The 37th Annual Conference of the Japanese Society for Artificial Intelligence, 2023

新聞と株式掲示板を用いた金融指標の予測と売買シミュレーション

Forecasting and Trading Simulation of Financial Indicators Using Newspapers and Stock Bulletin Boards

細川 蓮 *1 Ren Hosokawa
山田 優生 *2 Masaki Yamada
上田 健太郎 *3 Kentaro Ueda
諏訪 博彦 *3 Hirohiko Suwa
梅原 英一 *4 Eiichi Umehara
山下 達雄 *5 Tatsuo Yamashita
坪内 孝太 *5 Kota Tsubouchi
小川 祐樹 *2 Yuki Ogawa

*1立命館大学大学院 情報理工学研究科
Graduate School of Information Science and Engineering, Ritsumeikan University

*2立命館大学 情報理工学部
Dep. of Information Science and Engineering, Ritsumeikan University

*3奈良先端科学技術大学院大学 先端科学技術研究科
Nara Institute of Science and Technology

*4新潟国際情報大学 経営情報学部
Niigata University of International and Information Studies

*5ヤフー株式会社 Yahoo！JAPAN 研究所
Yahoo! JAPAN Corporation

Abstract: In stock investment, it is important for investors' profitability to predict the future trend of the market. The volatility index (VI) is one of the financial indices that expresses investors' psychological state toward the market. If investors can predict an increase in the VI, they may be able to reduce their investment risk. In this study, we predict the rise of the "Nikkei 225 VI Index" using both newspaper and stock bulletin boards. The results of this study show that our model is more prediction accurate than conventional models. In addition, we have validated our model by trading simulations to demonstrate its usefulness. As a result, our model yielded a profit of +745,000 yen.

1. はじめに

株式投資において，将来の市場の動向を推測することは，投資家の収益のために重要である．金融指標の一つに，投資家の市場に対する心理状態を表したボラティリティ・インデックス（以下，VI）という指標がある．VI の上昇を予測することができれば，投資家の投資リスクを低減できる可能性がある．しかし，株価の値動きは，効率的市場仮説 [Fama 70] によりランダムウォークするため，VI の予測は非常に困難な課題である．本研究では，マスメディアとソーシャルメディアの両方のメディアに着目する．新聞などのマスメディアの情報は，経済や金融に影響を与える出来事や社会情勢を把握するうえで有用であり，またソーシャルメディアのデータも社会の出来事だけでは捉えきれない投資家の関心や意見・感情を測るうえで有用なデータといえる．先行研究の多くはどちらか一方のメディアに焦点を当てて金融指標の予測を行っているが，これらの両メディアのデータを解析し統合して金融指標の予測に用いることで，社会情勢と投資家心理の両方を考慮した効果的な予測が可能になると考える．

本研究では，マスメディアとソーシャルメディアの両方のメディアに着目し，新聞の記事とインターネット株式掲示板（以下，株式掲示板）を用いて，日本における代表的な VI 指数である日経平均ボラティリティ・インデックス（以下，日経平均 VI） の上昇を予測する．さらに，提案手法の有効性を示すためにロングストラドル戦略を用いたオプション取引による売買シミュレーションによる検証も行う．

連絡先: 細川 蓮，立命館大学大学院情報理工学研究科
〒 525-8577 滋賀県草津市野路東 1 丁目 1 － 1
E-mail: is0474iv@ed.ritsumei.ac.jp

2. 関連研究

2.1 マスメディアを用いた金融指標の予測

金融指標の予測において，マスメディアのテキスト情報を用いた研究は，数多く行われてきた．例えば，Peng ら [Peng 16] は，一般的な単語埋め込み手法と Deep Neural Network を適用し，金融ニュースを活用した株価の動きを予測するモデルを構築した．実験の結果として，金融ニュースを株価予測モデルに適用することで，過去の金融価格情報のみを用いたベースラインよりも，株価予測精度を大幅に向上させたことを示した．マスメディアのテキスト情報は，情報の信頼性が高く，ノイズとなる文書が少ないことから，金融指標の予測に広く応用されている．

