# 株式掲示板における投稿の信頼度予測

北陸先端科学技術大学院大学
先端科学技術研究科

靱　勝彦

令和元年９月

修士論文

# 株式掲示板における投稿の信頼度予測

1630401　靱　勝彦

主指導教員　Dam Hieu Chi
審査委員主査　Dam Hieu Chi
審査委員　伊藤　泰信
神田　陽治
姜　理恵

北陸先端科学技術大学院大学
先端科学技術研究科［知識科学］
令和元年８月

# Prediction of posting reliability on stock bulletin board

Katsuhiko Utsubo

Graduate School of Advanced Science and Technology,

Japan Advanced Institute of Science and Technology

August 2019

keywords:stock bulletin board, reliability, machine learning, decision tree

Although it is difficult to predict trends in stock prices and indices, but if these trends can be predicted, they can be used as a basis for investment management. Traditionally, stock price is predicted using quantitative information, but there is also a phenomenon called "material exhaustion" that the stock price drops even if the settlement information is good. In many cases it is not possible to explain by quantitative information alone. Here, "material expenditure" means that information affecting the market is exhausted and that future stock price increases will not be visible.

In recent years, methods for analyzing stock price trends from information such as news have also been studied. This research predicts stock prices from extracted words, or predicts stock prices from emotional attributes of extracted words. These have achieved some success. However, these methods do not verify the authenticity of the information. All information is handled uniformly. There are few researches that mention the credibility of information and analyze stock market trends.

In this research, we propose a method to predict the reliability of the contents posted on the stock bulletin board, which is qualitative information. The purpose of this research is to construct a model for analysis of stock price trends using the reliability of information.

The reliability of a post is quantified by the post evaluation value given to the post, which is used as a target variable. The forecasting model used the explanatory variable as the index of stock price, price return, stock price historical volatility, and turnover. Furthermore, the evaluation value of stock brand and the post evaluation value of the poster and the negative / positive value of the post contents of the bulletin board were used as explanatory variables. We constructed a model to predict objective variables from explanatory variables using a bulletin board and stock price data. The data is from January 2015 to December 2016 as the training data, and from January 2017 to June 2017 as the validation data.

As a result of examining the relationship between each explanatory variable and the post evaluation value, the following was found. When the price return or the stock price historical volatility becomes high, the posting evaluation value of the bulletin board rises. When the turnover goes up, posting evaluation value of the bulletin board decreases. This is a situation in which the price return is high and the fluctuation is strong, that is, a situation in which it is easy for investors to obtain a profit. It is speculated that this would

mean that posts with high post evaluation values will increase.

Also, it was found that the contributors are classified into groups with high and low post evaluation values. A positive correlation was found between the posting evaluation value of the bulletin board and the posting evaluation value of the poster. Furthermore, correspondence analysis was performed using the posting evaluation value of the bulletin board and the posting evaluation value of the poster. As a result, it was found that contributors with high post evaluation value gather on a bulletin board with high post evaluation value, and contributors with low post evaluation value gather in a bulletin board with low post evaluation value.

Furthermore, the negative-positive analysis of posts by natural language processing showed a positive correlation between the reliability of posts and the negative-positive value of posts, and it was found that the post evaluation value was higher for posts with positive emotions. In order to confirm this result, the actual posting contents were extracted 5 posts in the descending order of the post evaluation value and 5 posts in the low order, and each post was visually confirmed. The posts with low post evaluation value have many dirty words and symbols, and many posts do not receive a good impression. Conversely, posts with high post evaluation value are polite sentences, and they are post contents that have a good feeling of favor. This result is consistent with negative-positive analysis.

Next, we created a model that predicts post evaluation values using a binary classification of positive or negative post evaluation values. The model was constructed using a decision tree with the contribution evaluation value as the objective variable, the stock return, stock price historical volatility, turnover , the evaluation value of stock brand, the post evaluation value of the poster, and the negative value of the content of the contribution as explanatory variables. The correct answer rate of the model is 0.756, and the F value is 0.744, which makes it possible to predict the post evaluation value. Furthermore, when the decision tree model was visualized and details of the model were confirmed, it was found that the post evaluation value is determined only by the post evaluation value of the poster. That is, a post with a high post evaluation value is predicted to have a high post evaluation value, and conversely, a post with a low post evaluation value is predicted to have a low post evaluation value.

Using this model, we evaluated the forecasting performance of the price-earnings ratio on the next day with 40 highly reliable and 40 low unreliable contributors among regular contributors. When the post sentiment attached to the post is "want to buy" or "want to buy strongly", the post predicts that the price / earnings ratio will rise the next day. Conversely, when "I want to sell" or "I want to sell strongly", the post predicts that the price / earnings ratio will decline the next day. The accuracy rate of the prediction at this time was analyzed by binary classification. As a result, it was found that the accuracy rate of prediction of a highly reliable poster is 0.566, and the accuracy rate of a low reliability poster is 0.477, and the prediction accuracy rate of a highly reliable poster is high. Furthermore, as a result of conducting a chi-square test, it was shown that this accuracy rate difference has an advantage and the accuracy rate of the prediction of a highly reliable poster is high. From this, it can be said that highly reliable information can be obtained from a highly reliable person, and the prediction performance of the stock price of the highly reliable information is high.

From the above results, the model proposed by this study shows that it is possible to predict the reliability of posts by examining the reliability of posters. Furthermore, it is

possible to predict the price return of the next day from highly reliable posts.

By using the model proposed by this research, it is possible to extract highly reliable posts as a preliminary step of analysis of qualitative data. By extracting reliable information, it is possible to contribute to investors' judgments on stock investment.

# 株式掲示板における投稿の信頼度予測

靱　勝彦

北陸先端科学技術大学院大学

先端科学技術研究科

令和 元年 8 月

キーワード:株式掲示板, 信頼度, 機械学習, 決定木

株価や指数の動向を予測することは困難であるが、この動向を予測できれば投資家への運用の判断材料になる。従来、株価の予測には定量的な情報を用いて行われているが、決算情報が良くても株価が下がる「材料出尽くし」という現象もあり、定量的な情報だけでは説明がつかない場合も多い。ここで「材料出尽くし」とは、相場に影響する情報が出尽くしてしまい、今後の株価上昇が見えないことを言う。

近年では、ニュースなどの定性情報から株価動向を分析する手法も研究されており、抽出した単語から株価を予測するものや、抽出した単語の感情属性から株価を予測するものなど、様々な手法が一定の成果を上げている。ただし、これらの手法では、定性情報の真偽を確かめることなくすべて一律に扱っており、定性情報の信頼性に言及し、株価動向を分析したものは少ないといえる。

定性情報の信頼性の研究では、フェイクニュースの信頼度を分類する研究が行われており、ニューラルネットワークを用いて情報を分類するものや、定性情報の伝播状況から情報の信頼性を分類するものや、情報の発信者の信用履歴を用いて分類するものなど様々なものがある。しかし、フェイクニュースは様々な種類があり、それぞれが異なるテキストの指標を持っていると報告するものもあり、単一のアプローチでは難しいと言える。

SNS など、コミュニケーションツールの重要性はますます高まっている。特に、個人投資家にとっては、機関投資家に比較し、情報の取得量の格差は依然として大きい。また、個人投資家は情報を得るために、知識の交換の場として掲示板などの SNS を利用することが多い。そのため、株式掲示板を分析することにより、投資家の発言としての形式知と、実際の行動としての暗黙知を、掲示板の信頼度の分析という形で、信頼度を定量的に評価することが可能となり、知識科学的に意味があると言える。

本研究では、定性情報としての株式掲示板における投稿内容の信頼性を予測する手法を提案し、情報の信頼性を踏まえた株価動向の分析への手かがりとするモデルの構築を行うことを目的とする。

投稿の信頼度は、投稿に付与された投稿評価値で定量化し、これを目的変数とし、説明変数を株価の指標である、株価収益率、株価ヒストリカル・ボラティリティ、売買代金と、 銘柄の投稿評価値、投稿者の投稿評価値、掲示板の投稿内容のネガポジ値を説明変数として予測モデルの構築を、掲示板と株価データを用いて行った。データは 2015 年 1 月から 2016 年 12 月までを学習データ、2017 年 1 月から 2017 年 6 月までを検証データとした。

それぞれの説明変数と投稿評価値の関係を調べた結果、次のことがわかった。株価収益率または株価ヒストリカル・ボラティリティが高くなると、掲示板の投稿評価値は上昇し、売買代金が高くなると掲示板の投稿評価値は減少した。これは株価収益率が高く、変動が激しい状況、すなわち投資家の利益の得やすい状況になると、投稿評価値が高い投稿が増えるということになるのではなかと推測される。

また、投稿者は投稿評価値の高いグループと低いグループに分類されることがわかった。掲示板の投稿評価値と投稿者の投稿評価値には正の相関関係が見られ、さらに、掲示板の投稿評価値と投稿者の投稿評価値でコレスポンデンス分析を行った結果、投稿評価値の高い掲示板には投稿評価値の高い投稿者が集まり、投稿評価値の低い掲示板には投稿評価値の低い投稿者が集まることがわかった。

さらに、自然言語処理による投稿のネガポジ分析から、投稿の信頼度と投稿ネガポジ値には正の相関がみられ、ポジティブな感情の投稿ほど投稿評価値が高いことがわかった。この結果を確認するために、実際の投稿内容を投稿評価値が高い順に 5 投稿、低い順に 5 投稿抽出し、それぞれの投稿を目視にて確認したところ、投稿評価値の低い投稿は、汚い単語や記号を多用しておりあまりいい印象を受けない投稿が多く、逆に投稿評価値の高い投稿は、丁寧な文章であり、好感の持てる投稿内容であり、ネガポジ分析と一致するような結果となった。

次に、投稿評価値を投稿評価値が正か負かの 2 値分類で予測するモデルを作成した。モデルは投稿評価値を目的変数とし、株価収益率、株価ヒストリカル・ボラティリティ、売買代金、投稿者の投稿評価値、投稿内容のネガポジ値を説明変数として、決定木によるモデルを構築した。作成したモデルは正解率が 0.756、F 値が 0.744 となり、投稿評価値を予測することができるモデルとなった。さらに、決定木のモデルの可視化を行い、モデルの詳細を確認したところ、投稿評価値は投稿者の投稿評価値のみによって決定されることがわかった。つまり、投稿評価値の高い投稿者の投稿は投稿評価値が高いと予測され、逆に投稿評価値の低い投稿者の投稿は投稿評価値が低いと予測される結果となった。

このモデルを用い、常連投稿者のうち、信頼度の高い 40 名と信頼度の低い 40 名で、翌日株価収益率の予測性能を検証した。投稿に付与されている投稿感情が「買いたい」「強く買いたい」のときにその投稿は翌日株価収益率が上昇すると予測しているとし、「売りたい」「強く売りたい」のときにその投稿は翌日株価収益率が下降すると予測するとした時の、予測の正解率を、2 値分類により分析した。その結果、信頼度の高い投稿者の予測の正解率が 0.566、信頼度の低い投稿者予測の正解率が 0.477 となり、信頼度の高い投稿者の予測正解率が高いことがわかった。さらに、カイ二乗検定を行った結果、この正解率差には優位性があり、信頼度の高い投稿者の予測の正解率が高いことが示された。 このことから、信頼度

の高い情報は信頼度の高い人から得ることができ、その信頼度の高い情報の株価の予測性能は高いということが言える。

以上の結果から、本研究により提案するモデルは、投稿者の投稿評価値、すなわち投稿者の信頼度を調べることにより、将来に投稿された投稿の投稿評価値、すなわち掲示板に投稿された投稿の信頼度の予測に対して有効であるといえる。さらに、その信頼度から翌日株価収益率が予測可能であることを示した。

本研究が提案するモデルを用いることにより、定性データの分析の前段階として、信頼度の高い投稿を抽出することが可能である。信頼性の高い情報を抽出することで、投資家の株式投資への判断材料に貢献することができるといえる。

# 目 次

# 図 目 次

# 表 目 次

# 第1章　はじめに

本研究では株式掲示板を分析し、掲示板の投稿内容の信頼性を推測するモデルを構築することを目的とする。

## 1.1　研究の背景と目的

株価や指数の動向を予測することは困難であるが、この動向を予測できれば投資家への運用の判断材料になる。従来、株価の予測には定量的な情報を用いて行われているが、決算情報が良くても株価が下がる「材料出尽くし」[26] という現象もあり、定量的な情報だけでは説明がつかない場合も多い。

そのため、ニュースなどの定性情報から株価動向を分析する手法も研究されており、金融経済月報を用い単語の共起関係から主要単語の抽出を行なったのちに単語をグループ化し、さらに重回帰分析を用いて市場金利を予測したもの [21] や、株価の時系列回帰式にニュース記事を主成分分析し補正項を加えて株価を予測したもの [33] などがある。

SNS などの投資家自身が発信する定性情報を用いて株価を予測する研究も行われており、Twitter[1]の書き込み内容を感情属性に分類しその属性と株価収益率には相関関係があると報告している [3] ものや、掲示板に投稿された情報から、投稿者の書き込みの感情属性と翌日の株価収益率との相関分析を行っている [24] ものなどがある。また、個別の定性情報が与える影響に関する研究においては、「株式掲示板において投資家の投稿による行動異常度を測定することにより、相場操縦行為を発見可能」と報告 [23] するものや、「SNS によりトレードの内容を開示している投資家の中でも、優秀な成績のごく一部の投資家のみ参考にされており、すべての SNS 等で開示された情報が影響を及ぼしているというわけではない」という報告 [17] もあり、特定の書き込み情報が株価動向へ影響を与える可能性があると言える。

