

# LDA を用いた株式掲示板の投稿メッセージによる恐怖指数 上昇予測の提案

## Proposal of Prediction to Rise Volatility index based on Stock BBS using LDA

佐々木皓大<sup>1</sup> 諏訪博彦<sup>2</sup> 小川祐樹<sup>3</sup> 梅原英一<sup>1</sup> 山下達雄<sup>4</sup> 坪内孝太<sup>4</sup>

Kodai Sasaki<sup>1</sup>, Hirohiko Suwa<sup>2</sup>, Yuki Ogawa<sup>3</sup>, Eiichi Umehara<sup>1</sup>, Tatsuo Yamashita<sup>4</sup>, and Kota Tsubouchi<sup>4</sup>

<sup>1</sup> 東京都市大学 メディア情報学部

<sup>1</sup> Faculty of Informatics, Tokyo City University

<sup>2</sup> 奈良先端科学技術大学院大学 先端科学技術研究所

<sup>2</sup> Nara Institute of Science and Technology

<sup>3</sup> 立命館大学 情報理工学部

<sup>3</sup> Dep. of Information Science and Engineering, Ritsumeikan University

<sup>4</sup> Yahoo! JAPAN 研究所

<sup>4</sup> Yahoo! JAPAN Research

**Abstract:** There are many studies predicting a stock market using social media. Suwa et al. (2017) proposed a VI index prediction model. They assumed that changes sentiments of investors are topics change posted on social media. However, in their prediction model, the data to use verification is included in the data to use to develop their topic model. Hence, their model might be overfitting. Therefore, we propose a prediction model of VI index avoiding overfitting. We developed a program that applies new posting messages to topic models of a learning period. We created data for verification using this program. As a result, we found that a logistic regression using time series topics on the past seven trading days may predict a rise in the VI index.

## 1 はじめに

リーマンショックのような大きな相場変動の予測は、資産運用等において重要である。相場変動を表すリスク指数は、リターンの標準偏差である。リターンの標準偏差の指標として VI 指数が公表されており、VI 指数を予測することは重要である。

一方で、相場は投資家心理で動く。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。そこで本研究は、ソーシャルメディアの投稿内容を用いた VI 指数の上昇予測モデルを提案する。

## 2 先行研究

ソーシャルメディアに存在する情報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いる。丸山ら[1]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[2]は、丸山ら[1]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。山内ら[3]は、時系列トピックモデルを用いて株式市場を予測するモデルを提案した。彼らは、時間情報を考慮したトピックモデルを用いて、時間情報を持ったニュース記事に対してトピックを割り当て、記事集合内のトピックの時間発展を推定した。推定したトピックの時系列変化と東証株価指数 (TOPIX) のボラティリティとの関連を調べ、ボラティリティ時系列モデルの改善に用いた。結果、彼らはトピックのスコアの有用性を示した。Suwa et al.[4]は、インターネット株式掲示板に基づく VI 指数予測手法を提案した。彼ら

は、ソーシャルメディアのメッセージをトピックモデルにより指標化し、VI 指数の予測に用いた。Random Forest によるモデルでは、適合率 0.66 を得た。しかし再現率が 0.06 と非常に低い結果であった。一方、ロジスティック回帰でのモデルでは、適合率 0.45 及び、再現率 0.45 を得た。よって彼らは予測モデルの有効性を示した。また、推定の際に有用な特徴量としてトピック別投稿率に加え、直近 7 取引日の VI 指数の日別変動、および投稿数が有効であることを示した。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。Sasaki et al.[5]は、Suwa et al.[4]の提案した VI 指数予測モデルの妥当性を検証することを目的として、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を開発した。彼らは、オプション取引を用いて、予測モデルの上昇指示に基づきシミュレーションを行った結果、利益を出すことに成功した。これにより Suwa et al.[4]の予測モデルが有効な可能性があることを示した。彼らは、予測モデルの精度向上を課題とした。

そこで本研究は、Suwa et al. [4]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。そこで我々は、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

構築手法の概要を図 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推定し、予測モデルの検証に使用する。

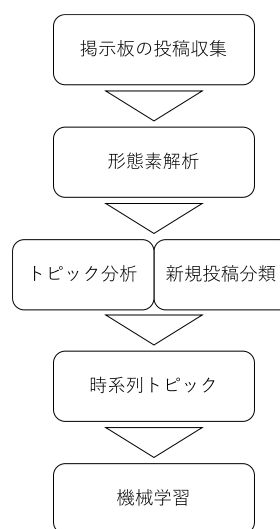


図 1 構築手法の概要

#### 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 トピック分析

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 et al.[6]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、 $\alpha=0.50$ 、 $\beta=0.10$ 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 新規投稿の分類

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。LDAは多項分布とディリクレ分布によって各データが生成されると仮定されている。しかし、本プログラムでは、一様分布によって生成されると仮定して推定を行う。

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。Suwa et al. [4]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種類を標準特微量と定義する。

- 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微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+標準特微量
2. 時系列トピック+標準特微量
3. トピック投稿数の前日差+標準特微量
4. 時系列トピックの前日差+標準特微量
5. トピック投稿数の前日比+標準特微量
6. 時系列トピックの前日比+標準特微量
7. 7取引日平均のトピック投稿数+標準特微量
8. 7取引日平均の時系列トピック+標準特微量
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+標準特微量
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+標準特微量
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+標準特微量
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+標準特微量
13. 全ての特微量

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。

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。

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

## 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、Sasaki et al.[5]の売買シミュレーション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、Suwa et al.[4]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、過学習を回避したロジスティック回帰と過去 7 取引日の時系列トピックによる VI 指数予測モデルの利用可能性を見出した。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 [7]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、今回の目的変数の定義では、VI 指数の過去 7 取引日の標準偏差を用いた。しかし「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。

## 謝辞

本研究の一部は科学研究費補助金基盤研究(C)(16K03820)の助成による。

## 参考文献

- [1] 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)
- [2] 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)
- [3] 山内海渡, 森本孝之. : 時系列トピックモデルを用いた株式市場の分析, 人工知能学会研究会 SIG-FIN-010-01, (2013)
- [4] Suwa, H, Ogawa, Y, Umehara, E, Kakiki, K, Yamashita, T, and Tsubouchi, K. : Develop Method to Predict the Increase in the Nikkei VI index, Proceedings of The 2nd International Workshop on Application of BigData for Computational Social Science in IEEE Bigdata 2017, (2017)
- [5] Sasaki, K, Hirose, Y, Umehara, E, Suwa, H, Ogawa, Y, Yamashita, T, and Tsubouchi, K. : Simulation of Volatility Trading using Nikkei Stock Index Option based on Stock

Bulletin Board, Proceedings of The Third International Workshop on Application of BigData for Computational Social Science in IEEE Bigdata 2018, (2018)

- [ 6 ] Blei, D, Ng, A, and Jordan, M. : Latent Dirichlet Allocation, Journal of Machine Learning Research, Vol.3, pp. 993-1022, (2003)
- [ 7 ] Andrew M, Dai, Christopher Olah, Quoc V. : Document Embedding with Paragraph Vectors, Proceedings of the NIPS Deep Learning Workshop, (2014)