，GMM ソフトクラスタリングを行うことでトピック分布を生成する．SGNS では隣接するトークンを考慮してトークンのコンテキストをベクトル化するため，LDA などで用いられる bag-of-words 手法の制限であるトークン順序の完全な無視を克服することが可能である．また，SCDV アルゴリズムにより獲得される文書ベクトルは多義語にも対応することができる．そのため，SCDV ベースでトピックとセンチメントを考慮した文書表現はトピックモデルベースでトピックとセンチメントを考慮した文書表現よりも有効な文書の表現が獲得され，金融市場予測タスクの精度向上につながると考えられる．

2.4 本研究の立ち位置

高精度な市場予測を達成するために，本研究では過去の金融時系列データの他に，ソーシャルメディアへの投稿文書を用いて機械学習モデルによる予測を行う．これにより，過去の金融時系列のみでは補足できない情報を予測に取り込むことを目指す．本研究では，予測に用いるソーシャルメディアとして「Yahoo!ファイナンス掲示板」を選択する．Yahoo!ファイナンス掲示板のユーザは株式に興味を持っている可能性が高く，その書き込みには，Twitter のような他のソーシャルメディアへの書き込みよりも，予測に有効な情報が含まれていると考えられる．

機械学習モデルへの入力のために，金融市場予測に特化したソーシャルメディア投稿文書埋め込み技術の開発を行う．従来研究より，金融市場予測タスクの場合，ソーシャルメディア文書に対しては投稿のセンチメント情報を用いることが有効であることが多く報告されている．さらに，トピックとセンチメントを共同で扱うことで更なる精度向上の可能性がいくつか報告されている．そこで我々はトピックとセンチメントの情報を共同で埋め込むことのできる新たな手法 SSCDV を提案する．提案手法 SSCDV は SCDV アルゴリズムを拡張することで開発され，事前に定義された感情辞書を用いて感情を明示的に埋め込むことを行う．トピックと感情を共同で扱う従来研究のアプローチは LDA ベースのアプローチであり，文書のコンテキストを一部破棄する恐れがある．我々の SSCDV は SGNS で学習された単語埋め込みベースでトピックを形成し，各単語ごとのトピック分布ごとの

重みつき和で最終的な文章埋め込みを生成する．そのため，SSCDV は LDA ベースの手法に比べ文書のコンテキストを可能な限り保持することができると考える．金融市場予測で用いられた既存の埋め込み手法と提案手法との比較を，感情値と話題とコンテキストの活用の観点から整理し，表 1 に示す．

本研究では市場予測タスクとして，日経平均 VI の大幅な上昇予測タスクを設定する．市場予測研究においては，n 日後の金融指数の値の予測タスク [32]，n 日後の金融指数の騰落方向の 2 値分類タスク [9, 10]，n 日後の金融指数の値が現在よりも「大きく上昇」「変化が少ない」「大きく減少」の分類タスク [33] などが，一般的に用いられるタスク設定である．本研究で，我々が設定するタスクは，n 日後の金融指数の値が現在よりも「大きく上昇」「変化が少ない」「大きく減少」の分類タスクに最も関係しており，より具体的には将来 5 日営業日以内に日経平均 VI が閾値以上上昇するか，上昇しないかの 2 値分類タスクを設定する．VI は恐怖指数とも呼ばれ，日経平均 VI が大きく上昇した際には日経平均株価が大きく下落することが経験的に知られている．そのため，将来の日経平均 VI 上昇の高精度な予測は，投資におけるリスク低減のために重要となる．

表 1 既存手法と提案手法の比較

手法	明示的な感情極性値の利用	話題の埋め込み	文脈の活用	既存研究
辞書ベース	+	-	-	[12, 13, 34]
bag-of-words	-	+	-	[6, 8, 4]
lda	-	+	-	[31, 35]
JST/TSLDA	+	+	-	[15]
word2vec/glove ベース	-	+	+	[36, 37]
Doc2Vec	-	+	+	[38, 39]
CNN/RNN ファミリー	-	+	++	[40]
BERT	-	+	++	[41, 42]
SCDV	-	+	+	[43]
SSCDV（提案手法）	+	+	+	

3. 問題設定

本章で述べる問題設定の概略図を図 1 に示す．過去の 2 種の金融時系列データ（図 1 A）と Yahoo!ファイナンス掲示板の日経平均株価スレッドへの投稿文書（図 1 B）を用いて VI の大幅な上昇を予測する問題を考える．2 種類の金融時系列データとは日経平均株価の始値と日経平均 VI の始値であり，それぞれ

$$\vec{Price} = \{Stock\_Open_1, Stock\_Open_2, ..., Stock\_Open_t, ...\} \tag{1}$$

$$\vec{VI} = \{VI\_Open_1, VI\_Open_2, ..., VI\_Open_t, ...\} \tag{2}$$

が観測できるとする．ただし，$Stock\_Open_t$ は市場が開いている日の t 日目における日経平均株価の始値であり，VI_Open_t とは市場が開いている日の t 日目における日経平均 VI の始値である（市場は土日祝は開いていないため金融時系列データが生成されない）．

また，Yahoo!ファイナンス掲示板の日経平均株価スレッドへの投稿文書

$$Doc = \{D_1, D_1, ..., D_k, ...\} \tag{3}$$

が観測できるとする．ただし D_k は k 日目の投稿文書である（投稿文書は市場が開いている開いていないに関わらず毎日存在する）. 投稿テキストデータは各種手法により分散表現に変換される（図 1 C）．k 日目の投稿文書から獲得された分散表現を $D\vec{Vec}_k$ と表記する．本研究では，$t-1$ 日目と t 日目の金融時系列データ情報と，それに対応する k 日目の投稿文書情報を用いて特徴セットを作り（5.2.2 節），作成した特徴セットを入力とした機械学習により，テスト日の将来 5 営業日以内に日経平均 VI が大きく上昇するかどうかの 2 値分類タスクを行う（図 1 D）. t 日目における正解ラベル $y_t \in \{0,1\}$ は将来 5 営業日以内に上昇する場合 1，将来 5 営業日以内に上昇しない場合 0 とラベリングされる（5.2.3 節）．本研究の目的は D_k から市場予測モデルに有効な分散表現 $D\vec{Vec}_k$ を獲得するアルゴリズムを提案することであり，その有効性は以下の 2 つのリサーチクエスチョンへの回答を通して評価する（図 1 E）.

RQ1. Prediction Accracy (Sec. 6.1): 提案手法により獲得された分散表現を特徴量として用いた場合，VI 予測モデルの精度は向上するか？

RQ2. Investment Simulation (Sec. 6.2): VI 予測モデルの予測結果に基づいた投資シミュレーションにおいて，いくら利益が獲得できるか？

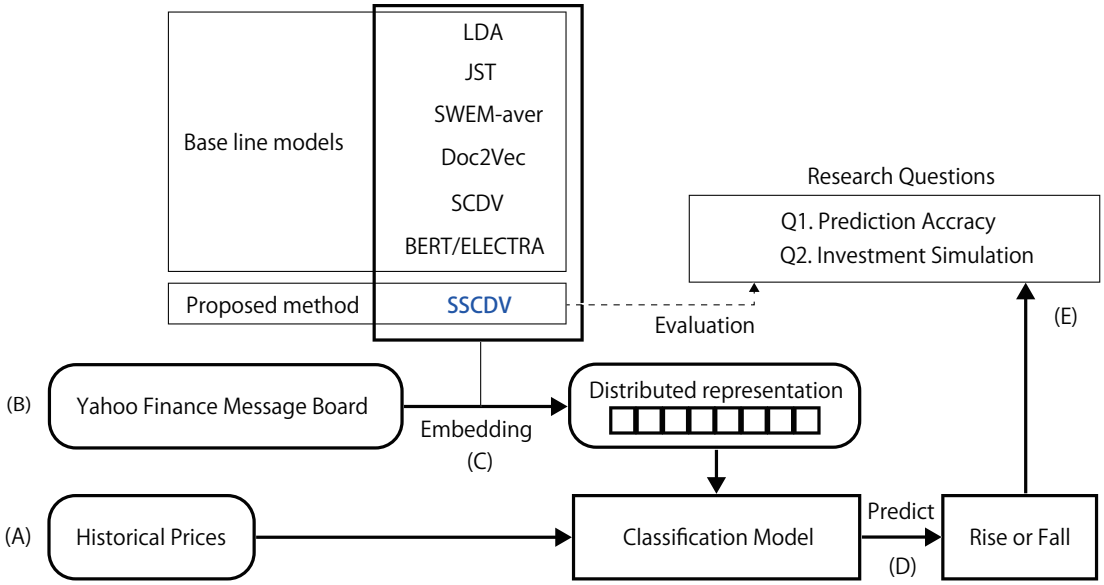

図 1 問題設定の概略図

4. ソーシャルメディア投稿文書からの金融市場予測に適した特徴抽出手法 SSCDV の提案

4.1 SSCDV の概要

前章の問題に対する解法を与えるため，Yahoo!ファイナンス掲示板データから文書分散表現を獲得する技術 SSCDV を提案する．提案手法の流れを図 2 に示す．SSCDV では感情辞書を用いることで，トピックと感情の混合クラスタに関するベクトルが生成されるため，SCDV では不可能であった，トピックと感情の情報の共同利用を可能としている．本章では，SSCDV の各構成要素について記す．