これらの研究では、定性情報自体に関しての分析をおこなっているものがほとんどであり、定性情報が集まる場としての掲示板自体の信頼度に関する研究は行われているものは少ない。掲示板自体の信頼度を測定することにより、その掲示板に集まった定性情報が株価に与える影響度を予め予測できれば、株価の予測に対して有意義であると言える。

本研究では、株式掲示板の書き込みと株価の関係及び、掲示板における投資家行動等を分析することにより、株式掲示板の信頼度の推計方法をモデル化することを目的とし、投資家の運用判断の材料に貢献する。

[1] https://twitter.com/

## 1.2 関連研究

定性情報の信頼度とは、新聞記事やSNS等の書き込みの内容がどれくらいの割合で正確な情報であるかということで定義されるが、その情報の真偽を確かめることは困難である。例えば、近年ではフェイクニュースを見分ける手法などが研究されている [16][4] が、その手法はまだ確立されているとはいいがたい。例えば、米国 Factcheck.org[2]では、人手によるニュース記事の真偽の検査を行っている。これらの手法では大量の情報を処理することができない。

自然言語処理などの手法により自動的に情報を分類する研究も行われており、ニューラルネットワークを用いて記事を多値クラス分類するもの [10][15] などがある。また、フェイクニュースの検出に関する研究では伝播状況を研究したもの [9] 等がある。発信者の信用履歴を記事に付与し LSTM（Long Short Term Memory）にて分類を行うことにより精度が向上すると報告 [7] するものもある。日本語の定性情報の研究では、日本語のフェイクニュースの分類を行なっている Factcheck Initiative Japan[3]にて人手でタグ付けした情報を用い、それを機械学習することによりモデルを作成し、そのモデルを用いて Twitter 投稿を対象にフェイクニュースかどうかのバイナリ分類を行っている [12] ものなどがある。

また、フェイクニュースは様々な種類があり、それぞれが異なるテキストの指標を持っている [14] とするものもあり、単一のアプローチでは難しいと言える。更に、テキスト分析、ネットワーク分析、知識データベースの組み込みや、言語的、対人的、文脈的な認識を十分に活用する必要がある [5] という報告もあり、まだフェイクニュースの発見手法、すなわち情報の信頼度を分類する手法は研究途上であると言える。

## 1.3 知識科学的な意義

SNS など、コミュニケーションツールの重要性はますます高まっている。特に、個人投資家にとっては、機関投資家に比較し、情報の取得量の格差は依然として大きい。また、個人投資家は情報を得るために、知識の交換の場として掲示板などの SNS を利用することが多い。しかしながら、SNS 等においては、風評など信頼度のおける情報かどうかわからない情報で溢れかえっている。例えば、風説の流布として日経 BP 社の記事 [27] に次のようなものがある。

> 数十名のメール会員に対して投資コンサルティングを行っていた広島市の男性が、2003 年 3 月中旬に、「ナスダック・ジャパン市場に上場しているソフト開発会社ドリームテクノロジーズの存立を左右するような悪材料があるため、明日の寄り付き（証券取引所で取引される最初の売買）で同社の株式の売り注文を出して下さい」という内容のメールを会員に送信した。このメールの効果で、ドリームテクノロジーズの株価は暴落し、男性はドリームテクノロジーズの株

[2] https://www.factcheck.org/
[3] http://fij.info

を安値で入手できた。その後男性は、悪材料は偽りだったとして株式の買い戻しを指示する電子メールを送信。安くなった株が再び高値になるよう誘導した。

このように、誤った情報で株価が左右されることもあることから、情報の信頼度を定量的に評価することは有用である。

さらに、個人投資家の意見交換場である、株式掲示板を分析することにより、投資家の発言としての形式知と、実際の行動としての暗黙知を、掲示板の信頼度の分析という形で、信頼度を定量的に評価することが可能となり、知識科学的に意味があると言える。

## 1.4 リサーチクエスチョンの設定

本研究では、信頼度の推計モデルを作成するにあたり、以下の MRQ と 3 つの SRQ を定義しこれらを解明することにした。

**MRQ:「投稿の信頼度予測はどのようモデル化され、そのモデルから株価は予測できるか？」**

株式掲示板の信頼度を定量的なモデルで表現し、そのモデルの株価予測性能を評価することを最終的な目標とする。作成した信頼度モデルから、掲示板の投稿の信頼度と翌営業日株価収益率の関係性を分析することにより、信頼度モデルの株価予測性能を評価する。

**SRQ1:「株価は投稿にどのように影響を与えるか？」**

株価に関係するニュースが発表されると、掲示板の書き込みが多くなることが経験的に知られている。株価の動きと掲示板の関係を捉えることにより、掲示板の投稿への信頼度への関係性を調べる。

**SRQ2:「投稿の信頼度はどのように分類されるか？」**

投稿の信頼度は、その投稿を行う投稿者と、その投稿内容に影響されると考えられる。同じような投稿内容でも、投稿者が異なれば、その投稿への信頼度か異なるのかどうかなど、投稿の信頼度に影響を与える条件を分類する。

**SRQ3:「投稿の信頼度はどのようにモデル化されるか？」**

掲示板の投稿の信頼度は、どのような説明変数と式を用いてモデル化できるかを分析し、信頼度のモデルの性能を評価する。

## 1.5　本論文の構成

2 章では研究対象の掲示板についての説明と用語の定義を行い、分析手法についての説明を行う。

3 章ではデータの分析にあたり信頼度の定義と、対象のデータの絞り込みについて行った。

4 章では投稿内容の分析を行い信頼度を推計するための特徴量を抽出し、信頼度のモデルの作成を行った。

5 章では信頼度モデルの結果を用い、投稿の信頼度と株価収益率との関係を調べた。

6 章では本研究の結論及び今後の発展的課題への言及を行った。

# 第2章　研究対象及び分析手法

## 2.1　用語の定義

本研究で用いる用語を表 2.1 に定義する。

表 2.1: 本研究で用いる用語の定義一覧

| 用語 | 定義 |
|---|---|
| 投稿 | 日時、文章、投稿評価、投稿感情、返信、投稿者を含む |
| 投稿内容 | 自然言語によって投稿に書かれた内容 |
| 投稿者 | 掲示板に投稿を行った人 |
| リプライ | 投稿に対し、返信された投稿 |
| 投稿評価数 | そう思う、そう思わないの数 |
| 投稿評価値 | 投稿に付与された「そう思う」「そう思わない」から計算される値<br>投稿評価値 $= \frac{\text{そう思う数} - \text{そう思わない数}}{\text{そう思う数} + \text{そう思わない数}}$ |
| 投稿感情 | 投稿に付与された投稿者の感情。「強く売りたい」「売りたい」「中立」「買いたい」「強く買いたい」の 5 段階。<br>付与されていない投稿もある。 |
| 投稿感情値 | 投稿感情から作成した辞書の単語に付与された値。値が大きいほど強く買いたい、小さいほど強く売りたい単語を表す。 |
| 学習データ | 2015 年 1 月 1 日から 2016 年 12 月 31 日までのデータ |
| 検証データ | 2017 年 1 月 1 日から 2017 年 6 月 30 日までのデータ |
| 調整済み学習データ | 学習データのうち、投稿評価値が正と負となる投稿が同数となるようにランダムに抜き出した 280,000 投稿 |
| 調整済み検証データ | 検証データのうち、投稿評価値が正と負となる投稿が同数となるようにランダムに抜き出した 60,000 投稿 |
| 上位 30 銘柄 | 学習データ期間において投稿数の多い上位 30 銘柄 |
| 常連投稿者 | 上位 30 銘柄の投稿者のうち、投稿の多い上位 800 名の投稿者 |
| 投稿者別<br>平均投稿評価値 | 常連投稿者の学習データ期間のそれぞれの投稿の投稿評価値の平均値 |
| 平均月次株価収益率 | 銘柄別の月次の株価収益率の平均値 |

| 用語 | 定義 |
|---|---|
| 平均月次株価 HV | 銘柄別の月次株価収益率の 12 ヶ月ヒストリカル・ボラティリティの平均値 |
| 平均月次売買代金 | 銘柄別の月次売買代金の平均値 |
| 銘柄の平均投稿評価値 | 銘柄の学習データ期間における投稿の投稿評価値の平均値 |
| ネガポジ辞書 | 単語の感情値を表し、-1 から 1 の値をとる。0 を中立、-1 をネガティブ、1 をポジティブとし、浮動小数で単語の感情を表す。 |
| 投稿ネガポジ値 | 投稿を形態素解析し、ネガポジ辞書から単語のネガポジ値を求め、投稿別にネガポジ値の平均したもの。 |

## 2.2 株式掲示板

投資家同士の情報共有の場として、投資家個人の感情や気持ち、会話をやり取りするツールとしては SNS が有用なツールである。SNS ツールには Twitter や Facebook[1]など様々な種類のものがある。Yahoo Japan 社[2]が提供する textream 掲示板は、株式専用の掲示板であり、各株式銘柄ごとに掲示板が設置されている。さらに textream 掲示板は、投資家個人を識別する個人 ID が付与されており、各投資家の分析を行うには非常に最適なツールである。そこで本研究では textream 掲示板（以下、株式掲示板）を分析することにした。

株式掲示板の構成を図 2.1 に示す。

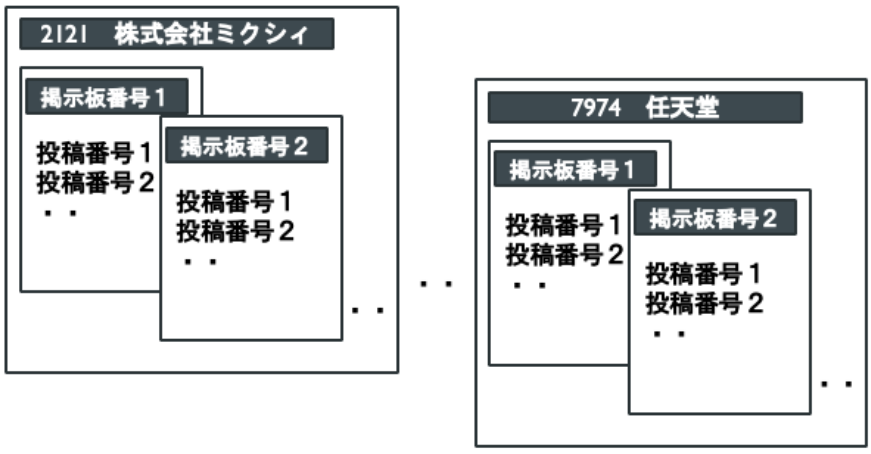


図 2.1: 株式掲示板の構成

株式掲示板は各銘柄ごとに、掲示板番号で区別される掲示板があり、掲示板の中の投稿は投稿番号で区別される。これらの、掲示板番号及び投稿番号は連番である。すなわち、掲示

[1] https://www.facebook.com/
[2] http://www.yahoo.co.jp/

板番号が若いほど古い掲示板であり、投稿番号が若いほど古い投稿である。また、投稿番号は連番であるが、投稿者自身が後ほど投稿を消すことも可能であるため、投稿番号の連番に抜けがある場合がある。

次に掲示板の例を図 2.2 に示す。投稿には、投稿番号、投稿内容、投稿日時、投稿者、投稿者感情、投稿評価、返信投稿番号が付与されている。


図 2.2: 株式掲示板の例

次に各項目の詳細な説明を示す。

①は投稿者 ID を示し、投稿を行った投稿者を表す。この投稿者 ID は全掲示板で一意となる。

②は投稿番号を示し、図 2.1 で示す投稿番号と同じである。

③は返信投稿番号を示し、この投稿番号に対する返信を示している。 1 つの投稿に複数の返信投稿を行うことができる。

④は投稿内容を示し、自然言語による投稿者の投稿を表す。投稿内容に画像を含めることができるが、画像は本研究の対象外とする。

⑤は投稿日時を示し、この投稿が投稿された日時を示す。投稿日時は分まで表示され、同じ分の場合には投稿番号が若いものが先の投稿となる。

⑥は投稿評価を示す。投稿評価とは投稿者及びその他掲示板を見ている人が評価値として各投稿に、「そう思う」もしくは「そう思わない」を付与している。ただし、必ずしも投稿に付与されているわけではない。

⑦は投稿感情を表す。投稿感情とは、投稿者が「強く売りたい」「売りたい」「様子見」「買いたい」「強く買いたい」の 5 つの感情の内どれかを投稿に付与することができる。

分析対象の株式掲示板のデータの取得にあたっては、Python2.7 を用いたプログラムを作成し、Web 画面をスクレイピングすることにより取得した。取得したデータは、HTML 形式のものであり、これを Python の XML ライブラリを用いて、個別のデータへと変換した。

変換したデータを SQLite[3]データベースに保存し利用した。SQLite とは軽量なデータベースであり、データベースの基本的な SQL 機能のみを有する。インストールも簡単であり、データの分析には適当であることから、本研究で用いることにした。

## 2.3 株価

株価データは、東京証券取引所（以下、東証）、札幌証券取引所（以下、札証）、名古屋証券取引所（以下、名証）、福岡証券取引所（以下、福証）にて、平日 9：00 から 15：00 まで（札証、名証、福証は 15：30 まで）に取引される株式の売買価格のデータから作成された 4 本値と出来高からなる。4 本値とは、取引期間のはじめについた値段（以下、始値）、取引期間中においての最高値（以下、高値）、取引期間中においての最安値（以下、安値）、取引期間の終了時の値段（以下、終値）で構成さる。売買高とは、取引された株式の数量を表す。

