ヤフー株式掲示板を用いたトピック分析による VI 指数予測モデルの開発

Development of The Prediction Model of VI index by Topic Analysis based on Yahoo! JAPAN Stock BBS

佐々木皓大*[1]　梅原英一*[1]　諏訪博彦*[2]　小川祐樹*[3]　山下達雄*[4]　坪内孝太*[4]
Kodai Sasaki　Eiichi Umehara　Suwa Hirohiko　Yuki Ogawa　Tatsuo Yamashita　Kota Tsubouchi
*[1] 東京都市大学　Tokyo City University
*[2] 奈良先端科学技術大学院大学　Nara Institute of Science and Technology
*[3] 立命館大学　Ritsumeikan University
*[4] Yahoo! JAPAN 研究所　Yahoo! JAPAN Research

要旨: ソーシャルメディアを用いて，株式市場を予測する研究は数多く存在する．中でも Suwa et al.(2017)は，投資家の気持ちの変化をソーシャルメディアに投稿される話題の変化と捉え，VI 指数予測モデルを提案した．本研究は，彼らのモデルにおける課題を解決し，予測モデルの構築を目指す．彼らのモデルでは，機械学習で推定する際に使用するトレーニングデータを作る過程で，テストデータとして使用した期間を含むトピックモデルが使われている．これでは，新規投稿に対する予測精度を検証する事が出来ない．そこで本研究では，新規投稿をトピックモデルに適応させるプログラムを開発する．さらに特徴量の見直しを行い，彼らのモデルとの予測精度を比較する．

キーワード: LDA，機械学習，株式指数，ソーシャルメディア

Abstract: There have been many studies on finance that involved investigating the factors that affect the stock price index by messages posted on social networking services. Especially Suwa et al. (2017) regarded changes in the sentiment of investors as changes in topics posted on social media and proposed a volatility index (VIX) prediction model. Their model has many problems. We solve its and aim for construction of the prediction model. Their model used topic model including period used as test data in the process of making training data to be used in estimating by machine learning. We develop a program to adapt new messages to topic models. Furthermore, we improve the feature quantity. We compare prediction accuracy with their model.

Keywords: LDA, Machine Learning, Stock Index, Social Media

1. はじめに

株式の投資は，年金基金の運用の中心であり，市場の予測をする事は，高齢化社会に対応するためにも重要な問題である．株価及び株式指標の予測は，これまで多くの研究がなされている．株価は投資家心理で動く可能性がある．投資家心理が表現されるひとつの媒体として，ソーシャルメディアがある．本研究は，ソーシャルメディアを用いたトピック分析による株価指数の上昇予測モデルを構築する．

2. 先行研究

機械学習を用いた株式市場の予測において，これまで数多くの研究が行われてきた．将来の取引価格の変化を，過去に発生した時系列パターンから予想・分析するテクニカル分析では，宮崎ら[1]が，深層学習の主要なモデルの一種である CNN を用いて過去の株価の値動きを画像に落とし込み，株価予測を行った．結果として，金融データに対する CNN の予測可能性を示した．

一方で，企業や市場の状態を分析するファンダメンタル分析も多く研究が行われている．ニュースと株式市場を分析した例で，五島ら[2]は，ニュースのポジネガ度合いを定量化し，ニュースと株式市場との関連性の分析を行なった．結果として彼らは，ニュースの配信に対し株式価格が短時間の間に反応していることを確認した．他にも，菅ら[3]は，高頻度データの分析を通じ，ニュース記事が投資家行動にどのように影響するのかを分析した．その結果，ニュース記事の配信が，より短期間に投資家行動に反映される可能性を示した．