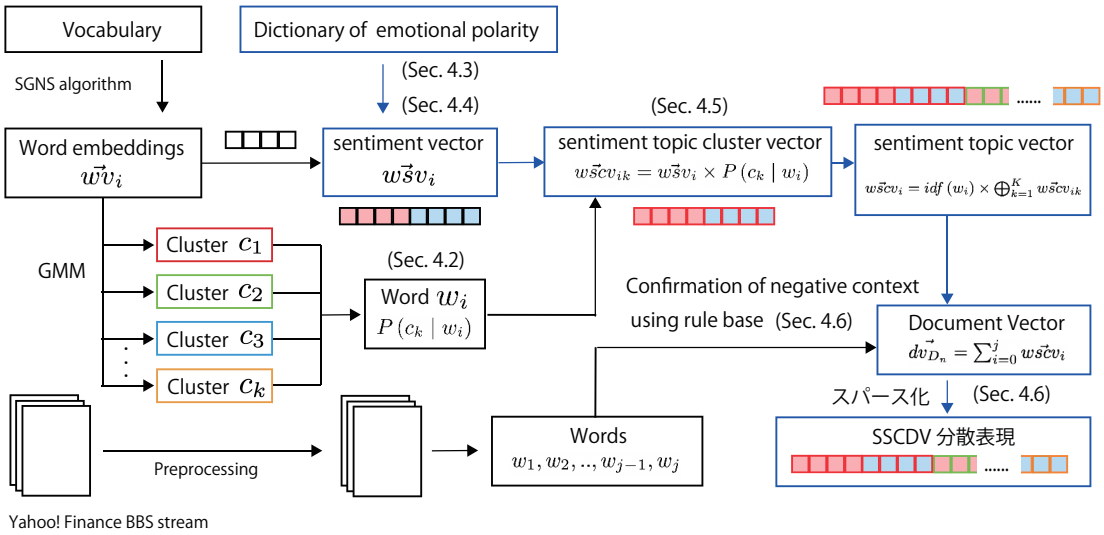

図 2 SSCDV による分散表現獲得のダイアグラム

4.2 単語ベクトルのクラスタリング

word2vec を SGNS アルゴリズムを用いて学習し，語彙 V の各単語について d 次元の埋め込みを獲得する．次に，1 つの単語が異なる意味を持つ多義語の特性を考慮するため，これらの単語埋め込みに対して GMM ソフトクラスタリングを用いたソフトクラスタリングを行う．結果として各単語の各クラスタへの所属確

率 $P(c_k \mid w_i)$ が得られる．この時，生成するクラスタ数 K は，SSCDV のハイパーパラメータである．

$$p(c_k = 1) = \pi_k \tag{4}$$

$$p(c_k = 1 \mid w) = \frac{\pi_k \mathcal{N}(w \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(w \mid \mu_j, \Sigma_j)} \tag{5}$$

4.3 単語ベクトルのセンチメントクラスへの所属確率の算出

各単語の極性を考慮するため，事前に構築された感情辞書から参照された極性値 s_i を用いて，各単語に対してポジティブクラスへの所属確率 $P(sc_{pos} \mid w_i)$ とネガティブクラスへの所属確率 $P(sc_{neg} \mid w_i)$ を以下のように算出する． 感情極性値は各単語に対し -1 から $+1$ までの実数値で割り当てられ，-1 に近いほどネガティブ，$+1$ に近いほどポジティブであることを示す．

$$P(sc_{pos} \mid w_i) = \tfrac{1+s_i}{2} \tag{6}$$

$$P(sc_{neg} \mid w_i) = 1 - P(sc_{pos} \mid w_i) \tag{7}$$

4.4 センチメントベクトルの生成

各単語の極性を明示的に文書分散表現に反映させるため，以下の処理により単語埋め込みからセンチメントベクトルを生成する．各単語の単語ベクトル $\vec{wv}_i$ に対し，ポジティブクラスへの所属確率 $P(sc_{pos} \mid w_i)$ とネガティブクラスへの所属確率（$P(sc_{neg} \mid w_i)$）で重み付けし，d 次元の 2 つのセンチメントクラスベクトル（$\vec{wsv}_{i(pos)}, \vec{wsv}_{i(neg)}$）を生成する．次に，生成した 2 つのセンチメントクラスベクトルを $2d$ 次元の埋め込みとして連結し，センチメントベクトル（$\vec{wsv}_i$）を得る．

$$\vec{wsv}_{i(pos)} = \vec{wv}_i \times P(sc_{pos} \mid w_i) \tag{8}$$

$$\vec{wsv}_{i(neg)} = \vec{wv}_i \times P(sc_{neg} \mid w_i) \tag{9}$$

$$\vec{wsv}_i = [\vec{wsv}_{i(pos)} \oplus \vec{wsv}_{i(neg)}] \in \mathbb{R}^{2d} \tag{10}$$

4.5 センチメントトピックベクトルの生成

極性が考慮された各単語に対し，所属トピックも考慮させるため，以下の処理によりセンチメントベクトルからセンチメントトピックベクトルを生成する．各センチメントベクトルに対し，K 番目のクラスタにおける確率分布 $P(c_k \mid w_i)$ で重み付けを行うことで K 個の異なるセンチメントトピッククラスタベクトル（$\vec{wscv}_{ik}$）を獲得する．次に生成した K 個のセンチメントトピッククラスタベクトルを $2dK$ 次元の埋め込みとして連結し，単語 w_i の逆文書頻度で重み付けを行うことでセンチメントトピックベクトル（$\vec{wscv}_i$）を獲得する．この結果，各単語に対応するベクトルは，その単語の感情極性とトピックと出現頻度で重み付けされたものとなる．

$$\vec{wscv}_{ik} = \vec{wsv}_i \times P(c_k \mid w_i) \tag{11}$$

$$\vec{wscv}_i = idf(w_i) \times \bigoplus_{k=1}^{K} \vec{wscv}_{ik} \in \mathbb{R}^{2dK} \tag{12}$$

4.6 センチメントトピックベクトルのスパース化

文脈を考慮したより正確な極性の反映と計算コストと複雑さの低減ために，ルールベースによる文脈判定とベクトルのスパース化の処理を経て最終的な文書分散表現を獲得する．はじめに．文書内の単語のセンチメントトピックベクトルの和をとることで文書ベクトル dv_{D_n} を得る．この時，より正確な極性を反映させるため，ルールベース（Algorithm: 1）により否定の文脈で現れたと判断されたトークンは，センチメントの重みを反転させたセンチメントトピックベクトルを使用した．SCDV 同様に，文書ベクトルの各成分について，その絶対値が閾値よりも小さいものを 0 にする．これにより文書ベクトルをスパース化した，Sentiment Sparse Composite Document Vectors（$\vec{SSCDV}_{D_n}$）を獲得する．閾値は SSCDV のパラメータとして設定され，具体的な閾値は以下の式により決定される．p はスパース化を行う際の閾値パラメータである．a_i はセンチメントトピックベクトルの i 番目の成分値であり，n はトレーニングデータセットの n 番目の文書を示す．また，Algorithm: 2 では SSCDV アルゴリズムを説明する．

$$a_i = \begin{cases} a_i \text{ if } |a_i| \geq \frac{p}{100} * t \\ 0 \text{ otherwise} \end{cases} \tag{13}$$

$$t = \frac{|a_{\min}| + |a_{\max}|}{2} \tag{14}$$

$$a_{\min} = \text{avg}_n \left(\min_i (a_i) \right) \tag{15}$$

$$a_{\max} = \text{avg}_n \left(\max_i (a_i) \right) \tag{16}$$

Algorithm 1 感情極性反転のための文脈判定

Data: 前処理済み（5.2.6 節: 前処理 B）投稿テキストの 1 文 $[w_1, w_2, ..., w_i, ...]$

Results: w_i が否定されているか，されていないかの判定

for *each word* w_i, w_{i+1} *in* $[w_1, w_2, ..., w_i, ...]$ **do**
  if i == 1 **then**
    Pass
  else if w_{i+1} の活用が「不変化型」かつ原型が「ん」 **then**
    w_i は否定後の文脈で使用と判定
  else if w_{i+1} の活用が「特殊・ナイ」かつ原型が「ない」 **then**
    w_i は否定後の文脈で使用と判定
  else if w_{i+1} の活用が「特殊・ヌ」かつ原型が「ぬ」 **then**
    w_i は否定後の文脈で使用と判定
  else
    w_i は通常の文脈で使用と判定
  end if
end for