また、株価には日々の営業日における 4 本値と出来高を扱う日次株価と、月ごとにまとめた 4 本値と出来高を扱う月次株価が存在する。本研究においては、個別の株価の細かい動きに対する分析を行うのではなく、掲示板における全体的な特徴を得るために、株価の日々の細かい動きを排除することができる月次株価を用いた。

月次株価は取引期間を 1 ヶ月として扱い、始値は月の初めの営業日における日次株価の始値、高値はその月の最高値、安値はその月の最安値、終値は月の最終営業日の終値を示す。出来高はその月に取引された株式の総数を示す。

株価データは、分析ソフトウエア R[4]の quantmod モジュール[5]を用い日次株価を取得した。さらに Python2.7 にて作成した変換プログラムを用いて、月次株価に変換した。

[3] https://sqlite.org/
[4] https://cran.r-project.org/
[5] https://www.quantmod.com/

## 2.4　分析期間

今回収集した株式掲示板の投稿及び株価データは、2015 年 1 月 1 日から 2017 年 6 月 30 日までの期間の東証、札証、名証、福証に上場している 4267 銘柄を対象とした。

また、株式掲示板の分析にあたり、株式掲示板の日付を株式の株式の取引が行われる営業日の日付とした。東証では平日の 9 時から 15 時（札証、名証、福証は 9 時から 15 時 30 分）まで株式が売買される。掲示板の書き込みも、株式の売買状況により影響を受けることを考えると、掲示板の分析を日別で行う場合、株式売買が行われる時間と行われない時間帯で分けて考える必要がある。

本研究では簡単のため、株式売買が行われる時間帯は当日の株価に影響さることや、次営業日の株式売買が行われるまでは当日の株価が掲示板に影響することから、株式掲示板の日付を次のようにした。

前営業日の 9 時から営業日の 8 時 59 分 59 秒までを前日日付とし、9 時から翌営業日の 8 時 59 分 59 秒までを当日日付とする。

取得したデータは 2015 年 1 月 1 日から 2016 年 12 月 31 日までのデータ（以下、学習データという）と、2017 年 1 月 1 日から 2017 年 6 月 30 日までのデータ（以下、検証データという）の 2 つに分け、学習データで分析を行いモデルを作成し、検証データでモデルの確からしさを検証することにした。

## 2.5　分析手法

本研究で用いるデータマイニングの手法の説明を行う。

### 単回帰分析

回帰分析とは、2 つの変数 $x, y$ が存在するときに、$x$ を説明変数、$y$ を被説明変数とし、$x$ と $y$ の関係を調べることで、$x$ で $y$ を表現する関係を求めること [29] である。$x$ と $y$ の組で表される標本データ $((x_1,y_1),(x_2,y_2),...,(x_n,y_n))$ を、式 2.1 で示される回帰式と、各データとの $y$ 軸方向の誤差の 2 乗和が最小となるような回帰係数 $a$,$b$ を求めることで、$y$ を $x$ で表現する。

$$y = a \times x + b \tag{2.1}$$

単回帰分析を用いることで、未知の　$x$ の値から $y$ の値を推測することが可能である。標本データに対する、回帰式の当てはまりの良さを表す指標としての決定係数 $R^2$(R-squared) は式 2.2 で示される。これは、推定値の偏差平方和を標本値の偏差平方和で除したものである。

$$R^2 = \frac{\sum(\hat{y_i} - \bar{\hat{y}})^2}{\sum(y_i - \bar{y})^2} \tag{2.2}$$

$\hat{y_i}$ は回帰式から計算された推定値を、$\bar{y}$ は標本平均を表す。決定係数は 0 から 1 の値をとり、1 に近いほど当てはまりがよいことを示している。

### 重回帰分析

重回帰分析は、単回帰分析における説明変数を複数個としたものである。すなわち、被説明変数 $y$ と、説明変数 $x_1$,$x_2$,...,$x_n$ との関係を調べることである。回帰式は、式 2.3 で示され、各データとの $y$ 軸方向の誤差の 2 乗和が最小となるような回帰係数 $a_1$,$a_2$,..,$a_n$,$b$ を求めることで被説明変数を、説明変数で表現することができる。

$$y = a_1 \times x_1 + a_2 \times x_2 + ... + a_n \times x_n + b \tag{2.3}$$

決定係数 $R^2$(R-squared) は単回帰分析と同様に求めることができる。重回帰分析の場合、説明変数を増やすほど決定係数が 1 に近づくことが知られている。このため、重回帰分析においては、式 2.4 に示す自由度調整済み決定係数 $R^{*2}$(Adjusted R-squared) を用いる。

$$R^{*2} = 1 - \frac{n-1}{n-p-1}(1 - R^2) \tag{2.4}$$

ここで、$n$ は標本データ数、$p$ は説明変数の数を示す。自由度調整済み決定係数は 0 から 1 の値をとり、1 に近いほど当てはまりがよいことを示している。

### 相関係数

相関係数とは、2 つの変数 $x, y$ が存在するときに、2 種類のデータの関係を表すことであり、式 2.5 で表される。

$$\begin{aligned} cor_{xy} &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}} \\ \overline{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ \overline{y} &= \frac{\sum_{i=1}^{n} y_i}{n} \end{aligned} \tag{2.5}$$

一方の変数の増加に連れて他方の変数も増加することを「正の相関関係」といい、逆に一方の変数の増加により他方の変数が減少することを「負の相関関係」という。2 つの変数を XY 平面上で散布図を作成した際に、直線的な傾向が強いもしくは弱いとき、それぞれ強い相関、弱い相関と表現される。相関係数は－ 1 から 1 の値を取り、相関の強さの目安を表 2.2

に示す。

表 2.2: 相関係数の目安。[20] より引用

| 相関係数の値 | 相関係数の強弱 |
|---|---|
| 1〜0.7 | 強い正の相関 |
| 0.7〜0.4 | 中程度の正の相関 |
| 0.4〜0.2 | 弱い正の相関 |
| 0.2〜-0.2 | ほとんど相関なし |
| -0.2〜-0.4 | 弱い負の相関 |
| -0.4〜-0.7 | 中程度の負の相関 |
| -0.7〜-1 | 強い負の相関 |

### コレスポンデンス分析

コレスポンデンス分析とは、それぞれカテゴリ分けした 2 つのデータ間の関係を調べる際に、2 つのデータのクロス集計結果を散布図にして見みやすくし、それぞれのカテゴリ間の関係を調べるための手法である [32]。

コレスポンデンス分析を行うことにより、各カテゴリ項目間の可視化を行うことができる。その反面、各カテゴリのサンプルサイズの影響が反映されないというデメリットもある。

### k 分割交差検定

k 分割交差検定とは、機械学習の際のデータ分割の手法である。与えられた標本データを k 個に分割し、その中の 1 つを検証用データとして用い、残りを学習データとして機械学習を行う（付録 C)。この機械学習を k 回繰り返して行い、その平均を用いて 1 つの推計結果とすることで、少ない標本データで偏りが少ない機械学習結果を得ることができることが特徴である。

### 決定木

決定木とは木構造により分類を行う機械学習の手法 [30] である。木構造の節が分類の基準となる属性を表し、葉が分類されるクラスを表現する。あらかじめ与えられた学習データにより、決定木の節となる属性の条件を決定し、新たに与えられたデータで検証を行う。

また、決定木には、学習データに適合しすぎてしまい、検証データでは適合精度が悪くなってしまう、「過学習」という問題が起こる可能性があるため、決定木においてはある地点で学習を打ち切る「枝刈り」を実行する場合が多い。

決定木は、分類性能は他の機械学習手法に比べて高くない場合が多いが、機械学習で分類される条件が明快であることが特徴である。

### 分類問題の評価手法

正と負の 2 値分類を行う際に、モデルの予測結果を評価する指標として、適合率、再現率、F 値で評価を行う。予測値と正解の組み合わせは、表 2.3 の組合せとなる。

表 2.3: 予測結果と正解の関係

| | 予測 | |
|---|---|---|
| | 正 | 負 |
| 正 | TP(TruePositive) | FN(FalseNegative) |
| 負 | FP(FalsePositive) | TN(TrueNegative) |

ここで、TP は正と予測して真の値が正、FP は正と予測して真の値が負、FN は負と予測して真の値が正、TN は負と予測して真の値が負を示す。この時次の評価尺度を定義する。

- 正解率（精度、Accuracy）　正や負と予測したもののうち正解したものの割合。

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.6}$$

- 適合率（Precision）　正と予測したもののうち、実際に正であるものの割合。0 から 1 の値をとる。

$$Precision = \frac{TP}{TP + FP} \tag{2.7}$$

- 再現率（Recall）　実際に正であるもののうち、正と予測したものの割合。0 から 1 の値をとる。

$$Precision = \frac{TP}{TP + FN} \tag{2.8}$$

- F 値（Fmeasure）　適合率と再現率の調和平均、0 から 1 の値をとる。

$$Fmeasure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{2.9}$$

適合率と再現率はトレードオフの関係、すなわち適合率を上昇させようとすると再現率が下がり、また再現率を上昇させようとすると適合率が下がる関係である。これらの調和平均をとる F 値を評価することにより、分類問題におけるモデルの性能が評価可能である。F 値は 0 から 1 の値を取り、1 に近いほどモデルの予測性能が高いことを示す。

### 形態素解析

形態素解析とは、文字の並びとして文やテキストが形態素と呼ばれる意味のある最小単位や形態素から構成される語に分割し、その品詞を明らかにすること [19] である。形態素解析処理では、辞書と呼ばれる語と品詞等が予め登録された情報が必要である。語は品詞によって分類され、その品詞から、内容語と機能語に分けられる。内容語とは物事の特徴や動作、状態などを表す語であり、機能語は助詞など単体としては意味を持たない語を表す。

本研究では、機能語は使用せず、内容語のみを形態素解析にて抽出し使用した。

### センチメント分析

センチメントとは、ブログや SNS で投稿されている定性情報である、投稿内容に込められた感情を分析することを言う。感情とは「感情表現辞典』[28] によると、「喜」、「怒」、「哀」、「怖」、「恥」、「好」、「厭」、「昂」、「安」、「驚」の 10 種類から構成されている。この辞典の例文を 10 種類の感情に分類する試みなどが行われている [1]。

本研究では、感情の種類をポジティブとネガティブの 2 種類に分類する辞書を作成し、投稿のセンチメント分析を行う。また、一般的な辞書として高村ら [11] が単語感情極性対応表から作成した辞書との比較を行うことにより、辞書による違いについて評価した。単語感情極性対応表とは、岩波国語辞書をリソースとして使用し、単語の感情極性を電子のスピン方向でモデル化したものであり、単語の感情を、-1 に近いほど消極的、+1 に近いほど積極的で表現した辞書である。

# 第3章　事前調査

本章では、textream 掲示板（以下、株式掲示板）の学習期間のデータの分析を行い、東証、札証、名証、福証に上場する全 4,267 銘柄から、銘柄別の掲示板の特性、傾向などを把握し、信頼度モデルを構築するための事前調査を行とともに、信頼度モデルの目的変数となる信頼度の定義も行う。

## 3.1　掲示板データの概要

掲示板は、日々大量の投稿者の書き込みが行われ、これらすべての掲示板の投稿内容を分析することは、計算機のリソースから考えても困難であり、またデータ数の多さから、分析においてノイズが多くなり正確な分析ができなくなる場合も多いと考えられる。 そのため、データの特性を調べて、分析対象を絞って分析を行う研究結果も数多く報告されている[24][18]。本研究では機械学習や統計処理を中心としたデータの分析を行うため、分析対象のデータ量が豊富にあることが望ましいが、ノイズによる影響は避けて分析を行うために事前調査として、データの大まかな特徴を掴み、分析対象を絞ることにする。

まず、取得した掲示板のデータの集計を行い、掲示板のデータを調査した。掲示板は 2 章で示したように、掲示板の投稿ごとに、投稿内容およびそれを投稿した投稿者が存在し、投稿者は複数の掲示板に複数の投稿を行うことができる。投稿者と投稿数という観点で分けて、全体の投稿数と投稿者数を調査することで、掲示板全体の規模感を掴むことができる。そこで、掲示板の学習データの投稿数、投稿者数を調べた結果を表 3.1 に示す。

表 3.1: 学習データの投稿数、投稿者数（2015 年 1 月から 2016 年 12 月）

| 項目 | 値 |
|---|---:|
| 投稿数 | 20,095,466 |
| 投稿者数 | 216,615 |

表 3.1 から、投稿数は 20,095,466 であり、学習データが 2015 年 1 月 1 日から 2016 年 12 月 31 日までの 2 年間であり、全 4,267 銘柄であることから、1 日あたりの 1 銘柄の投稿数は平均 6.45 であることがわかる。この平均値が掲示板の投稿数を代表値として用いかどうかということを調べるため、個別の銘柄の投稿数を調べてみたところ、例えば 2015 年 4 月 1

日と 2015 年 4 月 2 日の東京電力 (銘柄コード 9501) の投稿数はそれぞれ 107 および 86、日本水産 (銘柄コード 1332) の投稿数は 1 および 0 となっていることがわかった。このことから、投稿数は銘柄や日により偏りがあり、平均値を用いて分析することは適当ではないと言え、それぞれの個別の銘柄や日別に特徴量を取る必要があることがわかった。