中でも，近年は SNS 等のソーシャルメディアに存在する情報を分析する試みも数多く行われている．和田ら[4]は，株式掲示板の機械学習に用いる専門辞書の開発を行なった．Antweilr et al.[5]は，株式掲示板を基づく，機械学習を用いた米国株式指標の予測を行った．彼らは，インターネット株式掲示板の投稿内容には株式価格の予想可能性はないとしながらも，投稿数の増加はその後の株式価格変動率の上昇を予測し得るとした．丸山ら[6]は，機械学習を用いて，Yahoo! Finance 掲示板内の投稿数上位５０銘柄の株式指標の予測を行った．彼らは，投稿数がボラティリティや出来高の先行指標であることを示した．諏訪ら[7]は，[6]の結果を受け，市場全体を分析対象とすることを目的として，東証 1 部における投稿数及び強気指数によるポートフォリオを構築した．結果として，強気指数が市場全体で株価リターンと関係している可能性があることを示した．また，多様な情報が発信されるソーシャルメディアとして Twitter に着目した研究も多く存在する．Sprenger et al.[8]は，株に関係のある 25 万ツイートを分析した．その結果，ツイートの感情は株価の異常な変化や翌日の株価の変化との関連があることを明らかにした．さらに，投資利益とツイートの影響力とには相関があることを示した．Bollen et al.[9]は，ツイートを 6 種類の感情レベルに分類し，“calm”な感情は 2 日後，及び 5 日後のダウ平均株価と正の相関があることを示した．ソーシャルメディアの投稿内容を対象とする分析の多くは，投稿内容を強気や弱気などの感情に分類し，相場の分析を行うセンチメント分析が主流である．しかし，ソーシャルメディアには様々な投稿がされており，感情が定かではない投稿も数多く存在する．投稿内容をより詳細に分類し，分析する方法のひとつにトピック分析がある．トピック分析は，ソーシャルメディアなどから話題を抽出するトピックモデルを用いて分析する手法である．Suwa et al.[10]は，トピック分析を用いて，恐怖指数と呼ばれる VI 指数に着目した．インターネット株式掲示板を用いた VI 指数の予測手法を提案した．彼らの提案手法は，以下である．初めに，掲示板の投稿を形態素解析した．その出力データを LDA を使い 100 種の話題に分類した．次に，時系列のトピック生成確率を計算して，機械学習による VI 指数予測を行った．彼らの提案したモデルは適合率 0.45，再現率 0.45 を得た．しかし予測精度が低い点や，新規投稿の分類に関する点を今後の課題とした．

そこで本研究は，[10]での課題を解決し，予測モデルの構築を目指す．

彼らの分析では，機械学習で推定する際に使用するトレーニングデータを作る過程で，テストデータとして使用した期間を含むトピックモデルが使われている．したがって彼らの提案手法では，過学習を起こしている．そこで本研究では，過学習を回避する手法を提案するために，新規投稿をトピックモデルに適応させるプログラムを開発する．さらに，特徴量の見直しを行う．彼らは，特徴量を作る際に，各文書のトピックに閾値を設定した．閾値を超えたトピックのみを，各文書が持っているトピックとして定義した．本研究では，各文書のトピックに閾値を設定せず，全てのトピックを各文書が持っているトピックとして定義し特徴量を作成する．それにより過学習を回避した特徴量を作成する．作成した特徴量を機械学習に用いて，予測精度を測る．

3. 構築手法

本章では，ソーシャルメディアの話題を用いて VI 指数を予測するモデルの構築手法について述べる．

3.1. 概要

開発手法の概要を図 1 に示す．ソーシャルメディアの話題は，Yahoo!JAPAN の株式掲示板から取得する．理由は，株取引について活発な投稿がなされている代表的な掲示板だからである. 取得したメッセージから，投稿の内容を表す単語群を抽出するために，形態素解析を行う．形態素解析で得られた単語群から話題を抽出するために, LAD トピックモデルを用いてトピック分析を行う．これを基に，日別のトピック所属確率を集計する．日別のトピック所属確率を特徴量として，機械学習による VI 指数の上昇予測モデルを構築し評価する．