Algorithm 2 SSCDV Algorithm による文書分散表現の獲得

Data: Documents $D_n, n = 1..N$

Results: Documents vectors $\vec{SSCDV}_{D_n}, n = 1..N$

Obtain word vector ($\vec{wv}_i$), for each word w_i;
Obtain sentiment polarities value s_i, for each word w_i;
Calculate probabilities of belonging to the sentiment class $P(sc_{pos} \mid w_i)$ and $P(sc_{neg} \mid w_i)$, for each word w_i;
Calculate idf values, $idf(w_i)$, $i = 1..|V|$; /* $|V|$ is vocabulary size */
Cluster word vectors $\vec{wv}$ using GMM clustering into K clusters;
Obtain soft assignment $P(c_k \mid w_i)$ for word w_i and cluster c_k;
for *each word* w_i *in vocabulary* V **do**
    $\vec{wsv}_i = [\vec{wsv}_{i(pos)} \oplus \vec{wsv}_{i(neg)}] \in \mathbb{R}^{2d}$;
    for *each cluster* c_k **do**
        $\vec{wscv}_{ik} = \vec{wsv}_i \times P(c_k \mid w_i)$;
    end for
    $\vec{wscv}_i = idf(w_i) \times \bigoplus_{k=1}^{K} \vec{wscv}_{ik}$; /* $\bigoplus$ is concatenation */
end for
for $n \in (1..N)$ **do**
    Initialize document vector $\vec{dv}_{D_n} = \vec{0}$;
    for *word* w_i *in* D_n **do**
        if Determine that w_i is a negative word on a rule basis **then**
            for *each sentiment topic vector* $\vec{wscv}_i$ *component* a_i **do**
                if $\mathrm{mod}(i, 2d) \in (1..d)$ **then**
                    $a_i = a_i \times \frac{P(sc_{neg}|w_i)}{P(sc_{pos}|w_i)}$
                else
                    $a_i = a_i \times \frac{P(sc_{pos}|w_i)}{P(sc_{neg}|w_i)}$
                end if
            end for
        end if
        $\vec{dv}_{D_n} += \vec{wscv}_i$;
    end for
    $\vec{SSCDV}_{D_n}$ =make-sparse($\vec{dv}_{D_n}$);
end for

5. 評価実験

提案手法の評価のために，実世界の市場データを用いた評価実験を行う．

5.1 実験目的

今回の評価実験では、提案手法により獲得された分散表現が機械学習モデルを用いた市場予測に効果的であるかどうかを評価するために 3 章で述べた 2 つのリサーチクエスチョンへの回答を行う．

5.2 実験方法

評価実験は，実世界の市場データと掲示板投稿文書を用いて学習した機械学習モデルの評価タスクにおける予測精度の算出，および予測結果を用いた実世界における取引シミュレーションによる累積獲得利益を算出することで実施する．

5.2.1 データセット

本研究では，金融時系列データとして，日経平均株価と日経平均 VI を用い，テキストデータとして Yahoo!ファイナンス掲示板投稿文書を用いた．また，SSCDV による分散表現獲得のため，感情辞書として，ドメインに依存しないものと，金融ドメインに依存しているものの 2 種類を用いた．本項では使用した各データについて説明する．

日経平均株価： 株価データセットとして JPX クラウド証券から収集した日経平均株価の始値を使用する．これは日本の個別銘柄の中でも流動性の高い 225 銘柄から算出されている．収集期間は 2012/11/26-2020/09/30 である．

日経平均 VI： ボラティリティーデータセットとして日経新聞社から収集した日経平均 VI の始値を使用する．日経平均 VI は，大阪取引所に上場している日経平均先物および日経平均オプションの価格をもとに算出される．日経平均 VI は，投資家が日経平均株価の将来の変動をどのように想定しているかを表した指数で

あり，指数値が高いほど，投資家が今後，相場が大きく変動すると見込んでいることを意味する．収集期間は 2012/11/26-2020/09/30 である．

Yahoo!ファイナンス掲示板データ： Yahoo ファイナンス掲示板の日経平均スレッドに投稿された文書を使用する．対象期間は 2012 年 11 月 26 日から 2020 年 9 月 30 日であり，対象となる文書は 9,463,958 個であった．

なお，Yahoo!ファイナンス掲示板投稿データは投稿内容（テキスト情報）のみを扱っているため，個人情報（個人を特定可能な ID 等）は扱っておらず，解析は全て Yahoo!JAPAN 研究所サーバ内で個人を特定しない形で行われた．

感情辞書：感情に関する事前知識である感情極性値は、事前に構築された感情辞書を参照して得られる。市場予測に適切な感情知識はドメインによって異なる可能性があると考える．例えば，”反発”は一般的にネガティブを表すが，金融ドメインの文脈ではポジティブである可能性が高い．我々は、そのようなセンチメント事前知識の影響を分析するために、2 つのタイプの感情辞書を使用する。一つは，ドメインに依存しない感情辞書として高村らが公開しているものを使用した [44]．もう一つは金融ドメインに依存した感情辞書として東京大学の和泉研究室が公開しているものを使用した [45]．実際，”反発”についてこの 2 つの辞書を参照すると，ドメインに依存しない辞書ではその極性値が-0.500296 であるのに対し，金融ドメインに依存しない辞書では極性値は 0.540438974 であった．

5.2.2 特徴量生成

市場予測において，前日からの差分や変化率も重要な特徴量となると考える．そこで，日経平均株価，日経平均 VI，$\vec{SSCDV}_{D_t}$ のそれぞれに対して，前日差分と変化率を算出し新たな特徴量を生成した．機械学習モデルに入力される最終的な特徴量を表 2 に示す．市場が閉まっている日には金融時系列データは存在しないが，文書は毎日投稿されている．我々は文書を用いたより効果的な予測のために，市場が閉まる前日の特徴セットの作成には金融時系列データとの市場が閉まっている最終日の文書分散表現を利用することとした（図 3）．

表 2 機械学習モデルの入力とする特徴量一覧

Features	Calculation	Description
$Price$	$Stock\_Open_t$	Nikkei 225 opening price
$Price_{Diff}$	$Stock\_Open_t - Stock\_Open_{t-1}$	Nikkei 225 Difference from Previous Day
$Price_{ROC}$	$Stock\_Open_t / Stock\_Open_{t-1} - 1$	Nikkei 225 Percentage Change from Previous Day
VI	VI_Open_t	Nikkei 225 VI opening price
VI_{Diff}	$VI\_Open_t - VI\_Open_{t-1}$	Nikkei 225 VI Difference from Previous Day
VI_{ROC}	$VI\_Open_t / VI\_Open_{t-1} - 1$	Nikkei 225 VI Percentage Change from Previous Day
$SS\vec{C}DV$	$SS\vec{C}DV_{D_t}$	$SS\vec{C}DV$
$SS\vec{C}DV_{Diff}$	$SS\vec{C}DV_{D_t} - SS\vec{C}DV_{D_{t-1}}$	$SS\vec{C}DV$ Difference from Previous Day
$SS\vec{C}DV_{ROC}$	$SS\vec{C}DV_{D_t} / SS\vec{C}DV_{D_{t-1}} - 1$	$SS\vec{C}DV$ Percentage Change from Previous Day