また、投稿者数は 216,615 であり、投稿者一人当たりの投稿数の平均は 92.8 投稿であることがわかる。この投稿者あたりの投稿数を代表値として用いていいかどうか検討するため、個別の投稿者別に投稿数を調べてみたところ、例えば投稿者 ID が”5GOawZBguzNlNykNDQ–”の投稿数は 2,184、投稿者 ID が”v1vCplR3sjEIV1x9y2I-”の投稿数は 313 であることがわかった。このことからも、投稿者によってそれぞれの投稿数が大きく違い、平均値をそのまま用いることには意味がないことがわかる。

さらに、2015 年 1 月 1 日から 2016 年 12 月 31 日までの 2 年間であることを踏まえると、すべての投稿者が毎月投稿しているとは考えにくいと考えられる。すべての投稿者が毎月投稿しているとすると、毎月の平均投稿者数は学習データ期間の投稿者数 216,615 と同じになる。また、投稿者別に投稿数が異なることを踏まえると、毎月の投稿数が一定数の投稿者がいるとは言えないとも言える。そこで、投稿者が毎月投稿調べるために、掲示板の月別の投稿者数を調べた結果を、図 3.1 に示す。

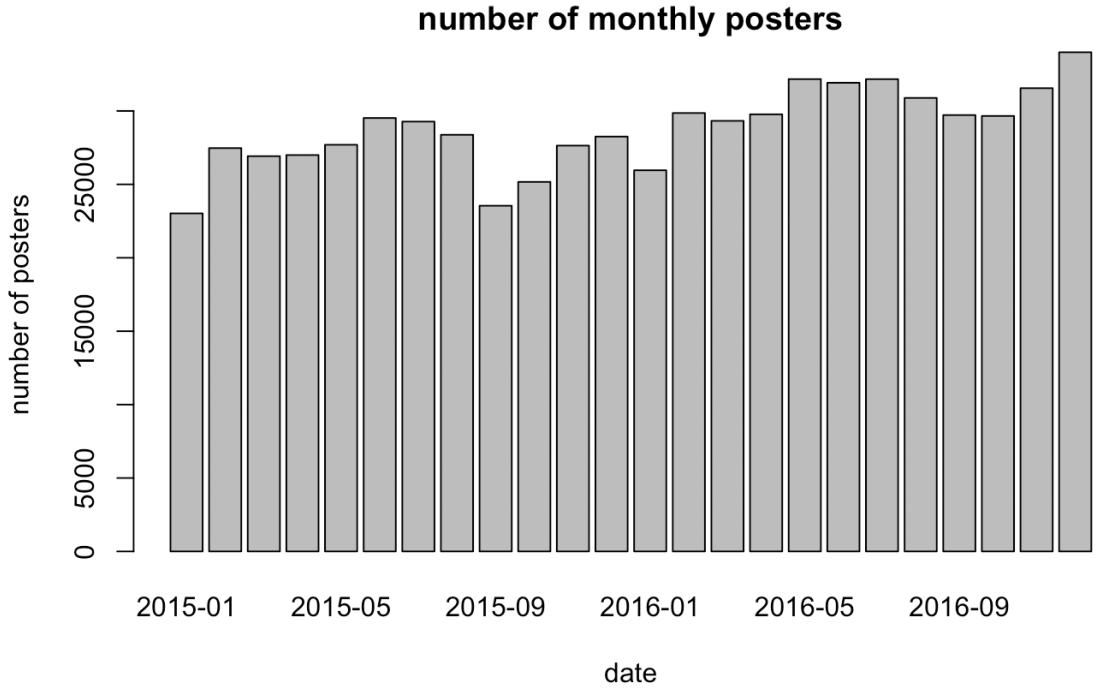


図 3.1: 掲示板への月別投稿者数（2015 年 1 月から 2016 年 12 月）

図 3.1 は、2015 年 1 月から 2016 年 12 月までの月ごとの全 4,267 銘柄への投稿者数を示したグラフである。同じ投稿者が一回でもその月に投稿した場合に 1 と数えており、月ごとに

どこかの銘柄の掲示板に投稿した投稿者の総数を表している。この図では、最も投稿者数が少ないのが 2015 年 1 月の 23,025 人、最も多いのが 2016 年 12 月の 33,998 であり、約 1.5 倍の差がある。こととことから、掲示板への投稿者数は月によって大きく異なることがわかる。投稿者数が増減するということは、投稿数も増減するのではないかと考えられる。そこで、同様に月ごとの投稿数を調べた結果を図 3.2 に示す。

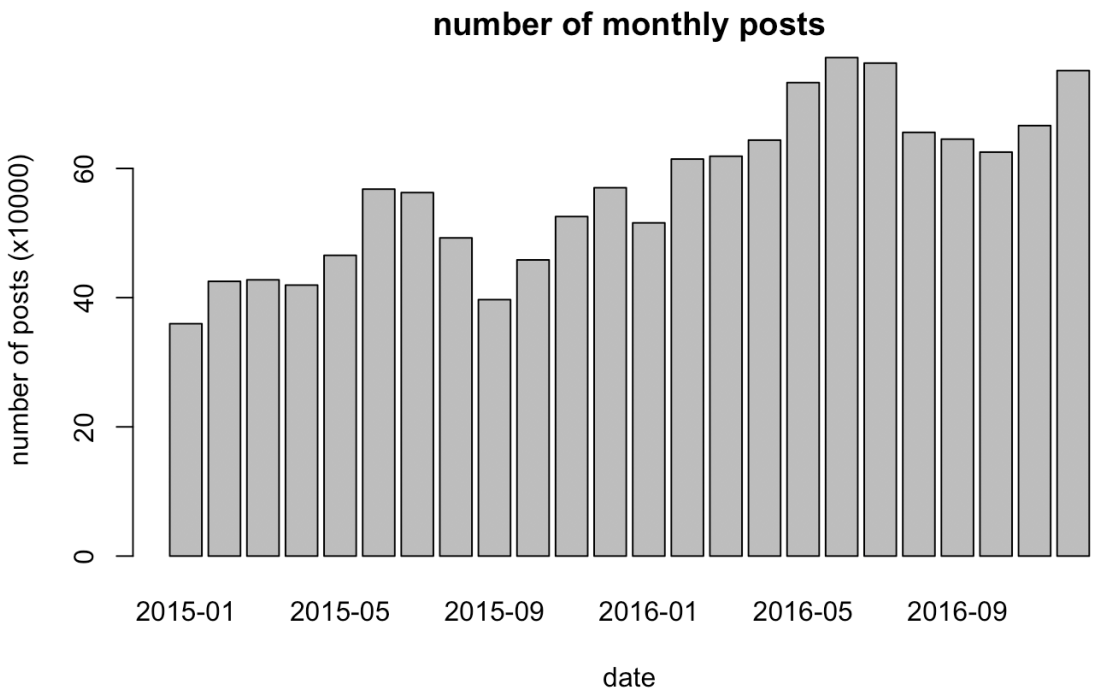


図 3.2: 掲示板への月別投稿数（2015 年 1 月から 2016 年 12 月）

図 3.2 は、2015 年 1 月から 2016 年 12 月までの月ごとの全 4,267 銘柄への投稿数を示したグラフである。この図から、月ごとに投稿数が大きく異なり、学習データ期間である 2015 年 1 月から 2016 年 12 月においては、2015 年 1 月が最も少なく約 36 万投稿であり、2016 年 6 月が約 77 万投稿となっており、ほぼ 2 倍の差があることがわかる。また、図 3.1 と図 3.2 を比較すると、同様の傾向があるように読み取れる。そこで、月別の投稿者数と投稿数の相関係数を調べた結果、0.963 の正の強い相関が見られた。

このことから、投稿者数と投稿数には相関が見られ、投稿者数が多くなると投稿数が増えることがわかった。ここで投稿者数の増減、すなわち投稿数の増減の外部要因を考えると、株価の動きに影響されているのではないかと推測される。そこで、次に株価と投稿数の関係を調べることにする。

株価は、全 4,267 銘柄の株価の動きを代表する指数である日経平均株価を用いることにした。日経平均株価とは、東証一部に上場する約 2,000 銘柄から各業種ごとに代表的な銘柄を

数銘柄づつ抽出し、全 255 銘柄で構成される指数であり、日本株式市場の株価の動きを代表していると言える。225 銘柄は毎年数銘柄づつ入れ替えが行われるが、入れ替わった際にも入れ替わる前の指数と連続性があるように調整されていることが特徴である [25]。そこで、月別投稿者数と日経平均株価の関係を調べ、散布図にした結果を図 3.3 に示す。

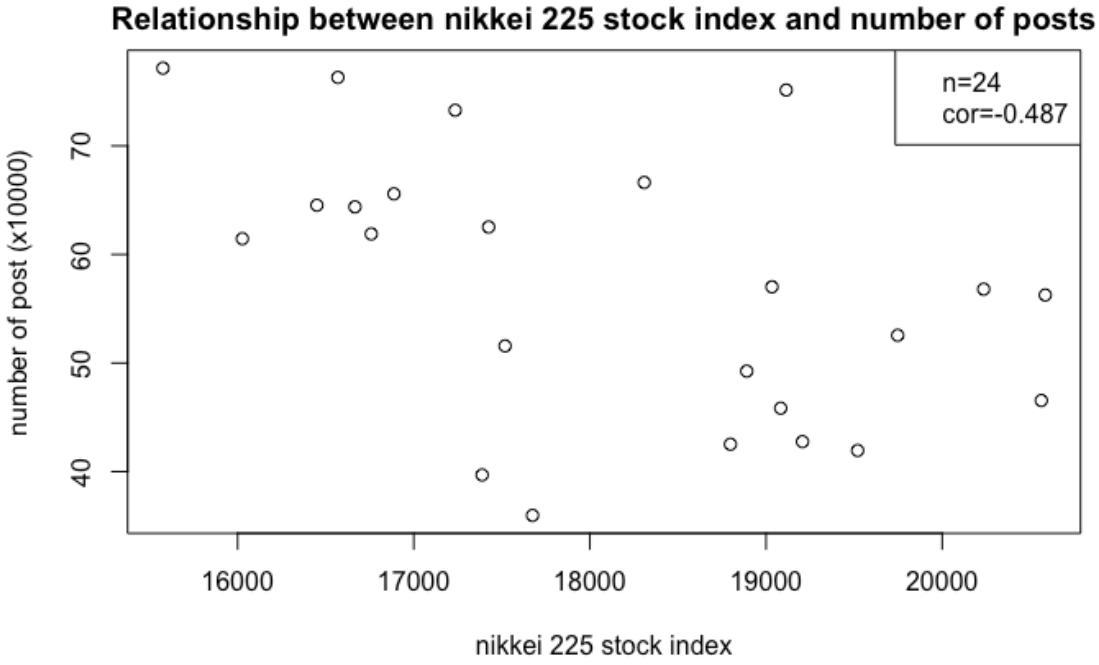


図 3.3: 掲示板への月別投稿者数と日経平均株価の関係（2015 年 1 月から 2016 年 12 月）

図 3.3 は、横軸に月次の日経平均株価、縦軸に月次の全銘柄の掲示板への投稿数を表している。図から、日経平均株価と月別投稿者数の関係には負の相関が見られる。全体の傾向として、株価が高い時には投稿数は少なく、株価が低い時に投稿数が多いことを示している。これは、株価が低い時に投稿者が多く、すなわち人が集まってくるということから、株価が低い時には人々が株式に興味があり、株価が高くなるにつれて、株式に興味がなくなっていくのではないかと推測される。この理由が正しいのであれば、個別の銘柄においても、株価が低い銘柄には掲示板への投稿が多く、株価が高い銘柄には掲示板への投稿が少なくなるのではないかと考えられる。

## 3.2 投稿数の調査

3.1 章から、株価の高低により掲示板の投稿数が影響されるのではないかと推測された。また、銘柄や日時ごとに投稿数が異なることもわかった。すなわち、掲示板によっては投稿数が人気のある掲示板、人気のない掲示板があり、掲示板ごとに投稿数が異なることが予想される。また、掲示板の信頼度を分析するために、銘柄を信頼度の指標として加えるために

は、ある程度の投稿数がないと正確に分析が行われないことが想定される。そこで、銘柄別の投稿数を調査することにより、本調査で対象とする銘柄を選別することにした。

銘柄別の投稿数の調査には、銘柄別の投稿数を集計し、その降順に並べた棒グラフと、その累積の構成比率を表現することができるパレート図を用いる。図 3.4 に全 4,267 銘柄の銘柄別の投稿数をパレート図に表わしたものを示す。