本研究では，分析対象データを前半と後半の大きく 2 つに分類する．前半のデータで，トピックモデルを作成し，VI 指数の上昇予測モデルを構築する．後半のデータは, 前半のデータで作成したトピックモデルで, 各文書に存在するトピック所属確率を計算し，予測モデルの検証に使用する．

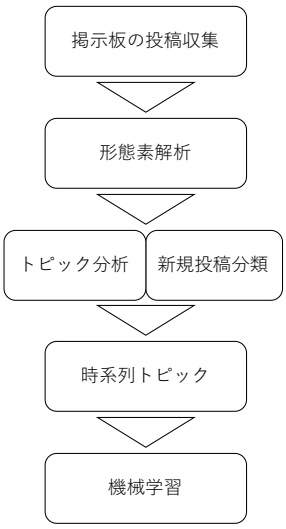

図 1 開発手法の概要

3.2. 分析対象

本研究で使用するデータは, Yahoo!JAPAN の株式掲示板内にある日経平均株価のスレッド内で投稿されたデータを用いた．分析期間は 2012 年 11 月 21 日から 2017 年 7 月 31 日である. この期間の投稿数は 4,738,275 件であった．このデータを用いて，共同研究先である Yahoo!JAPAN 研究所のサーバ内において解析した.

本研究は，日経平均 VI 指数の上昇を予測する．日経平均 VI 指数とは，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る. 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．対象期間の VI 指数を図 2 に示す.

図 2　VI 指数

3.3. トピックモデル

各投稿が，どのような話題を意味しているのかを判断するために，Blei ら[11]の LDA トピックモデルを用いる．これは，「潜在的ディリクレ配分法」と呼ばれ，文書の確率的生成モデルである．各文書には潜在トピックがあると仮定し，統計的に共起しやすい単語の集合が生成される要因を，この潜在トピックという観測できない確率変数で定式化する．本研究では，全データを前半と後半とに分ける．前半の期間は，2012 年 11 月 21 日から 2015 年 6 月 21 日とする．後半の期間は，2015 年 6 月 22 日から 2017 年 7 月 31 日とする.

3.3.1. 時系列トピック

前半のデータを使用して，トピックモデルを作成する．トピックモデルは，以下の 2 種類からなる.

- 各文書におけるトピックの所属確率
- 各単語におけるトピックの生成確率

トピック数を 100 として LDA 分析をする．トピックの所属確率を求めるためのパラメータは，一般的なテキスト分類においてデフォルトの値として用いられている，α=0.50，β=0.10 とする．1 日に投稿された文書のトピック所属確率を全て足し，その日の総投稿数で割ることにより, 日別のトピック所属確率を得る. この日別のトピック所属確率を時系列トピックとする. 時系列トピックの推移が，その日に議論された話題の推移と考えられる.

3.3.2. 新規投稿に対する分類

我々は，新規投稿をトピックに分類するために，新規投稿分類プログラムを開発した．このプログラムを用いて，3.3.1 節で求めた各単語におけるトピックの生成確率を利用して，後半のデータの各文書におけるトピックの所属確率を推定する．推定したトピックの所属確率から前半のデータと同様に，日別のトピック所属確率を得て，時系列トピックを作成する．これを検証用のデータとして使用する．予備実験として，推定性能を比較するために，推定したいトピック未知の文書ファイルを, LDA 分析に用いた文書ファイルで実行した．その結果，トピック番号の推定についてはトピック数 10 で 7~8 割，100 で 6 割程度になった．同一のデータを使用しても，推定自体が確率分布を用いるために誤差が生じてしまうと考えられる．このため，文書の中でトピックの所属確率が 1 番高いトピックをその文書のトピックとして採用した．この点に関しては，今後の研究課題である.

3.4. 機械学習

本研究では，対象期間の各日について，上昇・平穏の 2 クラスを定義し，代表的なアルゴリズムである Random Forest とロジスティック回帰を用いてモデルの構築をする.