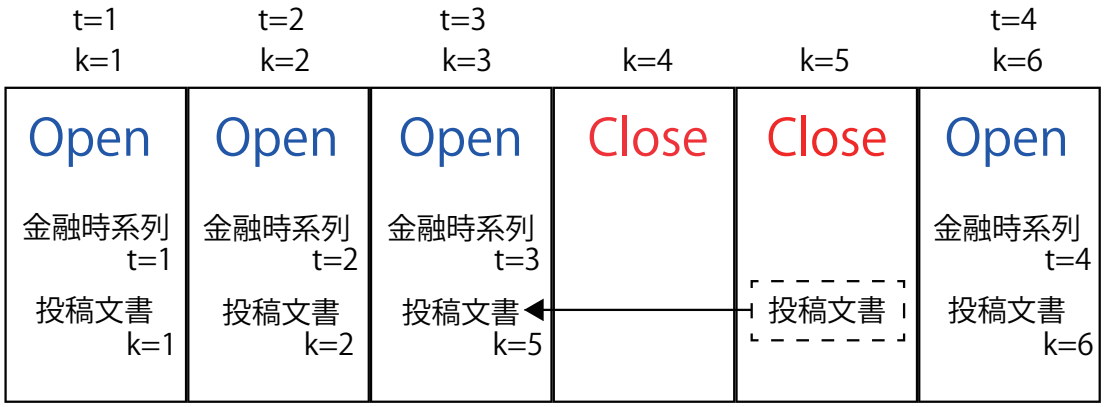

図 3 休日処理

5.2.3 ラベリング

表 2 に示した 9 種類の特徴量を用いた機械学習モデルにより，将来 5 営業日以内に日経平均 VI が閾値以上上昇（$y_t = 1$）するか上昇しないか（$y_t = 0$）を予測する 2 値分類タスクを設定する．ここで，y_t は t 日目のラベル．実験期間中のデータセットに対して先行研究 [46] と同様に以下の式に従いラベリングを行った．ここで T は実験期間中の営業日の日数，x_t は t 日目の日経平均 VI 始値の 1 日差分を表しており，$x_t = VI\_Open_t - VI\_Open_{t-1}$ である．時間窓 d は将来何日以内の上昇を予測するかを決定するパラメータであり，1 以上 T 以下の整数値をとる．i は

上昇の閾値を決定するパラメータであり，制約条件として正の実数値とする．本実験においてパラメータ d は $d=5$ とし，i は $i=2$ とした．このような条件のもとラベリングを行なった結果，VI 上昇日（$y_t=1$）とされた日数は 291 日であり，実験期間 1915 日の 15.2%である．また，テスト期間内で VI 上昇日（$y_t=1$）とされた日数は 119 日であり，テスト期間 929 日の 12.8%である．このことから我々のタスク設定におけるラベリングでは 2 つのクラスが非常に不均衡となっていることがわかる．実験期間における日経平均 VI の値の推移と正例（$y=1$）と判定された日についての関係を図 4 に示す．

$$y_t = \begin{cases} 1 \text{ if } \max\left(VI\_Open_{[t+1,t+d]}\right) - VI\_Open_t \geq i\sigma \\ 0 \text{ otherwise} \end{cases} \tag{17}$$

$$\text{where } \sigma = \sqrt{\frac{1}{T-1}\sum_{t=2}^{T-1}(x_t - \bar{x})^2} \tag{18}$$

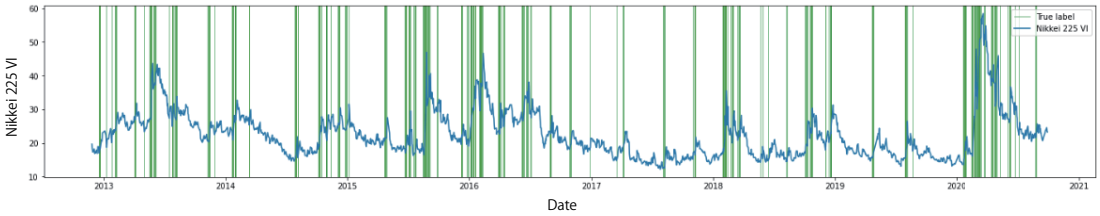

図 4 実験期間中の日経平均 VI の推移とラベルの正例

5.2.4 訓練・テストデータ

予測は機械学習モデルを通して行われる．機械学習モデルはランダムフォレストを採用した．学習および評価は，学習期間を固定し，1 日ずつずらしつつテストを行うローリングウィンドウ方式により行なった．具体的な方法を以下に示す．

1. 実験期間 T のうち，学習期間を n 日目から m 日目までの $m-n+1$ 日間として学習する．

2. 学習されたモデルを用いて $m+6$ 日目をテストする．

3. n と m を 1 だけ増加させる．

テスト日が実験期間の最終日に到達する，すなわち $m+6=T$ になるまで上記の (1-3) を k 回繰り返す．最終的な評価は k 回のテスト結果を集計したものとする．本研究では学習期間を約 4 年間である 980 日と固定した．そのため，$m-n+1=980$ を常に満たす．学習が必要なアルゴリズムは最初の 980 日分のデータを使用した．以後，この期間については開発期間と呼ぶ．

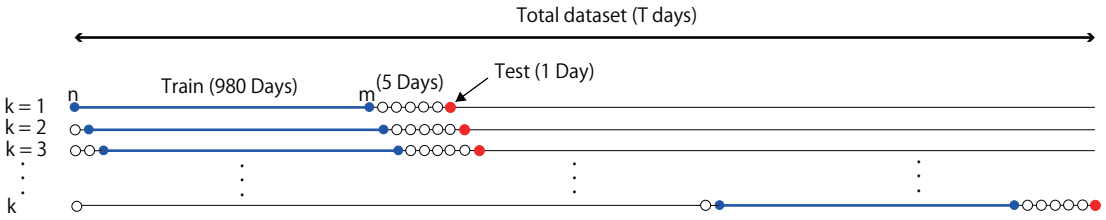

図 5 ローリングウィンドウ方式によるモデルの学習と評価

5.2.5 評価指標

我々は実験の評価のために，Accuracy, Precision, recall, F1-score, Matthews Correlation Coefficient（MCC）の 5 つの評価指標を使用する．MCC は 2 つのクラスが歪んでいる場合でも使用できるメトリックで，与えられた混同行列に対して定義される．

$$\frac{tp \times tn - fp \times fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}} \tag{19}$$

5.2.6 ベースライン

SSCDV により獲得された文書分散表現の有効性を評価するため，以下の手法 [16, 47, 48, 49, 17, 50, 51, 20] により獲得した文書分散表現を特徴量として使用した場合をベースラインとして比較する（つまり，獲得分散表現を $\vec{DVec}$ とした時，表 1 の特徴量セットの $\vec{SSCDV}$ に対応する部分が全て $\vec{DVec}$ に置き換わ

る）．また，価格データのみで予測した場合との比較も行い，ソーシャルメディアの有効性を確かめる．

入力テキストの前処理はモデルの特性に応じて以下の二種類の前処理を用意した．文脈を考慮するモデルである BERT，ELECTRA に関しては前処理 A を行なったテキストを用い，それ以外のモデルには前処理 B を行なったテキストを用いた．

前処理 A: テキストの正規化を行なったのち，形態素解析を行い，記号とストップワードを取り除いた．

前処理 B: テキストの正規化を行なったのち，形態素解析を行い，各文書から，名詞，動詞，形容詞の中で Subtype が数値，非自立，代名詞，接尾でないものを抽出し，ストップワードを取り除いた．

ストップワードには，公開されている日本語 Slothlib のデータ[1]に，日本語でよく使用される「ある、する、ちゃう、ない、なる、やる」の 6 つのワードを追加したもの使用した．形態素解析には Mecab[52] を使用し，辞書は，新規語や固有表現に強い NEologd[53] を使用した．なお、NEologd は 2020 年 5 月 21 日更新分を使用した．

- **Historical Only Method**: テキストデータを特徴量として用いず，金融時系列データ特徴量のみで予測を行うモデル．
- **LDA [16]**: テキスト解析に頻繁に用いられるトピックモデル．文書の潜在トピックへの所属確率を算出する．
- **JST [47]**: 文書の潜在トピックとセンチメントを共同で利用する LDA の派生モデル．各潜在トピックは感情分布に依存して生成されると仮定している．
- **SWEM-aver [48]**: word2vec や GloVe [54] などで得られた単語埋め込みの各配列の単純な平均や加算などにより，文や文書埋め込みとする手法．文書分類，シーケンスマッチング，タグ付けの 3 つの NLP タスクを含む 17 のデータセットにおける実験において，LSTM，CNN を用いた手法と同程度かそれ以上の精度を達成しており，単純な手法ながら，強力なベースライン．本

[1] http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/NLP/Filter/StopWord/word/Japanese.txt

研究では単語埋め込みの各配列の単純な平均により文書埋め込みを得る手法を採用する．

- **Doc2Vec [49]**: 文書埋め込み獲得のための word2vec の拡張手法．本研究では，文書 id と文書内の単語を用いて次の単語を予測するタスクを解くことで文書分散表現を獲得する PV-DM アルゴリズムを採用する．