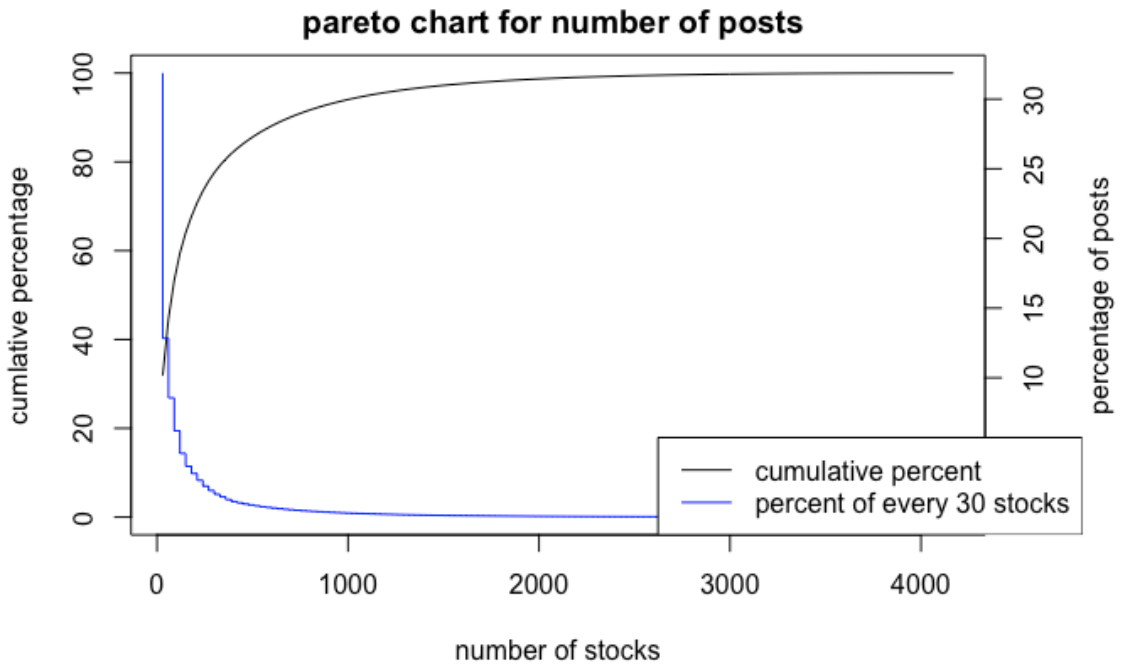


図 3.4: 銘柄の投稿数のパレート図（2015 年 1 月から 2016 年 12 月）

図 3.4 は、横軸方向に投稿を投稿数の多い順に並べ、x 軸の 0 の方向から 30 銘柄づつをまとめた投稿数の和を青線で繋いでいる。黒線は 30 銘柄ごとの青線の投稿数の累積和の比を示したものである。この図からは、黒線の左の支点は上位 30 銘柄の投稿数の総数の累積和の比を示しており、この点が約 30%の累積和の比を示している。このことから、投稿の多い上位 30 銘柄で全掲示板の投稿数の 30%を占めることを表していることがわかる。このように、掲示板への書き込みは特定の掲示板に集中する傾向があることがわかった。また、掲示板の分析にあたってはある程度のデータ量が必要なことから、本研究では、投稿数が多い上位 30 銘柄の掲示板に絞り、上位 30 銘柄に対してのデータ分析を行い、上位 30 銘柄のそれぞれの銘柄に対する掲示板の投稿の傾向を見ることにする。なお、上位 30 銘柄の銘柄名および、それぞれの銘柄への投稿数、投稿者数の一覧は付録 A に示した。

## 3.3　投稿者数の調査

Potthast[8] らは、フェイクニュースの発見手法として、大きく 3 つの分類を行なっている。その中に、Social Network Analysis というカテゴリがあり、これは SNS などにおいて、情報の伝達経路などを調べることにより、それがフェイクニュースかどうかというものを調べる手法であると定義している。この SNS における情報伝達の経路を調べるということは、すなわち、本研究における掲示板の投稿者間の経路を調べることで置き換えることが可能ではないかと考えられる。そこで、本章では投稿者を信頼度分析の指標に加えるために、投稿者の選別を行うことにする。

3.1 章で示したように、投稿者別に大きく投稿数が異なることが言えるため、ある程度の投稿数がある投稿者を、分析対象として絞り込む必要があると言える。そこで、各銘柄の掲示板あたりに参加する投稿者数とその投稿の量の関係を調べるため、学習データ期間の 2015 年 1 月から 2016 年 12 月までの月ごとに、全 4,267 銘柄の投稿者数と投稿数の関係を散布図にした結果を、図 3.5 に示す。

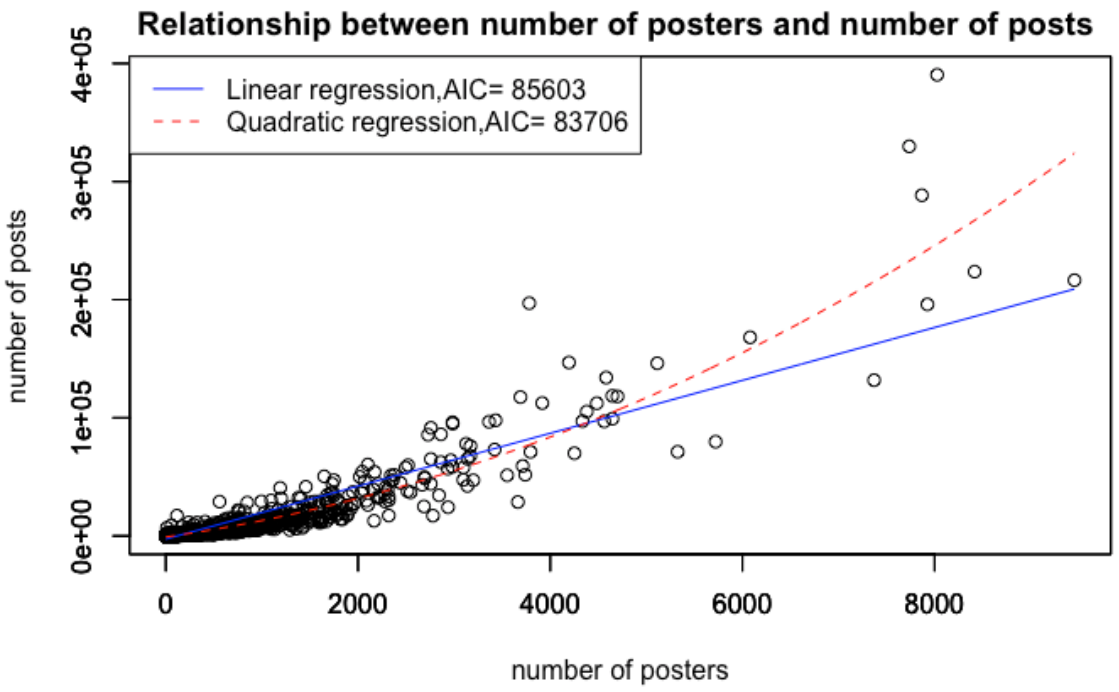


図 3.5: 銘柄別の投稿者数と投稿数の関係（2015 年 1 月から 2016 年 12 月）。図中の青線は投稿者数と投稿数の 1 次式による回帰線であり、赤点線は投稿者数と投稿数の 2 次式による回帰線である。また、それぞれの回帰式を AIC 規準量 [2] で評価した結果は、それぞれ 85,603 と 83,706 であり、この AIC 基準量が低いほどデータに対して当てはまりが良いことを示している。

図 3.5 は、横軸に投稿者数、縦軸に投稿数を示しており、銘柄別に 2015 年 1 月から 2016 年 12 月までの学習データ期間中の投稿者数と投稿数の関係をプロットした図である。また、1 次式での回帰線を青実線で、2 時式での回帰線を赤点線で示している。図から、2 次式による回帰線の AIC が 1 次式の AIC より低く、回帰線の当てはまりが良いことがわかった。すなわち、投稿者数が増えると投稿数は 2 次関数的に増える、すなわち、一人あたりの投稿数が増えることを意味すると言える。これは、投稿者数が多い銘柄の掲示板は、より投稿数が増え活発であると言える。ゆえに、投稿者数が多い掲示板、すなわち投稿数が多い掲示板を分析対象とすることは、データ量が多くなり、分析対象として適切であると言える。

銘柄は上位 30 銘柄に絞ることができたので、次に、分析対象とする投稿者を絞ることにする。上位 30 銘柄において、投稿者一人あたりの投稿数をヒストグラムにした結果を図 3.6 に示す。

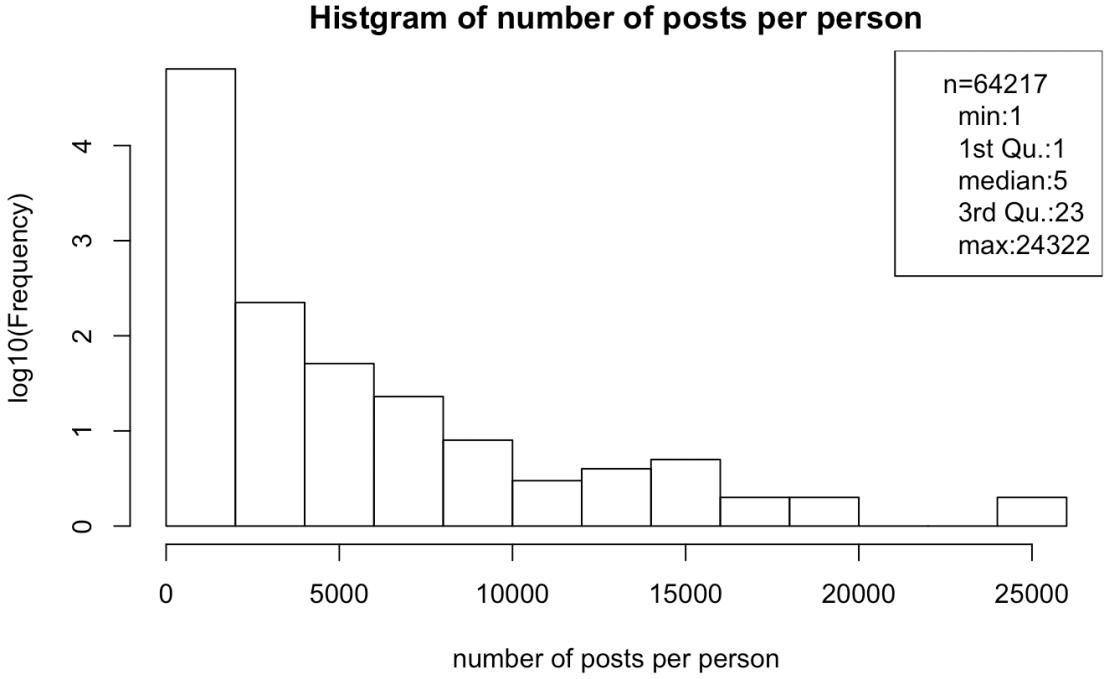


図 3.6: 投稿者一人当たりのの投稿数（2015 年 1 月から 2016 年 12 月）

図 3.6 から、投稿者一人当たりの投稿数は最小値（min）が 1、最大値（max）が 24,322 であり、かなりばらつきがある。また、データを小さい順に並び替えた際に、データの数で 4 等分した区切り線である四分位数の、25%タイルである第一四分位数（1st Qu.）は 1 となっている。このことは、データ数 64,217 のうち、1/4 である約 16,000 人は、2015 年 1 月から 2016 年 12 月を通して 1 投稿しかしていないことを示している。中央値（median）で

見ても投稿数が 5 であり、これらの少ない投稿数の投稿者を分析に含めることは適切であるとは言えない。投稿数の詳細を見るために、図 3.6 を一人あたりの投稿数別に表にした結果を、表 3.2 に示す。

表 3.2: 投稿者一人あたりの投稿数 (2015 年 1 月から 2016 年 12 月)

| 1 人あたり投稿数 | 投稿者数 | 投稿数合計 | 全投稿に対する割合 |
|---|---|---|---|
| 1 以上 | 64,217 | 4,361,251 | 1 |
| 500 以上 | 1,782 | 2,574,426 | 0.590 |
| 1000 以上 | 800 | 1,890,187 | 0.433 |
| 5000 以上 | 63 | 510,435 | 0.117 |

ここで、一人あたりの投稿数が 1,000 投稿であり、学習データの期間は 2 年であることを踏まえると、投稿者が平均的に投稿していると仮定すれば、ほぼ毎日投稿している計算となり、投稿の連続性などを見る観点から考えると、分析対象としてふさわしいと言える。表 3.2 から、1,000 投稿以上の投稿者は 800 名ではあるが、この 800 名で全投稿の 43.3%を占めることからも、分析対象としてふさわしいと考えられる。

よって、本研究の投稿者の分析対象を、1,000 投稿以上投稿している投稿者に絞ることにした。今後、この上位 800 名の投稿者のことを常連投稿者と呼ぶことにする。

次に、常連投稿者の投稿の Social Network Analysis を調査するため、投稿の返信に着目し、この返信率を調査することにした。ここで、投稿に対し返信がつくということは、その投稿を見た人がその内容に反応しているということであり、返信のつかない投稿に比較し、その投稿を見た人が内容に価値があると考えて返信しているのではないかとられる。そこで、投稿に対する返信率を常連投稿者と、常連投稿者以外の投稿者で比較した結果を表 3.3 に示す。

表 3.3: 投稿者種類別の投稿の返信率（2015 年 1 月から 2016 年 12 月）

| 投稿者種別 | 投稿数 ($A$) | 返信あり投稿数 ($B$) | 返信率（$\frac{B}{A}$） |
|---|---|---|---|
| 全投稿者 | 1,209,953 | 4,361,251 | 0.277 |
| 常連投稿者 | 584,514 | 1,771,771 | 0.330 |
| 非常連投稿者 | 625,439 | 2,589,480 | 0.242 |

表 3.3 は、常連投稿者と非常連投稿者を合わせた全投稿者、常連投稿者、非常連投稿者の 3 つの区分に分け、それぞれの投稿数と返信あり投稿数と返信率を表したものである。返信率とは、全投稿に対して、返信のあった投稿の割合を示している。全投稿者での平均返信率