2.2 ソーシャルメディアを用いた金融指標の予測

金融指標の予測において，ソーシャルメディアのテキスト情報を用いた研究も，これまでに行われてきた．Suwa ら [Suwa 17] は，株式売買に特化したソーシャルメディアを用いて，日経平均 VI の上昇を予測する手法を提案した．具体的には，ソーシャルメディアの投稿文書に対して，トピックモデル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，それらを日別にまとめ，機械学習モデルの入力とした．そして，日経平均 VI の上昇とその他の 2 値分類問題を設定し，実験の結果として，Precision， Recall ともに 0.45 の予測性能を得た．このように，ソーシャルメディアのテキスト情報は，投資家の意見や気持ちの情報も反映されており，金融指標の予測に有効である可能性を示している．

2.3 金融指標の予測の有用性に関する研究

金融指標の予測において，有用性を検証するために，売買シミュレーションによる評価を行った研究がある．例えば，Sasaki ら [Sasaki 20] は，Suwa ら [Suwa 17] の構築した予測モデルに対して，ロングストラドル戦略を用いたオプション取引による売買シミュレーションを行った． 実験の結果として，Suwa らのモデルの指示で取引した場合の損益は，+3021 円であった．ベンチマークの損益は，− 3,590 円であったことから，改善額は，+6,611 円であった．したがって，Suwa らの日経平均 VI 予測モデルが有効である可能性が確認された．

3. 分析手法

3.1 VI 予測モデルの構築

本研究における VI 予測モデルについての概要を図 1 に示す．新聞の記事及び，株式掲示板の投稿から文書の特徴を捉えるために，文書を日別にまとめ，形態素解析により形態素に分割する．それぞれの文書群から，トピックモデルを用いて，日別におけるトピックへの所属確率を獲得し，それらを特徴ベクトルとする．獲得したそれぞれの特徴ベクトルに，金融時系列データを加えて，機械学習モデルへの入力とする．最後に，機械学習アルゴリズムを用いて，先行研究 [Suwa 17] と同様に，投資リスクとなる VI 指数の大幅な上昇を行う．

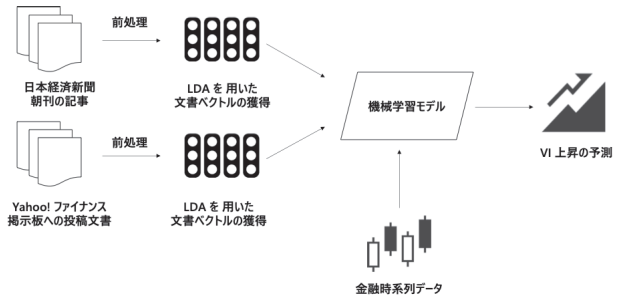

図 1: 本研究の分析概要図

3.2 特徴ベクトルの獲得

新聞記事及び，株式掲示板の投稿文書のそれぞれから特徴ベクトルを獲得する．具体的には，それぞれの文書群において，トピックモデルである LDA[Blei 03] を用いて，日別のトピックへの所属確率を算出し，これらを文書の特徴ベクトルとする．本研究における対象は，時系列データであるため，予測対象に未来の情報を含まないようにする．したがって，それぞれの文書において，図 2 のように，トピックモデルを構築する期間とトピックモデルを適用する期間を設定する．検証期間には，トピックモデルを適用する期間のみの特徴ベクトルを用いる．

3.3 オプション取引による売買シミュレーション

本研究で構築した予測モデルの有用性を検証するために，オプション取引に基づいた売買シミュレーションを行う．本研究における売買シミュレーションの位置付けを図 3 に示す．本研究で提案したモデルの予測対象は，日経平均 VI の上昇であり，日経平均株価の値動きに対応した証拠金取引が適切である．そのため，本研究における売買シミュレーションとして，日経 225 オプションを用いる．日経平均 VI の上昇は，将来の相場の変動可能性を示唆しているため，相場が大きく動いた際に利益が生まれる戦略であるロングストラドル戦略を採用

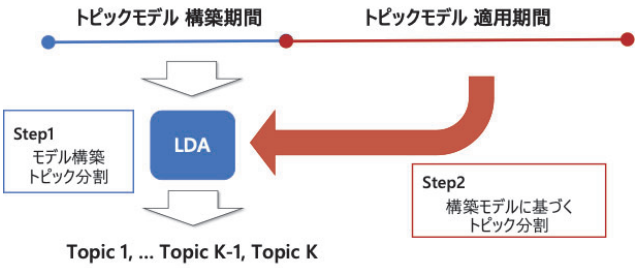

図 2: 特徴ベクトルの獲得手法

する．予測モデルの結果は，各営業日に対してモデルの確信度が出力されており，この値が閾値を超えた場合を「買いシグナル」とする．買いシグナルが発生した日に，オプションを購入し，その際の損益を計算する．