3.4.1. 目的変数

VI 指数の上昇について，当日の VI 指数が過去 7 取引日の標準偏差より 1.5 倍以上離れた日を，VI 指数が上昇する日と定義する．それ以外の日を平穏な日と定義する．対象期間である 2012 年 11 月から 2017 年 7 月までの 1130 取引日のうち，132 日がこの定義に当て

はまる日になった．[10]の目的変数と変えた理由は，彼らの定義では，対象期間の全期間の日別変化の平均を用いているため，過去の情報だけでなく，未来の情報も含めてしまっているためである．本研究では，予測日より過去の情報のみを利用しているため，定義として妥当と考えられる．

3.4.2. 説明変数

本研究では，説明変数として，トピック投稿数及び時系列トピックと VI 指数を基に 13 のモデルを作成した．トピック投稿数及び時系列トピックは以下の 12 種類である．

- トピック投稿数
- 時系列トピック
- トピック投稿数及び時系列トピックの前日差
- トピック投稿数及び時系列トピックの前日比
- 7 取引日平均のトピック投稿数
- 7 取引日平均の時系列トピック
- 当日のトピック投稿数(時系列トピック)と過去 7 取引日平均のトピック投稿数(時系列トピック)との差
- 当日のトピック投稿数(時系列トピック)と過去 7 取引日平均のトピック投稿数(時系列トピック)との比

総投稿数および VI 指数の日別変動は以下の 12 種類であるこの 12 種類を標準特徴量と定義する．

- 総投稿数
- VI 指数
- 総投稿数及び VI 指数の前日差
- 総投稿数及び VI 指数の前日比
- 7 取引日平均の総投稿数
- 7 取引日平均の VI 指数
- 当日の総投稿数(VI 指数)と過去7取引日平均の総投稿数(VI 指数)との差
- 当日の総投稿数(VI 指数)と過去7取引日平均の総投稿数(VI 指数)との比

トピック投稿数及び時系列トピックと標準特徴量を組み合わせて以下のモデルを作成する．

1. トピック投稿数+標準特徴量
2. 時系列トピック+標準特徴量
3. トピック投稿数の前日差+標準特徴量
4. 時系列トピックの前日差+標準特徴量
5. トピック投稿数の前日比+標準特徴量
6. 時系列トピックの前日比+標準特徴量
7. 7 取引日平均のトピック投稿数+標準特徴量
8. 7 取引日平均の時系列トピック+標準特徴量
9. 当日のトピック投稿数と過去 7 取引日平均のトピック投稿数との差+標準特徴量
10. 当日の時系列トピックと過去 7 取引日平均の時系列トピックとの差+標準特徴量
11. 当日のトピック投稿数と過去 7 取引日平均のトピック投稿数との比+標準特徴量
12. 当日の時系列トピックと過去 7 取引日平均の時系列トピックとの比+標準特徴量
13. 全ての特徴量

これにより予測モデルを構築する．

4. 分析結果

推定期間の 2015 年 6 月 22 日から 2017 年 7 月 31 日の中で取引日は 518 日だった．しかし，過去 1 週間の情報を用いるため，最初の 7 取引日のデータは棄却している．対象期間の 511 日中，上昇と定義した日は 59 日であった．予測精度の結果を表 1 に示す．

表 1 予測精度の結果

	Random Forest			Logistic regression		
	適合率	再現率	F 値	適合率	再現率	F 値
1	0.87	0.89	0.84	0.81	0.71	0.75
2	0.81	0.88	0.83	0.79	0.18	0.16
3	0.84	0.88	0.83	0.79	0.81	0.80
4	0.87	0.89	0.84	0.82	0.83	0.82
5	0.78	0.88	0.83	0.79	0.81	0.80
6	0.90	0.89	0.84	0.81	0.80	0.80
7	0.86	0.89	0.84	0.79	0.70	0.74
8	0.85	0.88	0.84	0.73	0.14	0.09
9	0.86	0.89	0.84	0.80	0.82	0.81
10	0.85	0.88	0.85	0.83	0.77	0.80
11	0.90	0.89	0.84	0.80	0.83	0.81
12	0.87	0.89	0.85	0.83	0.75	0.78
13	0.90	0.89	0.84	0.81	0.72	0.76