は 0.277 と、投稿の約 1/4 に返信がついていることを示している。また、常連投稿者での返信率は 0.33、非常連投稿者の返信率は 0.242 と返信率に差があることが読み取れる。この差が偶然のものであるかどうかを、$\chi$ 二乗検定を行い調べた結果、有意水準 5%で常連投稿者と非常連投稿者の返信率に違いはないという帰無仮説は棄却され、常連投稿者の投稿は返信率が高いことがわかった。よって、上位 30 銘柄の常連投稿者の投稿は、非常連投稿者の投稿よりも、返信する価値があるすなわち内容があると言えることから、分析対象として、常連投稿者の投稿に絞って行うことは意味があると言える。よって、以降では上位 30 銘柄の常連投稿者の投稿を分析する。

## 3.4 信頼度の定義

投稿の信頼度を定義するにあたり、投稿に対して定量的な値が付与されている必要がある。2 章から、本研究の研究対象とした株式掲示板には、各投稿に対し、その投稿を見た人の評価である「そう思う」「そう思わない」という評価（以下、投稿評価という）が定量的に付与されている。この評価は、その投稿を見た人がそれぞれ 1 回だけ評価可能である値として付与されているものであり、その投稿の評価として定義するにはふさわしいと言える。

そこで、本研究においては、信頼度の定義として、掲示板の投稿に付与されている投稿評価を用いることにした。この投稿評価のそれぞれの値を式 3.1 で定量化することで、投稿 $p$ に対する評価を一つの値で評価可能である投稿評価値 ($EvaluationValue_p$) として定義した。

$$EvaluationValue_p = \frac{\text{そう思う数}_p - \text{そう思わない数}_p}{\text{そう思う数}_p + \text{そう思わない数}_p} \tag{3.1}$$

ここで、投稿評価は、掲示板を見た多数の参加者がその投稿内容に対し、「そう思う」、「そう思わない」という 2 値の値を付与したものであり、その掲示板の投稿においての真偽の評価を多数の参加者がその投稿評価によって行っているとみなせば、投稿評価値は投稿の信頼度であると言える。投稿評価値 $EvaluationValue_p$ は－ 1 から 1 の値を取り、1 に近いほど「そう思う」の評価が多く、-1 に近いほど「そう思わない」の評価が多くなることを意味している。

そこで、この投稿評価値を投稿の信頼度と定義し、これを予測するモデルを作成することにする。モデル化は、投稿評価値を非説明変数とし、それを予測するための説明変数を定義し、学習データをあてはめてモデル化することで可能である。投稿には投稿者、投稿内容、銘柄などの特徴量があるが、次章において、説明変数となる投稿の特徴量を選別し、信頼度モデルを構築する。

# 第4章　信頼度モデルの構築

本章においては、3 章で選別した上位 30 銘柄の 800 名の常連投稿者の投稿に対して、掲示板の投稿、投稿者、株価との関係を分析し、信頼度の予測モデルの構築を行うことを目的とする。モデルの構築に先立ち、4.1 章では銘柄別の投稿評価値と株価との関係、4.2 章では投稿評価値と常連投稿者の関係、4.3 章では投稿評価値と投稿内容の関係を調べることにより、投稿評価値を予測するための説明変数となる特徴量を抽出することを目的とする。最後に、4.4 章では、抽出した特徴量を用いてモデル式を作成し、その予測の当てはまり度を検証することで、モデルを評価する。

## 4.1　銘柄分析

### 銘柄別の信頼度

図 3.3 から、月ごとの投稿者数と日経平均株価との関係を調べた結果、株価の高低により投稿数が異なることがわかった。このことは株価が低くなっている銘柄は投稿数が多く、株価が高くなっている銘柄は投稿数が少なくなることを意味している。また、図 3.5 からは、投稿者数が増えるとその投稿数は 2 次関数的に一人あたりの投稿が増えていることを示しており、投稿数が増えるすなわち株価が低い場合には、投稿者一人あたりの影響が大きくなるといえる。このようなことから、株価が低い銘柄の場合には投稿者が多くしかも一人あたりの影響度が大きい投稿が増え、それ故に投稿評価値のばらつきが大きくなるのではないかと推測される。

そこで、銘柄別の投稿評価値を調べることにより、銘柄というと特徴量が投稿評価値への説明変数になるかどうかを確認することにした。学習データを用い、投稿に「そう思う」、「そう思わない」が付与されている投稿を抽出した。その後、それぞれの銘柄ごとに投稿の投稿評価値を集計し、銘柄別に投稿評価値の最大値、最小値を求めた。さらに平均値と四分位数を求め、これらをボックス図にした図を、図 4.1 に示す。

図 4.1 は、横軸に銘柄コードを、縦軸に投稿評価値をとり、上位 30 銘柄のそれぞれの投稿評価値のボックスと平均値を示しており、上端に近いほど投稿評価値が高く、下端に近いほど投稿評価値が低いことを意味している。また、各銘柄の赤丸に示す投稿評価値は、学習データのデータ期間（2015 年 1 月から 2016 年 12 月）における投稿評価値の平均値であり（以下、平均投稿評価値）、どの銘柄も正の値であるが銘柄ごとにその値が異なることがわかる。例えば、フューチャーベンチャーキャピタル（銘柄コード 8462）は平均投稿評価値が

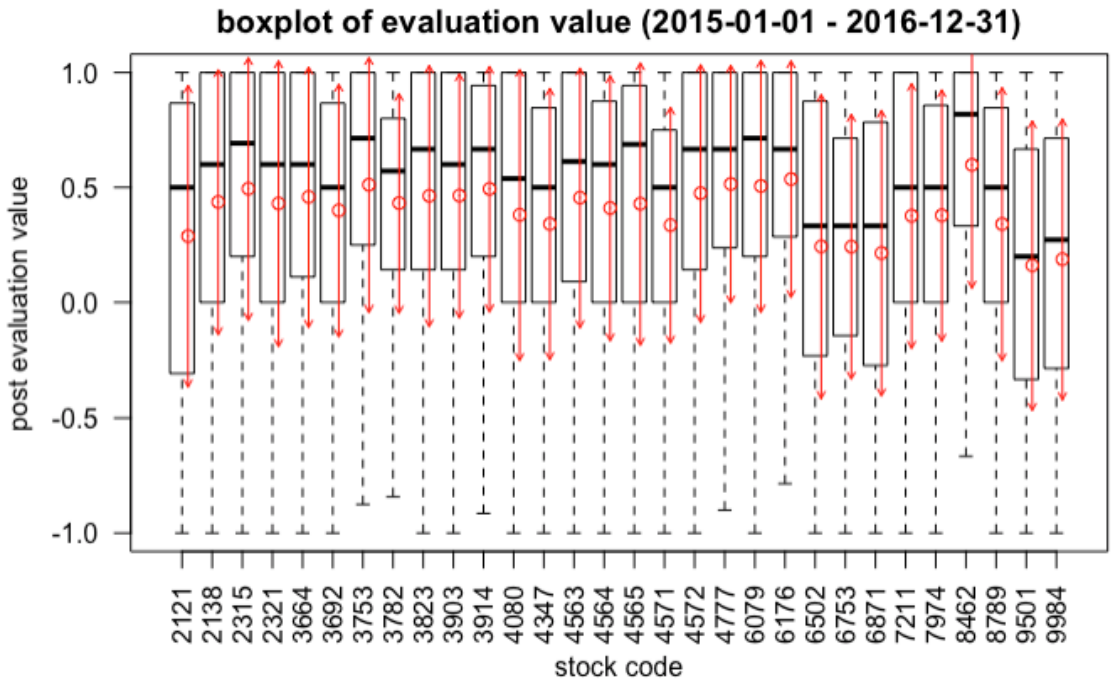


図 4.1: 銘柄別投稿評価値のボックス図（2015 年 1 月から 2016 年 12 月）。点線の上端、下端は投稿評価値の最大値、最小値を示す。ボックスの上端、下端は第一、第三四分位数を示し、横線は中央値を示す。赤丸は平均値を示し、上下へ向かう矢印は 1 標準偏差の範囲を示す。

約 0.6 程度と他の銘柄に比較し高いが、この銘柄は 2015 年 1 月の株価が 622 円、2016 年 12 月の株価が 1,649 円と、学習期間中に大きく株価が上昇している銘柄である。これに対し、東京電力（銘柄コード 9501）は平均投稿評価値が約 0.2 と他の銘柄に比較し低いが、東京電力の株価は 2015 年 1 月の株価が 502 円、2016 年 12 月の株価が 472 円となっている。また、この期間中 890 円まで上昇する局面もあったが、期間を通して株価が低く抑えられていたといえる。このように、期間中の株価の動きの違いにより、平均投稿評価値に影響が出たのではないかと考えられる。すなわち、銘柄別の平均投稿評価値は、銘柄の特徴を表しており、各投稿の投稿評価値の予想の特徴量として用いることは適切であると言える。よって、銘柄の平均投稿評価値 $StockReliability$ は投稿評価値の予測モデルの説明変数として、4.4 章で用いることにする。

### 株価指標と信頼度

株価から銘柄の特徴を表す指標は数多くあるが、その中に株価収益率と株価ヒストリカル・ボラティリティ、売買代金という 3 つの指標がある。株価収益率とは、ある期間中に株価がどれだけ変化したかという指標であり、短期で評価するときは日次の株価収益率、長期で評価するときは月次株価収益率を用いる。また、株価ヒストリカル・ボラティリティとは、ある連続した期間の株価収益率の標準偏差であり、いわゆる株価の値動きの荒さを数値化し

たものである。売買代金とは、株価の取引において売買が成立した株価とその株数をかけたものであり、実際の取引で動いた金額のことである。ある期間中の売買代金は、その銘柄の取引の活発さを表す指標となっており、売買代金が高いほど取引が活発になされていることを示している。これらの 3 つの指標は各銘柄ごとに異なり、それぞれもしくは組み合わせたものを銘柄の特徴として利用している。

図 4.1 から、銘柄ごとの期間中の株価の動きと平均投稿評価値が異なるということが推測された。株価の動きとはそれぞれの銘柄ごとに異なり、その株価の動きの指標である、株価収益率や株価ヒストリカル・ボラティリティ、売買代金という株価の指標が平均投稿評価値に影響することも考えられる。そこで、株価収益率と株価ヒストリカル・ボラティリティおよび売買代金と平均投稿評価値の関係を調べ、それらの指標が平均投稿評価値を予測するためのモデルの特徴量となるかどうかを確認する。

### 株価収益率

まず、株価収益率と平均投稿評価値の関係を調べるため、それぞれの銘柄の株価収益率を求める。本研究では長期で評価するために、月次株価収益率を求めるが、月次株価収益率は日次の株価収益率を元に次のように計算される。月次株価とは、月はじめの営業日の日次株価の始値を月次株価の始値、月中の営業日全ての日次株価の高値のもっとも高い値段を月次株価の高値、月次営業日全ての日次株価の安値のもっとも安い値段を月次株価の安値、月の最終営業日の日次株価の終値を月次株価終値とし、月中の日次株価の出来高の総和を月次株価の出来高である。

次に、月次株価収益率は次の方法で求めることができる。時系列で与えられた $t$ 月の株価を $Price_t$ とすると、$t$ 月の株価収益率 $Return_t$ は式 4.1 で表される。

$$Return_t = \frac{Price_t - Price_t}{Price_{t-1}} \tag{4.1}$$

式 4.1 で計算されるそれぞれの銘柄のそれぞれの月の月次株価収益率を、学習データの期間である 2015 年 1 月から 2016 年 12 月までの 24 個の株価収益率の平均値を平均株価収益率（以下、平均月次株価収益率）として用い、学習データ期間の平均の投稿評価値である平均投稿評価値との関係を調べることにより、銘柄ごとの平均月次株価収益率が平均投稿評価値に影響するかどうかを銘柄ごとに調べることができる。そこで、学習データ期間における平均月次株価収益率と平均投稿評価値の関係を図 4.2 に示す。

図 4.2 は、横軸に平均月次株価収益率、縦軸に平均投稿評価値を表し、各銘柄別に学習データ期間の平均月次株価収益率と平均投稿評価値の関係をプロットしたものである。図中の直線はプロットを最小二乗法による回帰直線を引いたものであり、その当てはまりの指標である自由度調整済み決定係数（Adj-R2) は、0.336 となった。また、平均月次株価収益率と平均投稿評価値の相関係数は 0.599 となり、正の相関が見られた。これらより、ばらつきは大