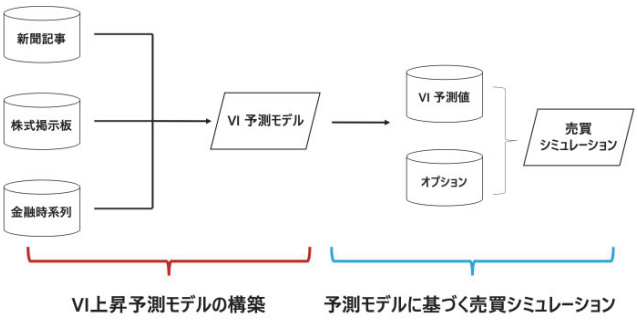

図 3: 売買シミュレーションの位置付け

4. 実験

4.1 データセット

4.1.1 日本経済新聞

本研究に使用する新聞記事のデータとして，日本経済新聞のテキストデータ *1 を用いた．2012 年 11 月 26 日から 2020 年 09 月 30 日（2,866 日）までに日本経済新聞の朝刊記事を用いる．集約した結果，2,866 日のうち 1,841 日分の新聞記事，440,416 件を獲得した．

4.1.2 Yahoo!ファイナンス掲示板

本研究に使用するソーシャルメディアのデータとして，Yahoo!ファイナンス掲示板の投稿文書を用いた．2012 年 11 月 26 日から 2020 年 09 月 30 日までに Yahoo!ファイナンス掲示板における日経平均株価スレッドに投稿された投稿文書を集約した結果，9,463,597 件であった． なお，Yahoo!ファイナンス掲示板におけるデータは，投稿文書のみを使用し，個人情報（個人を特定できる ID など）は，扱わない．

4.1.3 金融時系列データ

本研究の予測に使用する金融時系列データとして，日経平均株価，日経平均 VI の始値のデータを用いた．これらのデータは，JPX データクラウド *2 から収集した．収集期間は，2012 年 11 月 26 日から 2020 年 10 月 07 日（1920 営業日）である．各メディアの文書における収集期間と異なるのは，最後の最後の 5 営業日を投資リスクのラベル付けに使用するため，実験期間は，さらに最初の 5 営業日を除いた 2012 年 12 月 03 日から

*1 https://www.nikkeimm.co.jp/service/detail/id=225
*2 http://db-ec.jpx.co.jp/item/C430509.html

2020 年 09 月 30 日（1,910 営業日）である．一般に，日次ベースの金融商品の予測では，金融商品の営業日のみを対象としている．しかし，新聞には不定期の休刊日があり，新聞の特徴ベクトルを獲得できない日が存在する．本研究では，新聞及びソーシャルメディアの特徴ベクトル，金融時系列データの全てが存在する日のみの予測とする．ゆえに，本研究における実験期間は，2012 年 12 月 03 日から 2020 年 09 月 30 日（1841 日営業日）で，検証期間は，2017 年 02 月 08 日から 2020 年 09 月 30 日（851 営業日）である．4.3.1 で示す投資リスクの正例の日数は，128 日で，約 15.0%の不均衡なデータであることを示している．また，本研究に使用する売買シミュレーションに使用する金融時系列データとして，オプションティックデータと先物ティックデータを用いた．これらのデータも，JPX データクラウドから収集した．ティックデータは，過去全ての銘柄の売買について成立した時間や価格が記録されたデータであり，これらを用いることで，過去に遡り，理論上の仮想の取引が行うことができる．

4.2 文書の前処理

新聞記事及び株式掲示板の投稿文書における前処理として，それぞれの文書の URL，HTML，改行コードを除去した．さらに，半角ひらがな，半角カタカナを全角に変換し，全角英数字を半角に変換した．その後，形態素解析を行い，形態素に分割し，名詞，動詞，形容詞の中で数値， 非自立，代名詞，接尾でないものを抽出した. 形態素解析には，MeCab[Kudo 06] を使用し，辞書には，新語や固有表現に強い NEologd [Sato 15] を使用した．なお，NEologd は，2020 年 5 月 21 日更新分を使用した. また，ストップワードの除去を行う．ストップワードのリストには，京都大学が公開している．Slotlib[*3] のテキストデータに「ある」,「する」,「ちゃう」,「ない」,「なる」,「やる」を追加したものを用いた．