- **Bidirectional Encoder Representations from Transformers（BERT） [17]**: Transformer と呼ばれる，アテンション機構をもつニューラルネットワーク構造を用いたモデル．様々な NLP タスクで高いパフォーマンスを発揮しており，テキストマイニングにおいてデファクトスタンダートとなってきている．本研究では日本語 Wikipedia コーパスで事前学習されたモデル[2]を使用．24 のレイヤー，1024 次元の隠れ層，16 の Attention heads から構成されている．

- **Efficiently Learning an Encoder that Classifies Token Replacements Accurately(ELECTRA）[55]**: BERT の事前学習手法を Generator と Discriminator を用いた手法で改善した，BERT の派生モデル．様々な日本語表現が混在しており，Wikipedia 文書よりもソーシャルメディア文書に近いコーパスである，日本語ブログコーパス（YACIS corpus [51]）で事前学習されたモデル [50] を使用する．

- **SCDV [20]**: 提案手法の拡張元となった手法．GMM ソフトクラスタリングを用いて word2vec で得られた各単語埋め込みを拡張することで，単語の所属トピックを考慮した文書分散表現を獲得する．20Newsgroup データセットを用いた文書分類タスクにおいて，lda や PV-DM などを含む 17 のベースラインモデルよりも高いパフォーマンスであったと報告されている．

[2]https://huggingface.co/cl-tohoku/bert-large-japanese

5.2.7 パラメータ設定

設定された各種パラメータについて述べる．SCDV,SWEM-aver,SSCDV アルゴリズムのための word2vec は，学習コーパスを開発期間中の投稿テキストとし，300 次元の埋め込みのために SGNS algorithm（Window-size=5）にて学習された．SSCDV のスパース化に用いる値 t(式 14) は開発期間内に対してあらかじめ計算されたもので固定した．SSCDV のハイパーパラメータである，クラスタ数 K とスパース化閾値 p は先行研究において最もパフォーマンスが良いと報告されている $K = 60, p = 4$ を使用した．SCDV に対しても同様にハイパーパラメータ K,p は $K = 60, p = 4$ を使用した．クラスタ数を一致させるため，LDA および JST のトピック数は SSCDV および SCDV の K の値と同じく 60 にセットした．機械学習モデルの学習時は毎回のトレーニングデータに対して前方クロスバリデーションを行い，グリッドサーチによりハイパーパラメータの決定を行なった．

6. 結果

本章では、評価実験の結果について述べる。

6.1 日経平均 VI 上昇予測の精度（RQ1）

リサーチクエスチョン（RQ1）への回答を通して，提案手法 SSCDV で獲得された分散表現の市場予測における有効性を評価する．

RQ1. Prediction Accracy (Sec. 6.1): 提案手法により獲得された分散表現を特徴量として用いた場合，VI 予測モデルの精度は向上するか？

SSCDV による分散表現を用いたモデルと他のベースラインモデルにより獲得された分散表現を用いたモデルの精度を比較した結果を表 3 に示す．SSCDV では感情辞書を二種類用意し，比較対象とした．ドメインに依存しない辞書を使用した場合を SSCDV(Lex. N) と表記し，金融ドメインのための辞書を用いた場合は SSCDV(Lex. F) と表記する．本研究のタスクは VI の大幅な上昇を予測するものである．そのため，Precision，Recall が共に高い値であることが望ましく，すなわち，F-1 Score と MCC が高いことが予測モデルに求められている．SSCDV(Lex. N) は他のベースラインと比較して，最も精度の高い Precision，F-1 Score，MCC を達成することができた．また，本研究において最もパフォーマンスの良かった SSCDV(Lex. N) の予測結果について，ラベルが正例（$y = 1$）の日（緑）と，上昇すると予測（$\hat{y} = 1$）した日（赤）を日経平均 VI の推移（青）とともに図 6 上に示し，予測が真陽性であった日（紫）を日経平均 VI の推移（青）とともに図 6 下に示す．図 6 より，日経平均 VI が将来的に大きく上昇する日を一定の精度で予測できていることがわかる．

表 3 VI 予測モデルの精度

Model	Accuracy(↑)	Precision(↑)	Recall(↑)	F1-score(↑)	MCC(↑)
Historical only method	0.581	0.177	**0.622**	0.276	0.132
LDA [16]	0.738	0.226	0.429	0.296	0.166
JST [47]	0.643	0.164	0.437	0.239	0.077
SWEM-aver [48]	0.716	0.234	0.538	0.327	0.205
Doc2Vec [49]	**0.755**	0.215	0.345	0.265	0.132
BERT [17]	0.704	0.225	0.538	0.318	0.193
ELECTRA [50]	0.649	0.184	0.504	0.269	0.122
SCDV [20]	0.718	0.220	0.471	0.300	0.168
SSCDV(Lex. N)	0.716	**0.238**	0.555	**0.333**	**0.215**
SSCDV(Lex. F)	0.717	0.229	0.513	0.317	0.192

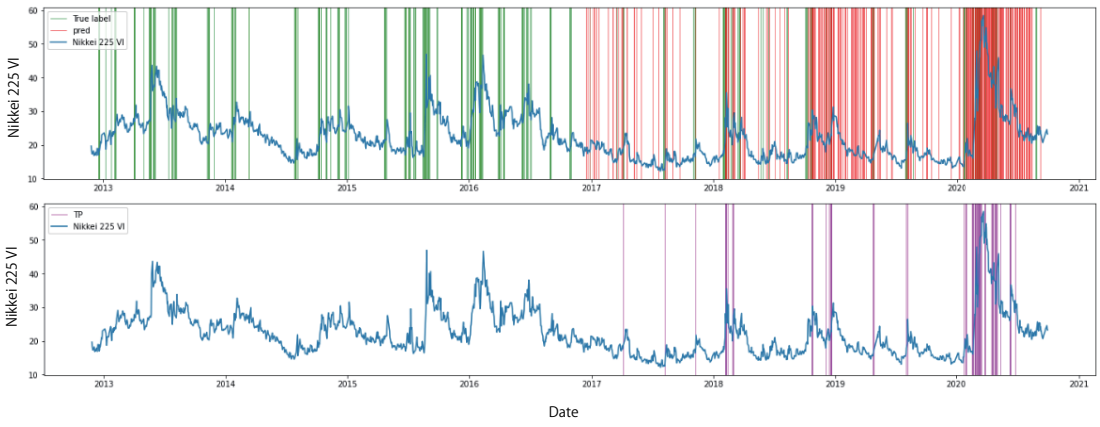

図 6 2 値分類の正例とモデルの予測と日経平均 VI の関係.

6.2 モデルの予測結果を用いた取引シミュレーションによる累積獲得利益（RQ2）

リサーチクエスチョン（RQ2）への回答を通して，提案手法 SSCDV の市場予測における有効性を評価する．

RQ2. Investment Simulation (Sec. 5.2): VI 予測モデルの予測結果に基づいた投資シミュレーションにおいて，いくら利益が獲得できるか？

投資戦略の一つにロングストラドル戦略と呼ばれる戦略が存在する．この戦略では，市場が大きく変動することが利益獲得の条件となる．我々のモデルは投資リスク低減のために設計され，VI の大幅な上昇を予測するモデルではあるが，この戦略を用いることで，我々のモデルの予測が投資における利益獲得にもつながる可能性がある．そこで，予測結果を用いた投資シミュレーションにより，実世界での利益獲得の側面から提案モデルを評価する．投資シミュレーションの結果を表 4 に示す．投資シミュレーションは佐々木らの戦略 [56] に従った．ただし，決済を行う閾値を 750 円に設定した．投資シミュレーションの結果，BERT を用いたモデルが最も高い累積利益を獲得した．我々の提案モデルは F-1 Score や MCC では最も高い精度となっていたが，累積獲得利益に関してはプラスではあったものの他のベースラインを上回ることはなかった．獲得利益を最大化するためには，直接獲得利益を最大化するためのタスク設定を行う必要があると考えられる．

表 4 投資シミュレーションによる累積獲得利益

Model	累積獲得利益 /円	取引した日の平均獲得利益 /円
Historical Only Method	-7,874,570	-19,400
LDA [16]	+1,722,870	+7,980
JST [47]	-1,262,320	-4,130
SWEM-aver [48]	+7,751,860	**+29,250**
Doc2Vec [49]	+201,530	+1,110
BERT [17]	**+7,953,150**	+28,920
ELECTRA [50]	+5,649,970	+17,880
SCDV [20]	+3,173,060	+12,900
SSCDV(Lex. N)	+1,249,880	+4,700
SSCDV(Lex. F)	+2,544,860	+9,900