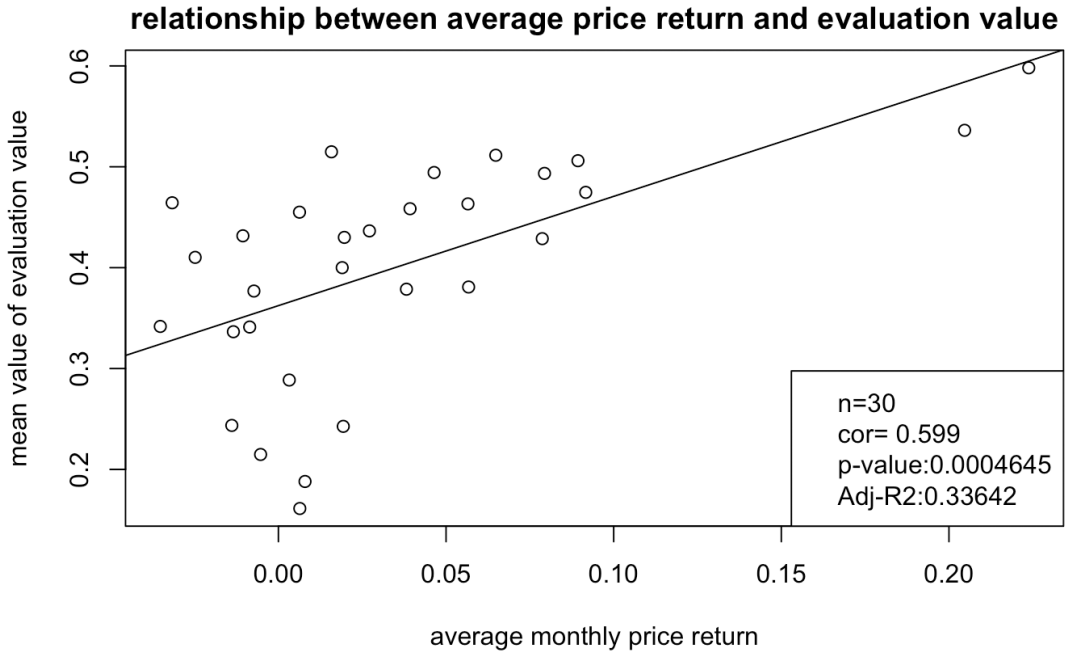


図 4.2: 平均月次株価収益率と平均投稿評価値の関係（2015 年 1 月から 2016 年 12 月）

きいものの、平均月次株価収益率が上昇すると、掲示板の平均投稿評価値が上昇することを意味している。すなわち株価が上昇すればするほど投稿に対する投稿評価値が高くなることがわかる。これは、株価が上昇することによりより信頼度の高い投稿がなされていることを意味しているといえる。例えば、株式の投資家は株価が上昇すれば利益が増える、すなわちポジティブな感情での投稿が増えるのではないかと推測され、ポジティブな感情は投稿の信頼度が高くなるのではないかと思われる。投稿の感情と投稿評価値の関係に関しては後ほど調べることにする。

以上のことから、銘柄別の平均月次株価収益率は平均投稿評価値との相関関係を持ち、投稿における投稿評価値の予測に用いる特徴量、平均月次株価収益率 $StockReturn$ として、説明変数に用いることは適切であると考えられる。この変数は、4.4 章で投稿評価値の予測モデルに用いることにする。

### 株価ヒストリカル・ボラティリティ

次に、株価ヒストリカル・ボラティリティと平均投稿評価値の関係を調べるため、月次株価ヒストリカル・ボラティリティを銘柄別に求めた。本研究では、月次株価収益率の 12 ヶ月をヒストリカル・ボラティリティとして用いることにした。この理由として、月次ヒストリカル・ボラティリティには 12 ヶ月のような短期から 72 ヶ月のような長期の期間が用いられることが多いが、学習データ期間は 2015 年 1 月から 2016 年 12 月までの 24 ヶ月であることから考えて、24 ヶ月間でも毎月の変化を細かく捉えることができる 12 ヶ月ヒストリカ

ル・ボラティリティを用いた。月次 12 ヶ月株価ヒストリカル・ボラティリティは、式 4.2 に示す株価収益率の計算期間 $term$ を 12 とした、$t$ 月の 12 ヶ月間ヒストリカル・ボラティリティ $Volat_t$ で計算した。

$$
\begin{aligned}
Volat_t &= \sqrt{\frac{12}{term-1}\sum_{k=1}^{term}(Return_k - \overline{Return_t})} \\
\overline{Return_t} &= \frac{\sum_{k=t-term}^{t}(Return_k)}{term}
\end{aligned}
\tag{4.2}
$$

ここで、$\overline{Return_t}$ は計算期間 $term$ 中の、$t$ 月の株価収益率の平均値を表している。

式 4.2 で計算されるそれぞれの銘柄のそれぞれの月の月次株価ヒストリカル・ボラティリティを、学習データの期間である 2015 年 1 月から 2016 年 12 月までの 24 個の株価ヒストリカル・ボラティリティの平均値（以下、平均月次株価 HV）を用い、学習データ期間の平均の投稿評価値である平均投稿評価値との関係を調べることにより、銘柄ごとの平均月次株価 HV が平均投稿評価値に影響するかどうかを銘柄ごとに調べることができる。そこで学習データ期間における平均月次株価 HV と平均投稿評価値の関係を図 4.3 に示す。

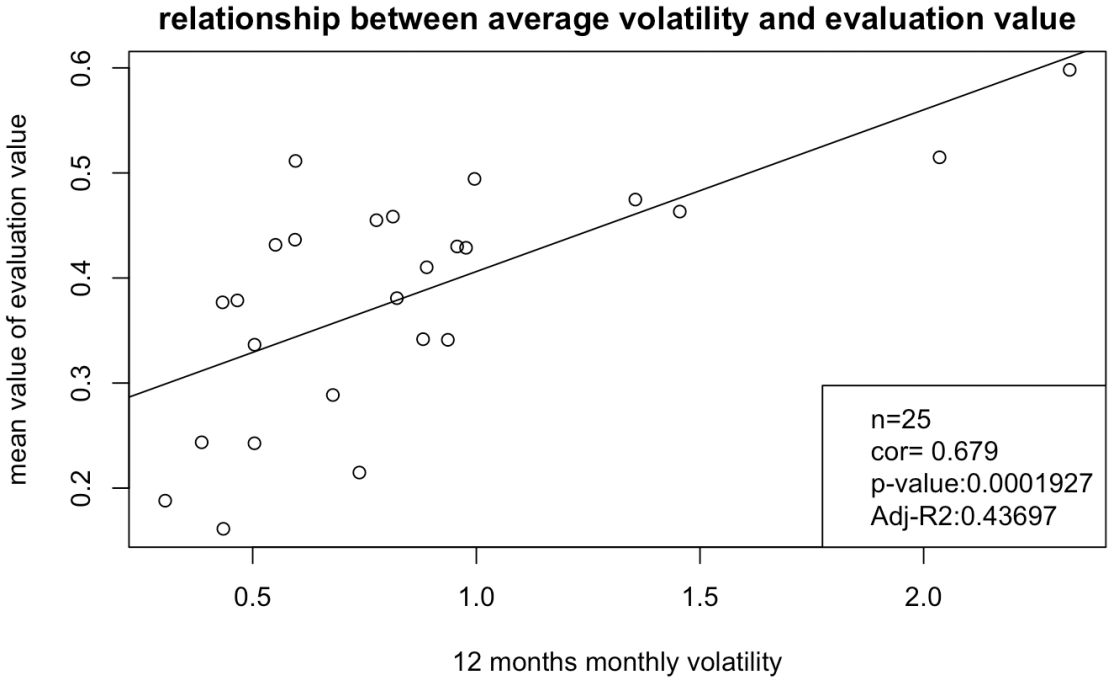


図 4.3: 平均月次株価 HV と平均投稿評価値の関係（2015 年 1 月から 2016 年 12 月）

図 4.3 は、横軸に平均月次株価 HV、縦軸に平均投稿評価値を表し、各銘柄別に学習データ期間の平均月次株価 HV と平均投稿評価値の関係をプロットしたものである。なお、12 ヶ月ヒストリカル・ボラティリティを用いているため、学習デーが期間 2015 年 1 月から 2016

年 12 月までのデータにおいて、12 ヶ月ヒストリカル・ボラティリティを計算するためには、株価が 2014 年 1 月から 2016 年 12 月まで存在しないと計算できない。そのため、本研究において平均月次株価 HV は上位 30 銘柄中 25 銘柄のみ計算したものを用いることにした。

図中の直線はプロットを最小二乗法による回帰直線を引いたものであり、その当てはまりの指標である自由度調整済み決定係数（Adj-R2) は、0.437 となった。また、平均月次株価 HV と平均投稿評価値の相関係数（cor）は 0.679 となり、正の相関が見られた。これらより、ばらつきはあるが、平均月次株価 HV が上昇すると、銘柄の平均投稿評価値が上昇することを意味している。すなわち株価の値動きが荒くなればなるほど投稿に対する投稿評価値が高くなることがわかる。これは、株価の値動きが大きいほど信頼度の高い投稿がなされていることを意味しているといえる。これは次のように推測される。投資家は株価の動きが小さい場合には、短期間の株式の売買で株価の上昇による利益（キャピタルゲイン）を出すことが難しくなる。逆に、株価の動きが大きい場合には、キャピタルゲインを得る機会が増えるということを意味している。そのため、株価の動きによる利益を得やすい銘柄の方が、よりポジティブで信頼性の高い投稿が増えるのではないかと推測される。

以上のことから、銘柄別の平均月次株価 HV は平均投稿評価値との相関関係を持ち、投稿における投稿評価値の予測に用いる特徴量、平均月次株価 HV$StockVolat$ として、説明変数に用いることは適切であると考えられる。この変数は、4.4 章で投稿評価値の予測モデルに用いることにする。

### 売買代金

最後に、売買代金と平均投稿評価値の関係を調べるため、月次売買代金を銘柄別に求めた。株価収益率や株価ヒストリカル・ボラティリティと合わせるため、月次の売買代金を計算し用いることにする。売買代金とは、取引の行われた価格とその取引量である株数の積により求められ、月間のそれらのすべての和が月次売買代金となる。しかし、本研究においては、日中の株式の取引の詳細なデータを得ることができなかったため、次の方法で月次売買代金を求め利用することにする。まず、日次の売買代金を求めるため、式 4.3 に示す $t$ 月の $d$ 番目の日の日次株価の終値 $ClosingPrice_{td}$ に、売買高 $Volume_{td}$ を乗じたものを日次売買代金 $Turnover_{td}$ を銘柄ごとに日次で求めた。

$$Turnover_{td} = ClosingPrice_{td} \times Volume_{td} \tag{4.3}$$

次に、銘柄ごとに求めた $t$ 月の日次売買代金を、式 4.4 に示すように、その月で合計したものを月次の売買代金 $Turnover_t$ として用いることにした。

$$Turnover_t = \sum (Turnover_{td}) \tag{4.4}$$

式 4.4 で計算されるそれぞれの銘柄のそれぞれの月の月次売買代金を、学習データの期間である 2015 年 1 月から 2016 年 12 月までの 24 個の月次売買代金の平均値（以下、平均月次売買代金という）を用い、学習データ期間の平均の投稿評価値である平均投稿評価値との関係を調べることにより、銘柄ごとの平均月次売買代金が平均投稿評価値に影響するかどうかを銘柄ごとに調べることができる。そこで学習データ期間における平均月次売買代金と平均投稿評価値の関係を図 4.4 に示す。

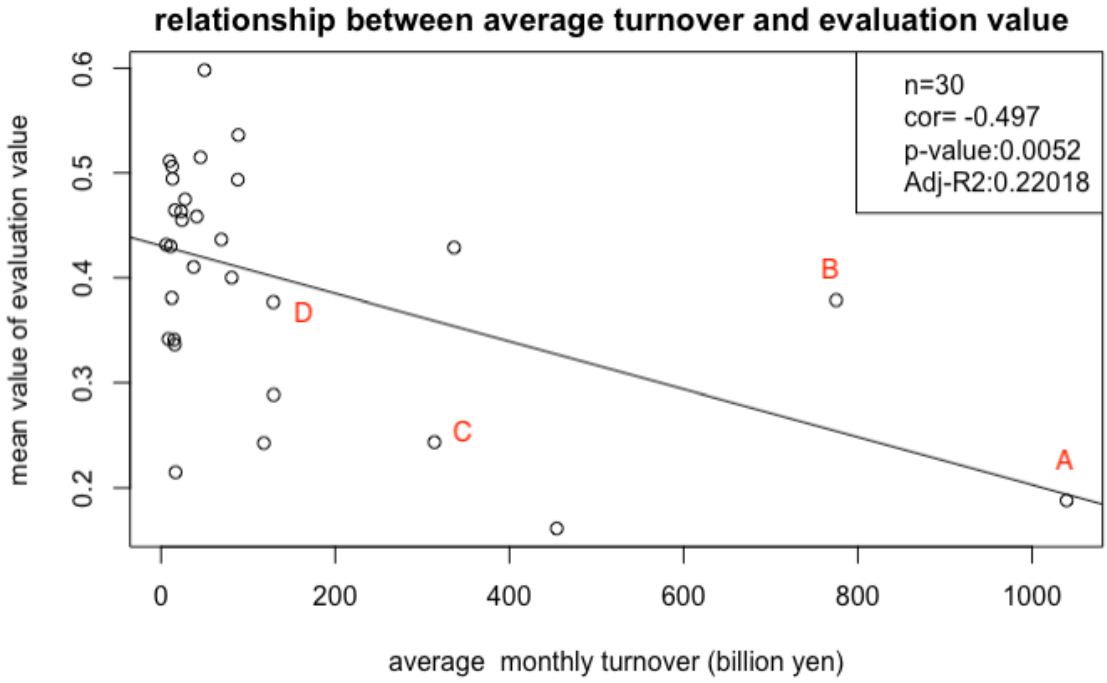


図 4.4: 売買代金と投稿評価値の関係（2015 年 1 月から 2016 年 12 月）。

図 4.4 は、横軸に平均月次売買代金、縦軸に平均投稿評価値を表し、各銘柄別に学習データ期間の平均月次売買代金と平均投稿評価値の関係をプロットしたものである。図中の直線はプロットを最小二乗法による回帰直線を引いたものであり、その当てはまりの指標である自由度調整済み決定係数（Adj-R2) は、0.220 となった。また、平均月次売買代金と平均投稿評価値の相関係数（cor）は-0.497 となり、正の相関が見られた。これらより、ばらつきは大きいが、平均月次売買代金が上昇すると、掲示板の平均投稿評価値が下落することを意味している。すなわち株が活発に取引されればされるほど、注目度が増えるほど投稿に対する投稿評価値が低くなるということになる。これを確かめるため、A、B、C、D の銘柄に着目して詳細を調べてみる。