Random Forest による最も高い適合率は，ケース 6,11,13 の 0.90 だった．しかしこれらの結果は，上昇日の予測結果ではない．上昇日の予測結果を表 2 に示す．

表 2 上昇日の予測結果

	Random Forest			Logistic regression		
	適合率	再現率	個数	適合率	再現率	個数
1	0.75	0.05	3/4	0.15	0.32	19/126
2	0.25	0.22	1/4	0.12	0.92	54/469
3	0.50	0.02	1/2	0.09	0.07	4/44
4	0.75	0.05	3/4	0.22	0.20	12/54
5	0.00	0.00	0/1	0.11	0.08	5/47
6	1.00	0.03	2/2	0.16	0.17	10/62
7	0.67	0.03	2/3	0.11	0.22	13/118
8	0.50	0.07	4/8	0.11	0.93	55/490
9	0.60	0.05	3/5	0.16	0.14	8/51
10	0.50	0.08	5/10	0.22	0.39	23/104
11	1.00	0.02	1/1	0.11	0.07	4/35
12	0.71	0.08	5/7	0.20	0.39	23/117
13	1.00	0.02	1/1	0.16	0.34	20/122

最も高い適合率は，Random Forest のケース 6,11,13 で 1.00 になった．しかし，再現率が 2%~3%と非常に低い結果になった．一方で最も高い再現率は，ロジスティック回帰のケース 2 で 0.92 だった．しかし適合率が非常に低い結果となった．適合率が 0.20 以上かつ再現率が 0.10 以上のケースを抽出すると，Random Forest のケース 2，ロジスティック回帰のケース 4,10,12 となり，その中で再現率が一番高い値で 0.39 だった．

5. 考察

機械学習によって VI 指数の予測を行なった結果，Random Forest では，ケース 6,11,13 で 1.00 になった．しかし，再現率が著しく低いため，本研究の提案手法としては採用しない．理由は，危機管理としては意味がある可能性があるが，今回はトレーディング目的のため再現率もある程度必要と考えるからである．その為，適合率が 0.20 以上かつ再現率が 0.10 以上のケースを抽出する．その結果 Random Forest のケース 2，ロジスティック回帰のケース 4,10,12 となり，その中で再現率が一番高い値は，ロジスティック回帰のケース 10,12 の 0.39 だった．本研究で採用するモデルは，ロジスティック回帰のケース 10 とする．本モデルの適合率及び再現率は，0.22，0.39 となった．一方で本研究の上昇日の比率は，0.12(59/511)である．この結果から，予測モデルが有効な可能性がある．予測精度向上については今後の課題である．このモデルを使って，廣瀬らの売買シミュレーション[12]等において投資指示を出し，利益を出すことができるかの検証を行うことが今後の研究課題である．

本研究の結果と，[10]の結果を比較すると，本研究のモデルの適合率及び再現率は低い結果となった．しかし，本研究は，過学習を回避しているため，この結果は妥当であると考えられる．

6. 結論

本研究では，トピック分析を用いて VI 指数予測モデルの構築を行った．我々は，トピック分析の際に，新規投稿のトピックを分類するプログラムを開発した．さらに正解ラベルを過去情報のみで作成した．これにより完全に過去のデータのみで予測モデルを構築することが可能となった．このプログラムを用いて，機械学習によるクラス推定を行なった結果，適合率 0.22，再現率0.39 を得た．一方で，本研究の上昇日の比率は，0.12(59/511)である．この結果によって，過学習を回避した VI 指数の予測モデルが有効な可能性がある．恐怖指数と呼ばれる VI 指数を予測することにより，株価の大きな変動を予測できる．これにより，年金基金等が使う株式リスクモデルが構築できる．しかし，このモデルを実用化するためには，まだ数多くの課題が残っている．