7. 考察

本章では我々のモデルをより深く考察するため行なった，SHAP による実験結果とハイパーパラメータの変動による精度変化に関して述べる．

7.1 SHAP 値によるモデル解釈

一般的に高度な機械学習モデルは特徴量の複雑な関係を効率的に学習し，高いパフォーマンスを発揮する一方で，モデルの予測がどのように行われたのかについてはブラックボックス化される．機械学習モデルによる金融市場予測を投資における意思決定支援のためのアプリケーションとして捉えた場合，このブラックボックス化はモデルの予測に対する信頼を低下させてしまうという問題につながる．ある特徴量がモデルの予測値にどれだけ影響を与えるかを説明する SHAP を用いて，我々のモデル SSCDV がどのように VI の予測に影響を与えたのかを解釈することを試みた．SSCDV の最終的なベクトルは解釈可能性を残しており，このことは他のモデルよりも優れた点の一つである．各日の予測モデルの SHAP 値の結果を Fig. 7 に示す．SHAP 値は d 次元（センチメントトピックベクトルに対応する次元）ごとに加算したものである．我々はこの図における 2 つの興味深い点を説明する．1 つはソーシャルメディアの投稿トピックにはどの期間にも一貫して影響を与えるトピック（e.g, (a-e)）が存在するということである（Case A）．もう一つは，ある時期から突然，予測に影響を与えるトピック (f-g) が存在することである (Case B)．さらに，f と g は共にトピックのネガティブが予測に強く影響を与えていることが分かる．興味深い点は f と g が影響を与える時期は 2020 年 1 月 29 日から後の時期であり，これは新型コロナウィルスについて議論が始まった少し後の時期ということである．そこで，(a-g) に対応するクラスタの上位語を調査した．結果を表 5 に示す．CaseA の中には政治関係 (c:Topic22) や金融関係 (d:Topic26)，スポーツ関係 (e:Topic60) のようなクラスターが確認された．CaseB の中には戦争関係 (f:Topic60) や金額，人数関係 (g:Topic51) のクラスターが確認された．

金額，人数関係 (g:Topic51) のトピックが突然予測に影響を与え始めたことに関

して，直感的にはこの時期，コロナ関係で毎日感染者数が話題に上っていたことが影響している可能性がある．また，戦争関係 (f:Topic60) のトピックが突然予測に影響を与え始めたことに関しては，直感的には，この時期，ロシアのウクライナ侵攻についての話題が影響している可能性がある．実際の投稿文を確認し，何が影響していたのかを今後調査する必要がある．

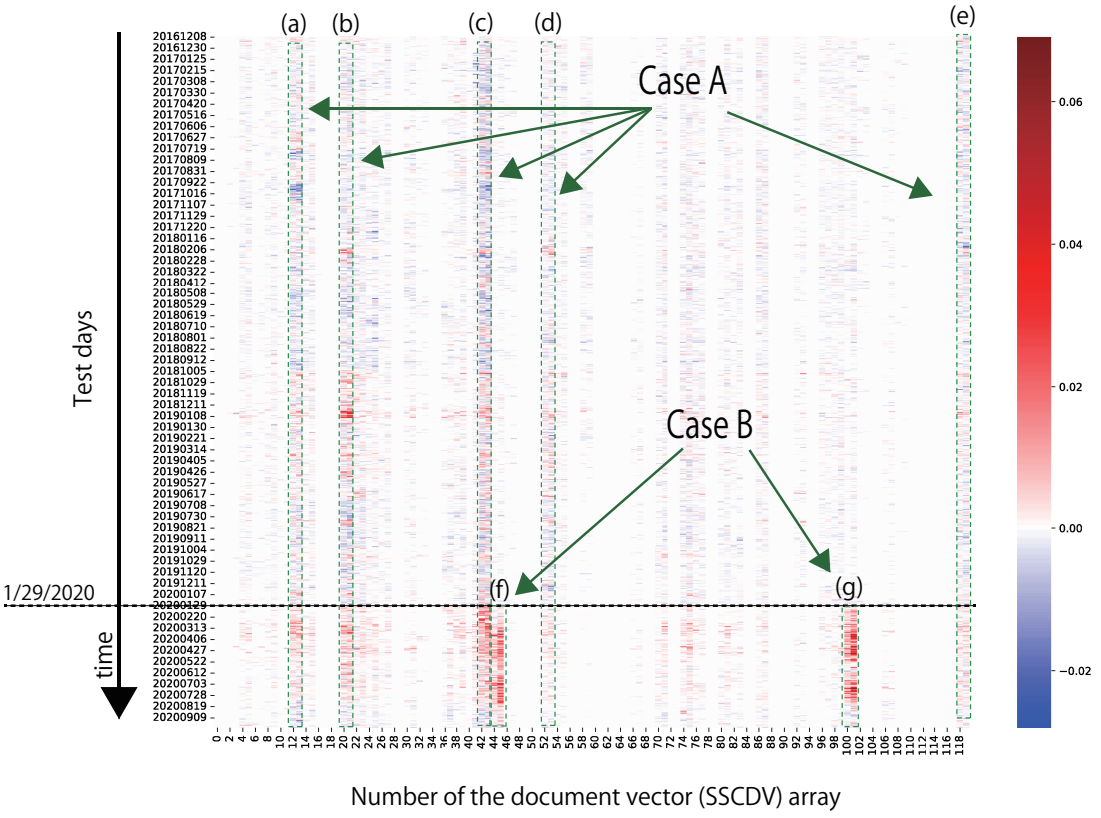

図 7 各トピックとセンチメントに対応する shap 値

表 5 各クラスタの上位語

(a):Topic7(駅名，場所名)	(b):Topic11(人名)	(c):Topic22(政治)	(d):Topic26(金融)	(e):Topic60(スポーツ)	(f):Topic60(戦争)	(g):Topic51(金額，人数)
品川駅	増子輝彦	国際調査報道ジャーナリスト連合	市中銀行	打線	強襲揚陸間	900 億円
常盤	山口泰明	ICIJ	当座預金	主審	早期警戒機	12 億円
上野駅	井上英孝	租税回避地	信用リスク	FC ソウル	艦載機	8900 万円
佐賀駅	遠藤敬	パナマ文書	貸出金利	開幕戦	ミサイル駆逐艦	26 倍
常盤線	坂本哲志	バージン諸島	金融派生商品	3 位決定戦	護衛艦	70 万人
新橋駅	中川俊直	バハマ	国債市場	ワールドシリーズ	攻撃機	80 億円
上野東京ライン	竹本直一	金言	国債買い入れ	バンクーバー五輪	艦載	7100 万円
寝台特急	小川勝也	租税回避	スイスフラン	公式戦	哨戒	8600 万
東海道	平野博文	検証委員会	ギリシャ国際	ブラジル W 杯	哨戒機	36 億
南口	上野賢一郎	南ドイツ新聞	資本流出	試合	偵察機	18 億

8. 結論

本研究ではソーシャルメディア投稿文書用いて高精度に将来の VI を予測することを目的とし，トピックと感情の特徴を同時に埋め込む手法，SSCDV を提案した．SSCDV は，SCDV アルゴリズムを事前辞書を用いた感情クラス機能で拡張することで開発された．獲得した文書分散表現と金融時系列データを機械学習モデルに共同で学習させ，VI の上昇予測を行なった．予測結果を用いて，2 種類のリサーチクエスチョンへ回答することで，提案アルゴリズムの評価を行なった．

1 つ目のリサーチクエスチョンでは，提案アルゴリズムを用いた VI 予測モデルの予測精度の観点から提案手法を評価した．我々のモデルは，Precision が 0.238，F-1-score が 0.333，MCC が 0.215 という予測精度となり，様々なベースラインモデルと比較して，最も高いパフォーマンスを示した．文書のトピックと感情の情報を共同で埋め込みに反映させることで，市場予測に有効なシグナルを捉えることができたと考えられる．

2 つ目のリサーチクエスチョンでは，提案アルゴリズムを用いた VI 予測モデルの実利的有効性の観点から提案手法を評価した．我々の構築した VI 予測モデルは市場の価格変動が大きくなることを予測している．そのため，価格変動が大きくなるときに利益が獲得できるロングストラドル戦略を用いた投資シミュレーションを行った．投資シミュレーションの結果，我々のモデル（SSCDV(Lex. F)）は，累積獲得利益が+2,544,860 円となり，提案アルゴリズムを用いた VI 予測モデルによる利益獲得の可能性を示した．ソーシャルメディア投稿文書を用いない Historical only method の予測結果を用いた投資シミュレーションでは累積獲得利益が-7,874,570 円となっており，ソーシャルメディア投稿文書を用いたことで利益獲得につながったと考えられる．しかし，我々のモデルは，予測精度の観点では他の手法と比べ最高のパフォーマンスを達成したものの，投資シミュレーションによる利益獲得の観点では，最も高い利益獲得とはならなかった．これは，我々のモデルが VI 予測に対して最適化されているためと考えられる．そのため，直接利益獲得を目指すタスク設定にすることで，獲得利益が増加すると考えられる．