4.3 VI の上昇を予測するモデルの構築

4.3.1 VI の大幅な上昇の定義とラベリング

投資リスクとなる VI の大幅な上昇を以下のように定義し，ラベリングを行った．

投資リスクの定義

「今後，5 営業日以内に日経 VI が閾値以上上昇する」

上記の投資リスクの定義に当てはまることを正例として，データにラベリングを行った．ここでの閾値とは，VI の 1 日差分の 2σ とし，σ は次のように表される．

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

(T : 実験期間中の営業日の日数, x : 日経 VI の 1 日差分)

4.3.2 特徴量の作成

新聞記事及び株式掲示板の投稿文書から獲得した特徴ベクトル，金融時系列データ，発行数や投稿数の情報などから，先行研究と同様に以下の特徴量を生成した.（表 1）

4.3.3 機械学習の手法

4.3.2 で作成した特徴量を機械学習のアルゴリズムへの入力とし，予測を行う．本研究で使用する機械学習アルゴリズムとして，LightGBM[Ke 17] を使用する．本研究での対象は時系列データであるため，直近のデータがより重要な役割を担う可能性がある．そのため，学習期間を固定する手法（図 4）を用いる．実験期間 T 日のうち，学習期間を n 日目から m 日目

*3 http://svn.sourceforge.jp/

表 1: 追加した特徴量の種類

特徴量
1 日差, 5 日差
1 日比, 5 日比
2 日間移動平均, 5 日間移動平均
2 日間移動平均差, 5 日間移動平均差
2 日間移動平均比, 5 日間移動平均比

までの $m-n+1$ 日間として学習する．本研究では，学習期間を 2 年（490 営業日）と固定する．評価には，$m+6$ 日目の 1 日を使用する．また，本研究での投資リスクの正例の数は，約 15.0% の不均衡なデータであるため，正例と負例を 1:1 にダウンサンプリングを行い，学習する．

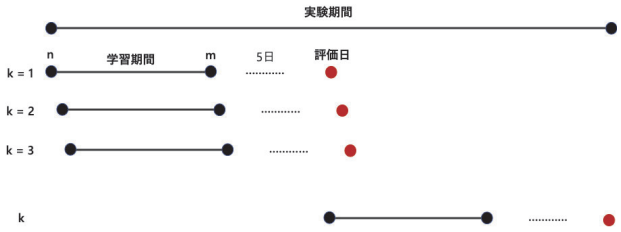

図 4: 予測モデルの方法と評価

4.4 売買シミュレーションのシナリオ

本研究における提案手法の有用性を示すために，売買シミュレーションの取引ルールについて述べる．予測モデルの出力をその日の売買指示とし，正例を買いシグナルと定義する．買いシグナルが出なかった場合には，取引は行わない．以下に，取引ルールを示す．

1. 買いシグナルが出力された日の日経 225 の始値をもとに ATM（アット・ザ・マネー）のコールオプションとプットオプションの銘柄コードを求める．
2. プットオプションとコールオプションの銘柄コードと取引日をもとに，現在値種別が始値である現在値時刻と現在値を取得．
3. 各銘柄の始値が決定された時間が，10 分以内の誤差であれば，各銘柄を 1 枚ずつ買い，ロングストラドル戦略（ポジション）を建てる．
4. 時間誤差が 10 分より大きい場合，始値がついた時間が遅い時間の銘柄をもとに，その銘柄の始値以降の時間誤差 10 分以内で，もう一方の銘柄の中値があるか探し，見つかればポジションを建てる．
5. 10 分以内で中値が見つからない場合，最も近い時間の中値をもとに，もう一方の銘柄の中値を探す．
6. 4 の作業を繰り返し，取引開始から 1 時間経っても見つからない場合，取引不成立とする．
7. ポジションを立てた日の日経 225 の始値から 500 円以上離れた日，もしくはポジション建てを行った日の次の日から，5 営業日経った日を清算日とし，ポジションを立てた際の買ったオプションを転売し，損益を確定する．

5. 結果

5.1 VI 指数上昇予測の結果

本研究における VI 指数の上昇予測の結果を表 2 に示す．評価指標として，正例に対する 適合率, 再現率, F 値を用いる．また，本研究の予測対象は，不均衡データであるため， PR-AUC も用いる．本研究では，提案手法の有効性を示すために,「金融時系列データのみで学習させたモデル」,「新聞の特徴ベクトルと金融時系列データで学習させたモデル」,「掲示板の特徴ベクトルと金融時系列データで学習させたモデル」との精度比較を行う．結果として，提案手法が最も予測精度が高い結果となった．