7. 今後の課題

今後の課題としては，第一に，予測精度の向上のため，特徴量や機械学習手法の見直しを行う必要がある．本研究では LDA トピック分析の結果を特徴量として使用した．しかし，学習期間に全く現れなかった話題で，推定期間よく議論された話題が出てきた場合に対応できない．そこで，Doc2Vec を用いる方法を検討する．Doc2Vec は，文書を低次元のベクトルに変更する手法である．新規投稿の分類も比較的容易である．さらに LDA との分類比較をした研究[13]によると，Doc2Vec の分類精度の方が高いという結果を示している．機械学習手法の見直しとしては，特徴量の計算も機械学習に任せる手法を検討する．具体的には，時系列データを扱う必要がある為，機械学習に RNN(Recurrent Neural Network)の一種である LSTM(Long short-term memory)を用いて行うことが考えられる．第二に，今回の正解ラベルの定義では，「平穏」が続いているのにも関わらず，VI 指数の全体を見ると明らかに上昇している場合に対応できていない．ゆえに，正解ラベルの定義として「上昇」「徐々に上昇」「その他」の 3 ラベルとする方法が考えられる．「徐々に上昇」は連検定等を用いて定義する方法が考えられる．

参 考 文 献

[1] 宮崎邦洋, 松尾豊, “Deep Learning を用いた株価予測の分析”, 人工知能学会全国大会論文集, Vol.31, 2017.

[2] 五島圭一, 高橋大志, 寺野隆雄, “ティックデータを用いたニュースと株価との関連性分析”, 第 30 回人工知能学会全国大会論文集, 2016.

[3] 菅愛子, 高橋大志, “高頻度データを通じたニュースと株式市場の関連性の分析”, 証券アナリストジャーナル, Vol.56, No.12, pp.15-24, 2018.

[4] 和田英一, 諏訪博彦, 小川祐樹, 太田敏澄, “専門辞書を用いたテキストマイニングによるインターネット株式掲示板の投稿分析に関する研究”, 経営情報学会秋季全国大会, pp.115-118, 2012.

[5] Antweiler, W, Frank, M, Z, “Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards”, Journal of Finance, Volume 59, No.3, pp1259-1294, 2004.

[6] 丸山健, 梅原英一, 諏訪博彦, 太田敏澄, “インターネット株式掲示板の投稿 内容と株式市場の関係”, 証券アナリストジャーナル, Vol. 46, No. 11-12, pp.110‐127, 2008.

[7] 諏訪博彦, 梅原英一, 太田敏澄, “インターネット株式掲示板の投稿内容分析に 基づくファクターモデル構築の可能性”, 人工知能学会論文誌, Vol. 27, No. 6, pp.376-383, 2012.

[8] Sprenger, T, O, Tumasjan, A, Sandner, P, G,“Tweets and Trades: the Information Content of Stock Microblogs”,

European Financial Management, Volume 20, Issue 5, pp.926–957, 2014.
[9] Bollen, J, Maoa, H, Zeng, X, “Twitter mood predicts the stock market”, Journal of Computational Science, Volume 2, Issue 1, pp.1–8, 2011.
[10] Suwa, H, Ogawa, Y, Umehara, E, Kakiki, K, Yamashita, T, Tsubouchi, K, “Develop Method to Predict the Increase in the Niei VI index”, Proceedings of The 2nd International Workshop on Application of BigData for Computational Social Science in IEEE Bigdata 2017,2017.
[11]Blei, D, Ng, A, Jordan, M, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol.3, pp. 993-1022, 2003.
[12] 廣瀬由衣, 梅原英一, “日経平均オプションを使った株式掲示板に基づくボラティリティトレーディングの売買シミュレーションの開発”, 第 24 回社会情報システム学シンポジウム講演予稿集, 3-3, 2018.
[13] Andrew M, Dai, Christopher Olah, Quoc V, “Document Embedding with Paragraph Vectors”, Proceedings of the NIPS Deep Learning Workshop, 2014.