さらに，SSCDV により獲得された分散表現と予測の関係を説明するために，SHAP を用いてその関係の可視化を行った．その結果，どのトピック，感情が予測

に影響を与えているのかを時系列で追うことに成功した．これにより，ソーシャルメディアの投稿トピックにはどの期間にも一貫して影響を与えるトピックと，ある時期から突然予測に影響を与えるトピックが存在することを確認した．

このように，我々の提案した SSCDV アルゴリズムは，文書から金融市場予測に有効な分散表現を獲得できている．そのため，本研究で提案した文書分散表現は，他の高度な機械学習アルゴリズム（LSTM や XGBoost[57]，TabNet[58]）を用いた市場予測や，テキストに対するエンコードを必要とする市場予測のためのアーキテクチャ [40, 59, 60] に拡張できると考える．さらに，本研究では VI の予測に焦点を当てたが，提案した埋め込み技術は株価予測や，ポートフォリオ管理といった金融市場予測に関わるその他のタスクにおいても有効な特徴量となる可能性があり，様々な方面での活用が期待できる．

最後に，提案アルゴリズム SSCDV の課題や今後の展望について述べる．本研究では，計算コストの観点から開発期間に含まれる語彙で獲得した単語埋め込みを使用したため，新規に現れる語彙やトピックに対応できていないという課題を残している．また，本研究では，2 種類の感情辞書を用いて，感情極性（Positive,Negative）とその強度の割り当てを行なった．近年の感情分析の研究 [61, 62] においては，より詳細な感情の識別や感情強度の推定にも焦点が当てられており，それらの活用は，提案アルゴリズムを用いた市場予測において，さらなる精度向上の可能性を残す．しかし，本研究では，2 種の感情辞書を用いての感情の組み込みしか実験できておらず，他の手法における感情の組み込みは未だ検証できていない．提案アルゴリズムにおいて，他の手法による感情の組み込みをすることで，どのように市場予測の精度が変化するかの検証は，今後の課題である．さらに，SSCDV アルゴリズムのトピック数とスパース化閾値は事前に決定される必要があるものとなっている．更なる精度向上のためには，新規語への対応や，時期に応じた最適なトピック数，スパース化閾値の決定を時系列に伴い動的に行なう必要があると考えられ，これらは今後の課題とする．本研究では，ソーシャルメディア文書として Yahoo!ファイナンス掲示板への投稿文書を用いた．しかし，ソーシャルメディアには他にも Twitter やブログなどが存在する．各メディアによってユーザの性質が異なると考えられ，ユーザの性質の違いから生まれる書き込み情報の違い

は市場予測において異なる影響を与えると考えられる．そこで，他メディアデータを用いて市場予測の実験，評価を行い，Yahoo!ファイナンス掲示板を用いた場合との比較分析を行うことも今後検討していく．

謝辞

本論文の執筆，および研究をすすめるにあたり，様々な方々に御協力を賜りました．ここに謝意を添えて御名前を記させていただきます．

安本慶一　教授には，研究全般に関し，多大なるご指導・ご助言を賜りました．また，充実した研究環境の整備など，研究活動を手厚くご支援いただきました．感謝の意を表すとともに，心より厚く御礼申し上げます．

荒牧英治　教授には，ご多忙の中，論文審査委員を引き受けてくださった上で，発表の場においては様々なご助言をいただきました．感謝の意を表すとともに，心より厚く御礼申しあげます．

諏訪博彦 准教授には，研究の初期段階から的確なご指導，およびご指摘をいただきました．特に，研究に行き詰まった際や論文の執筆の際に，多くのご助言を賜りました．感謝の意を表すとともに，心より厚く御礼申しあげます．

松田裕貴 助教には，日常的に研究室内の様々なことに対してご助言を賜りました．感謝の意を表すとともに，心より厚く御礼申し上げます．

金岡恵 事務補佐員，山内奈緒 事務補佐員には，学会や出張に関する事務処理を始め，研究生活の様々な場面でご支援いただきましたこと，謹んで感謝申し上げます．

立命館大学理工学部　小川祐樹　講師には，研究に関する多数のご助言をいただきました．感謝の意を表すとともに，心より厚く御礼申し上げます．

新潟国際情報大学経営情報学部　梅原英一　教授には，研究を進めるにあたり多数のご助言をいただきました．中でも金融に関する手法において根本的な議論，説明をしていただきました．心より厚く御礼申し上げます．

坪内孝太研究員，山下達雄研究員を始めとする Yahoo!JAPAN 研究所の皆様には，研究所サーバ内においてデータ分析するにあたり，多大なご助力をいただいたのみならず，研究に対しても的確なご指摘をいただきました．心より感謝申し上げます．

また研究全般において，的確なアドバイスをくださった先輩の皆様，共に研究生活を過ごしたユビキタスコンピューティングシステム研究室の同輩，後輩には，公私ともにお世話になりました．心より感謝申しあげます．

最後に，今日まで学生生活を様々な面から支えてくださった家族に心より感謝申し上げます.

参考文献

[1] Yaser S. Abu-Mostafa and Amir F. Atiya. Introduction to financial forecasting. *Applied Intelligence*, Vol. 6, pp. 205–213, 1996.

[2] 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*, WWW '21, p. 1–10, New York, NY, USA, 2021. Association for Computing Machinery.

[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. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '19, p. 2376–2384, New York, NY, USA, 2019. Association for Computing Machinery.

[4] Baohua Wang, Hejiao Huang, and Xiaolong Wang. A novel text mining approach to financial time series forecasting. *Neurocomputing*, Vol. 83, pp. 136–145, 2012.

[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, Online, July 2020. Association for Computational Linguistics.

[6] Yauheniya Shynkevich, T.M. McGinnity, Sonya A. Coleman, and Ammar Belatreche. Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning. *Decision Support Systems*, Vol. 85, pp. 74–83, 2016.

[7] Junwen Duan, Yue Zhang, Xiao Ding, Ching-Yun Chang, and Ting Liu. Learning target-specific representations of financial news documents for cumulative abnormal return prediction. In *Proceedings of the 27th International Conference*

on Computational Linguistics, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[8] Robert P. Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.*, Vol. 27, No. 2, mar 2009.

[9] Twitter mood predicts the stock market. *Journal of Computational Science*, Vol. 2, No. 1, pp. 1–8, 2011.

[10] Bing Li, Keith C.C. Chan, Carol Ou, and Sun Ruifeng. Discovering public sentiment in social media for predicting stock movement of publicly listed companies. Vol. 69, No. C, 2017.

[11] Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. Sentiment analysis of twitter data for predicting stock market movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pp. 1345–1350, 2016.

[12] Tien Thanh Vu, Shu Chang, Quang Thuy Ha, and Nigel Collier. An experiment in integrating sentiment features for tech stock prediction in twitter. In *Proceedings of the workshop on information extraction and entity analytics on social media data*, pp. 23–38, 2012.

[13] Yahya Eru Cakra and Bayu Distiawan Trisedya. Stock price prediction using linear regression based on sentiment analysis. In *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 147–154, 2015.

[14] Weiling Chen, Chai Kiat Yeo, Chiew Tong Lau, and Bu Sung Lee. Leveraging social media news to predict stock index movement using rnn-boost. *Data Knowledge Engineering*, Vol. 118, pp. 14–24, 2018.

[15] Thien Hai Nguyen and Kiyoaki Shirai. Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China, July 2015. Association for Computational Linguistics.

[16] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022, 2003.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[18] Ankit Kumar, Piyush Makhija, and Anuj Gupta. Noisy text data: Achilles' heel of BERT. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pp. 16–21, Online, November 2020. Association for Computational Linguistics.

[19] Buddhika Kasthuriarachchy, Madhu Chetty, Adrian Shatte, and Darren Walls. From general language understanding to noisy text comprehension. *Applied Sciences*, Vol. 11, No. 17, 2021.

[20] Dheeraj Mekala, Vivek Gupta, Bhargavi Paranjape, and Harish Karnick. SCDV : Sparse composite document vectors using soft clustering over distributional representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[21] Douglas A. Reynolds. Gaussian mixture models. In *Encyclopedia of Biometrics*, 2009.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing*

Systems - Volume 2, NIPS'13, p. 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

[23] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc.