表 2: 各予測モデルの精度結果

モデル	適合率	再現率	F 値	PR-AUC
金融時系列のみ	0.15	0.55	0.24	0.17
新聞+金融時系列	0.16	0.66	0.25	0.17
掲示板+金融時系列	0.17	0.60	0.26	0.18
新聞+掲示板+金融	0.19	0.71	0.30	0.19

5.2 売買シミュレーションにおける結果

本研究における売買シミュレーションの結果を表 3 に示す．売買シミュレーションにおける月別の損益平均と標準偏差を表 4 に示す．本研究の有効性を示すために，前節同様のモデルとの比較を行う．また損益のベースラインとして，取引可能日に毎日取引を行うモデル（VI 予測に基づかない取引）と，仮に日経 VI 平均上昇予測の的中確率が，100%であったときの取引（正解ラベルによる売買指示）の結果も示す．

結果として，提案手法の累計損益は，+745,000 円となり，収益を得られた．また，ベースラインにおける比較として,「売買指示なし」と比較すると，-7,206,000 円の投資リスクの回避を行うことができている．一方で,「目的変数による売買指示」と比較すると，+20,550,000 円の利益を取り逃がしている．

一方で，最も収益得られたモデルは,「掲示板の特徴ベクトルと金融時系列データで学習させたモデル」であり，提案モデルよりも高い収益が確認された．これらは，表 4 の月別損益の標準偏差から，提案モデルの損益のばらつきが大きいためだと考えられる．

表 3: 各モデルにおける売買シミュレーションの結果

モデル	取引回数	累計損益（千円）
売買指示なし	833	− 6,461
正解ラベルによる売買指示	124	+21,295
金融時系列のみ	472	+2,709
新聞+金融時系列	483	− 1,991
掲示板+金融時系列	462	+3,525
新聞+掲示板+金融時系列	432	+ 745

6. おわりに

本研究では，日経平均 VI の上昇における予測に対して，マスメディアとソーシャルメディアの両方のテキスト情報に着目した新たな手法を提案した．結果として，提案モデル（新聞+掲示板+金融時系列）の精度が最も高く，日経平均 VI における上昇予測に両メディアを用いることが有効である可能性を新たに示した．また，本研究における有効性を確認するために，オプション取引によるロングストラドル戦略の売買シミュレーションを行った．結果として，本研究の提案モデルは，+745,000 円の収益を得ることが確認された．

表 4: 月別損益の平均と標準偏差

モデル	平均（千円）	標準偏差（千円）
売買指示なし	-147	1,580
正解ラベルによる売買指示	+483	883
金融時系列のみ	+62	1,143
新聞+金融時系列	-45	1,204
掲示板+金融時系列	+80	905
新聞+掲示板+金融時系列	+17	1,062

今後の課題として，主に 2 点ある．第一に，予測モデルの解釈性である．相場によって，予測に有効だったメディアやそのトピックについて，明らかにできていない．第二に，売買シミュレーションにおける詳細な分析である．今回の検証では，予測モデルの正例を売買指示とし，決済閾値を 500 円とした．売買指示を予測モデルの出力を確率値の出力を用いて調整することで，提案モデルの収益を改善できる可能性がある．

参考文献

[Fama 70] Eugene F. Fama.: Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, Vol. 25, No.2, pp. 383–417 (1970)

[Peng 16] Yangtuo Peng, Hui Jiang.: Leverage Financial News to Predict Stock Price Movements Using Word Embeddings and Deep Neural Networks. *North American Chapter of the Association for Computational Linguistics*, pp. 374-379 (2016)

[Suwa 17] 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, *2017 IEEE International Conference on Big Data*, pp. 3133-3138 (2017)

[Sasaki 20] 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. *Hawaii International Conference on System Sciences*, pp. 1-9 (2020)

[Blei 03] D. Blei, A. Y. Ng and M. Jordan.: Latent dirichlet allocation, *Journal of Machine Learning Research*, Vol.3, pp.993-1022 (2003)

[Kudo 06] T.Kudo.: Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.jp* (2006)

[Sato 15] Toshinori Sato.: Neologism dictionary based on the language resources on the web for mecab (2015)

[Ke 17] Guolin Ke, Qi Meng, Thomas Finley, et al.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree, *Advances in Neural Information Processing Systems 30* pp.3149-3157 (2017)