[24] Xiaowei Lin, Zehong Yang, and Yixu Song. Short-term stock price prediction based on echo state networks. *Expert Systems with Applications*, Vol. 36, No. 3, Part 2, pp. 7313–7317, 2009.

[25] Wei Bao, Jun Yue, and Yulei Rao. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLOS ONE*, Vol. 12, No. 7, pp. 1–24, 07 2017.

[26] Qianggang Ding, Sifan Wu, Hao Sun, Jiadong Guo, and Jian Guo. Hierarchical multi-scale gaussian transformer for stock movement prediction. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 4640–4646. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Special Track on AI in FinTech.

[27] Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. Enhancing stock movement prediction with adversarial training. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 5843–5849. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[28] Jintao Liu, Hongfei Lin, Xikai Liu, Bo Xu, Yuqi Ren, Yufeng Diao, and Liang Yang. Transformer-based capsule network for stock movement prediction. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pp. 66–73, 2019.

[29] Rui Cheng and Qing Li. Modeling the momentum spillover effect for stock prediction via attribute-driven graph attention networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, No. 1, pp. 55–62, May 2021.

[30] Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Alexander Dür, Linda Andersson, and Allan Hanbury. Volatility prediction using financial disclosures sentiments with word embedding-based IR models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[31] 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 (Volume 2: Short Papers)*, pp. 24–29, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[32] Christoph Kilian Theil, Samuel Broscheit, and Heiner Stuckenschmidt. Profet: Predicting the risk of firms from event transcripts. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 5211–5217. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[33] Zihao Zhang, Stefan Zohren, and Stephen Roberts. Deeplob: Deep convolutional neural networks for limit order books. *IEEE Transactions on Signal Processing*, Vol. 67, No. 11, pp. 3001–3012, 2019.

[34] Zane Turner, Kevin Labille, and Susan Gauch. Lexicon-based sentiment analysis for stock movement prediction. *Journal of Construction Materials*, Vol. 2, No. 3, pp. 3–5, 2021.

[35] Hirohiko Suwa, Yuki Ogawa, Eiichi Umehara, Kento Kakigi, Keiichi Yasumoto, Tatsuo Yamashita, and Kota Tsubouchi. Develop method to predict the increase

in the nikkei vi index. In *2017 IEEE International Conference on Big Data (Big Data)*, pp. 3133–3138, 2017.

[36] Xi Zhang, Yunjia Zhang, Senzhang Wang, Yuntao Yao, Binxing Fang, and Philip S. Yu. Improving stock market prediction via heterogeneous information fusion. *Knowledge-Based Systems*, Vol. 143, pp. 236–247, 2018.

[37] Nader Mahmoudi, Paul Docherty, and Pablo Moscato. Deep neural networks understand investors better. *Decision Support Systems*, Vol. 112, pp. 23–34, 2018.

[38] Ryo Akita, Akira Yoshihara, Takashi Matsubara, and Kuniaki Uehara. Deep learning for stock prediction using numerical and textual information. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pp. 1–6, 2016.

[39] Xuan Ji, Jiachen Wang, and Zhijun Yan. A stock price prediction method based on deep learning technology. *International Journal of Crowd Science*, Vol. ahead-of-print, , 03 2021.

[40] Hongfeng Xu, Lei Chai, Zhiming Luo, and Shaozi Li. Stock movement predictive network via incorporative attention mechanisms based on tweet and historical prices. *Neurocomputing*, Vol. 418, pp. 326–339, 2020.

[41] Matheus Gomes Sousa, Kenzo Sakiyama, Lucas de Souza Rodrigues, Pedro Henrique Moraes, Eraldo Rezende Fernandes, and Edson Takashi Matsubara. Bert for stock market sentiment analysis. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 1597–1601, 2019.

[42] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. Finbert: A pre-trained financial language representation model for financial text mining. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 4513–4519. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Special Track on AI in FinTech.

[43] 上田健太郎, 諏訪博彦, 小川祐樹, 梅原英一, 山下達雄, 坪内孝太, 安本慶一. 日経 vi 予測のためのソーシャルメディアの感情とトピックを用いた文書分散表現獲得手法の提案. 社会システムと情報技術研究ウィーク (WSSIT2022), SIG-SAI, 2022.

[44] Hiroya Takamura, Takashi Inui, and Manabu Okumura. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[45] Tomoki Ito, Hiroki Sakaji, Kota Tsubouchi, Kiyoshi Izumi, and Tatsuo Yamashita. Text-visualizing neural network model: Understanding online financial textual data. In Dinh Phung, Vincent S. Tseng, Geoffrey I. Webb, Bao Ho, Mohadeseh Ganji, and Lida Rashidi, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 247–259, Cham, 2018. Springer International Publishing.

[46] Kentaro Ueda, Kodai Sasaki, Hirohiko Suwa, Yuki Ogawa, Eiichi Umehara, Tatsuo Yamashita, Kota Tsubouchi, and Keiichi Yasumoto. Prediction of nikkei vi increase for reducing investment risk using yahoo! japan stock bbs. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT '21, p. 126–133, New York, NY, USA, 2022. Association for Computing Machinery.

[47] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. New York, NY, USA, 2009. Association for Computing Machinery.

[48] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[49] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, Vol. 32 of *Proceedings of Machine Learning Research*, pp. 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR.

[50] Shogo Shibata, Michal Ptaszynski, Juuso Eronen, Karol Nowakowski, and Fumito Masui. Development and performance evaluation of electra pretrained language model based on yacis large-scale japanese blog corpus [in japanese]. In *Proceedings of The 28th Annual Meeting of The Association for Natural Language Processing (NLP2022)*, pp. 1–4, 2022.

[51] Michal Ptaszynski, Pawel Dybala, Rafal Rzepka, Kenji Araki, and Yoshio Momouchi. Yacis: A five-billion-word corpus of japanese blogs fully annotated with syntactic and affective information. In *Proceedings of the AISB/IACAP world congress*, pp. 40–49, 2012.

[52] T. KUDO. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.jp*, 2006.

[53] Sato Toshinori. Neologism dictionary based on the language resources on the web for mecab, 2015.

[54] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[55] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[56] Kodai Sasaki, Hirohiko Suwa, Yuki Ogawa, Eiichi Umehara, Tatsuo Yamashita, and Kota Tsubouchi. Evaluation of vi index forecasting model by machine learn-

ing for yahoo! stock bbs using volatility trading simulation. In *HICSS*, pp. 1–9, 2020.

[57] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, p. 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[58] Sercan Ö. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, No. 8, pp. 6679–6687, May 2021.

[59] Huizhe Wu, Wei Zhang, Weiwei Shen, and Jun Wang. Hybrid deep sequential modeling for social text-driven stock prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, p. 1627–1630, New York, NY, USA, 2018. Association for Computing Machinery.

[60] Heyuan Wang, Tengjiao Wang, Shun Li, Shijie Guan, Jiayi Zheng, and Wei Chen. Heterogeneous interactive snapshot network for review-enhanced stock profiling and recommendation. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 3962–3969. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[61] Laura-Ana-Maria Bostan and Roman Klinger. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2104–2119, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[62] Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2095–2104, Online, June 2021. Association for Computational Linguistics.

研究業績

国際会議

1. Kentaro Ueda, Kodai Sasaki, Hirohiko Suwa, Yuki Ogawa, Eiichi Umehara, Tatsuo Yamashita, Kota Tsubouchi, and Keiichi Yasumoto. Prediction of Nikkei VI increase for reducing investment risk using Yahoo! JAPAN stock BBS. In IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT '21). Association for Computing Machinery, New York, NY, USA, 126–133. 2021

国内会議

1. 上田健太郎，諏訪博彦，小川祐樹，梅原英一，山下達夫，坪内孝太，安本慶一：金融指標予測のためのソーシャルメディアに適した分散表現獲得手法の検討，第 28 回社会情報システム学シンポジウム（iss28），沖縄，2022 年 1 月

2. 上田健太郎，諏訪博彦，小川祐樹，梅原英一，山下達夫，坪内孝太，安本慶一：日経 VI 予測のためのソーシャルメディアの感情とトピックを用いた文書分散表現獲得手法の提案，社会システムと情報技術研究ウィーク（WSSIT2022），SIG-SAI，北海道，2022 年 3 月

3. 細川蓮，小川祐樹，上田健太郎，諏訪博彦，梅原英一，山下達雄，坪内孝太：株式掲示板と新聞記事データを用いた VI 指数予測，2022 年度人工知能学会全国大会，京都，2022 年 6 月