A はソフトバンク（銘柄コード 9984）で、2015 年 1 月の株価が 6,963 円、2016 年 12 月の株価が 7,765 円であり、学習データ期間中の株価の騰落率は 11.5%と小さい。次に、B は任天堂（銘柄コード 7974）で、2015 年 1 月の株価が 11,430 円、2016 年 12 月の株価が 24,540 円であり、学習データ期間中の株価の騰落率は 115%と大きい。さらに、C は東芝（銘柄コー

ド 6501）で、2015 年 1 月の株価が 474.6 円、2016 年 12 月の株価が 283.1 円であり、学習データ期間中の株価の騰落率は▲ 40%と、株価が下がっている。

この、ABC の銘柄から見ると、上昇率の大きい B の銘柄が A の銘柄より投稿評価値が高く、騰落率が下落している C の銘柄より、若干ではあるものの A の銘柄の方が投稿評価値が高いため、本章の株価収益率と平均投稿評価値の関係からも説明がつく。しかし、D は三菱自動車（銘柄コード 7211）で、2015 年 1 月の株価が 1,004 円、2016 年 12 月の株価が 666 円であり、学習データ期間中の株価の騰落率は▲ 33.7%と大きく下落しているのにも関わらず、投稿評価値は B の銘柄とあまり変わらない結果となっている。

以上のことから、売買代金は、株価収益率と平均投稿評価値の関係からは説明のつかない特徴量を持っていると考えられるため、銘柄別の平均月次売買代金を、投稿における投稿評価値の予測に用いる特徴量 $StockTurnover$ として、説明変数に用いることは適切であると考えられる。この変数は、4.4 章で投稿評価値の予測モデルに用いることにする。

## 4.2 投稿者分析

3.3 章では、上位 30 銘柄に 1,000 投稿以上行なっている常連投稿者の返信率の分析を行い、常連投稿者の投稿は非常連投稿者の返信率より高いということがわかった。このことから、常連投稿者の投稿には返信するだけの価値のある情報が入っているのではないかと推測される。

また、Long ら [6] は、フェイクニュースの分類手法として、テキストデータを深層学習により分析する際に、著者のプロフィールをデータに混ぜることにより、フェイクニュースの分類の正解率が向上すると報告している。このことからも、投稿者の特性を調べその特徴量を投稿評価値の予測の説明変数として用いることには意味があるのではないかと推測される。そこで本章では、常連投稿者の特性を調べ、投稿評価値の予測モデルに利用する特徴量を抽出することを目的とする。

まず、800 名の常連投稿者の投稿評価値が、それぞれの投稿者別にどのような値となっているのかを分析するために、学習データ期間（2015 年 1 月から 2016 年 12 月）において、投稿の投稿評価値を、それぞれの常連投稿者別に計算し、その平均値を各常連投稿者別に算出し（以下、投稿者別平均投稿評価値という）、この値を常連投稿者別の特徴量として抽出し、評価した。

投稿者別平均投稿評価値をヒストグラムにした図を、図 4.5 に示す。

図 4.5 は、800 名の常連投稿者別の平均投稿評価値のヒストグラムを表しており、横軸の左は投稿評価値が低い常連投稿者を表し、右に行くほど投稿評価値が高い常連投稿者を表す。図から、ヒストグラムは 0 を中心とした左右対称の分布とはなっておらず、最小値が－0.7538、最大値が 0.8829、中央値が 0.4626 となっている。また、常連投稿者の投稿評価値の低い投稿者と高い投稿者が、図中の A と B の領域に存在することが読み取れる。A の領域は

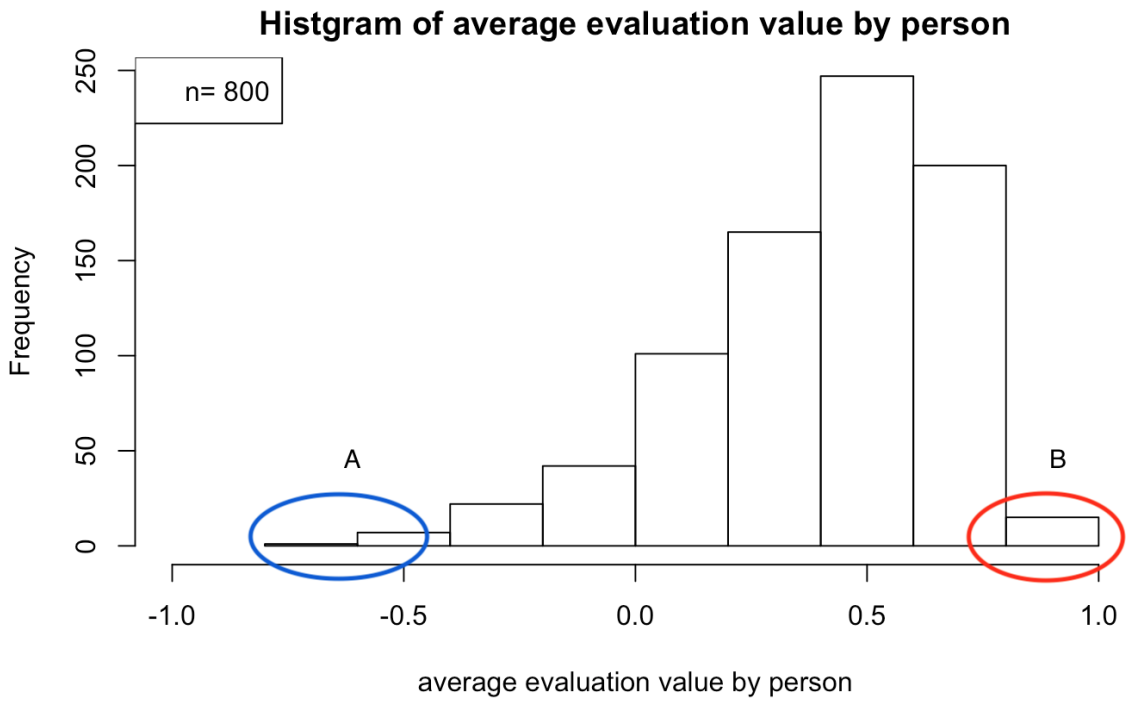


図 4.5: 常連投稿者の投稿者別平均投稿評価値のヒストグラム（2015 年 1 月から 2016 年 12 月）。A の領域は「そう思わない」と評価される投稿が多い投稿者、B の領域は「そう思う」と評価される投稿が多い投稿者を表す。中央値:0.4626

常に「そう思わない」と評価される投稿が多く、投稿者別平均投稿評価値が低い投稿者であり、B の領域は常に「そう思う」と評価される投稿が多く、投稿者別平均投稿評価値が高い投稿者であることを表していると推測される。これを調べるために個別に、最も投稿者別平均投稿評価値が高い投稿者 ID が”pOkHo5h_sS8CKupGDa3ySs5oQLI-”の投稿者（以下、高評価投稿者）と、最も投稿者別平均投稿評価値が低い投稿者 ID が”3KhnilVgsTFsbfgSZX4-”の投稿者（以下、低評価投稿者）を比較した結果を表 4.1 に示す。

表 4.1: 投稿者別平均投稿評価値の高い投稿者と低い投稿者の比較（2015 年 1 月から 2016 年 12 月）

| 投稿者 | 投稿数 | 投稿評価値 | | | |
|---|---|---|---|---|---|
| | | 第一四分位数 | 中央値 | 平均値 | 第三四分位数 |
| 高評価投稿者 | 1,140 | 0.8431 | 0.9524 | 0.8829 | 1.0000 |
| 低評価投稿者 | 1,452 | -0.8462 | -0.7576 | -0.7538 | -0.6757 |

表 4.1 から、高評価投稿者と低評価投稿者の投稿数はそれぞれ 1,140、1,452 と大きく変わらないが、平均値は 0.8829 、-0.7538 と大きく異なっている。また中央値で比較した時も、平均値と同様の結果となっている。さらに、高評価投稿者の第一四分位数は 0.8431 と評価値

が高いのに対し、低評価投稿者の第三四分位数は-0.7538 と大きく負の値となっている。これは、高評価投稿者はほぼ全ての投稿の投稿評価値が高い評価値を得ているが、低評価投稿者は全ての投稿に対して低い評価しか得ていないことを示している。

これは、投稿者別平均投稿評価値が投稿者の特徴を表しているということができる。つまり、投稿者別平均投稿評価値が低い人の投稿は常に投稿評価値が低く評価され、投稿者別平均投稿評価値が高い人の投稿は常に投稿評価値が高く評価された投稿になると推測される。

このことから、投稿者別平均投稿評価値を信頼度の予測モデルの特徴量として用いることは意味があると言える。この変数は、4.4 章で投稿者別平均投稿評価値 $UserReliability$ として予測モデルの説明変数に用いることにする。

## 4.3 投稿内容分析

4.2 章では、投稿者自身の指標である投稿者別平均投稿評価値を、投稿の信頼度の予測の説明変数として用いることが可能であると推測された。ここで、投稿者別平均投稿評価値とは、その投稿者が投稿した内容に基づいて算出された値であるということを踏まえると、投稿内容にその投稿者の特徴となる単語が存在するのではないかと考えられる。そこで、投稿内容を分析することにより、予測モデルの説明変数として用いることができるかどうか調べるために、投稿内容のテキストマイニングを行い投稿内容を定量化し、投稿評価値との関係を調べる。

投稿内容のテキストマイニングにはセンチメント分析のネガポジ判定という手法を用いて行う。センチメント分析とは、テキストを定量化する手法の一つであり、投稿の内容をいくつかの感情表現で代表させるという手法である。そのうち、ポジティブな内容であるか、ネガティブな内容であるかの 2 値分類を行う手法を本研究ではネガポジ判定と呼ぶことにし、投稿内容をネガポジ判定を行うことで定量化することにした。投稿のネガポジを判定するためには、投稿を形態素解析により単語に分割し、それぞれの単語がネガティブかポジティブかを判定する。ネガティブかポジティブかを元に、それぞれの単語を定量化した後に、それらを集計し投稿を定量化する。このネガポジ判定のためには、単語ごとにネガティブかポジティブかを判定した辞書（以下、ネガポジ辞書）を用いる必要がある。

本研究では、単語がネガティブかポジティブかの判定のために、ネガポジ辞書を作成し、それを用いて投稿のネガポジ判定を行い、それぞれの投稿のネガポジ（以下、投稿ネガポジ値）と投稿評価値との関係を調べた。

### ネガポジ辞書

ネガポジ辞書は単語ごとにネガティブかポジティブかを一対一に記述されている。ネガティブかポジティブかの判定には様々な方法があり、Turney ら [13] はインターネット上での単語の検索を用いた手法を提案しており、調べたい単語と例えばポジティブな単語である

good との出現回数とネガティブな単語である bad との出現回数の差を感情値の指標とする方法を提案している。また、高村ら [11] は、単語の感情を電子のスピン方向とみなし、シソーラスやコーパスによって構築された語彙ネットワークからネガティブな単語かポジティブな単語かを定量化している。坪内ら [22] は、本研究と同じ株式掲示板を用い、単語の抽出を N-Gram により行なったのち、「強く売りたい」、「売りたい」の投稿感情（以下、ポジティブ感情）がが付与されている投稿に現れる単語をそれぞれ-2、-1 の投稿感情値を付与し、「買いたい」、「強く買いたい」の投稿感情（以下、ネガティブ感情）が付与されている投稿に現れる単語をそれぞれ 1、2 の投稿感情値を付与し、L2 正則化回帰によりそれぞれの単語のスコアを求め、ポジティブ語のみの辞書とネガティブ語のみの辞書に分けて作成している。

本研究では、同じ掲示板を用いて分析を行っている坪内らの方法に習い、常連投稿者の 147,862 投稿の単語 (全単語数 34,365 語、内上位 80%の 3,100 語) を、L2 正則化回帰によりスコアを求め辞書を作成することにした。なお、本研究では一つの単語がネガティブとポジティブの両方に存在することを踏まえて、坪内らとは異なりネガティブとポジティブを分けずに L2 正則化回帰を行なった辞書を 1 つ作成した。なお、単語の抽出は、投稿の形態素解析を行なったのちに、ストップワード[1]と呼ばれる機能語を単語からとり除き、さらに品詞が名詞、動詞、形容詞のもののみを抽出し使用した。

形態素解析は、オープンソースプログラムである MeCab[2]を用いた。MeCab は辞書を用い、文章を形態素に分解するプログラムであり、MeCab から用いる形態素解析のための辞書は Neologd[3]を使用した。使用した Neologd は、新語に対応していることが特徴であり、週 2 回を目標に人手による更新処理が行われ、随時新語が追加されている。本研究では、研究者向けに辞書の更新を固定したバージョンである Neologd バージョン 0.0.5 を用いて形態素解析を行った。抽出した単語のネガティブ、ポジティブ、中立の単語の数を表 4.2 に示す。

表 4.2: ネガポジ辞書に含まれる単語数（2015 年 1 月から 2016 年 12 月）。L2 正則化回帰の係数が 0 より大きいものをポジティブ単語、0 より小さいものをネガティブ単語、0 のものを中立単語とした。

| ポジティブ単語数 | 中立単語数 | ネガティブ単語数 |
|---:|---:|---:|
| 1,253 | 290 | 1,057 |

表 4.2 では、L2 正則化回帰の結果、係数が 0 となったものを中立単語として数えた。その結果、ポジティブ単語 1,253 とネガティブ単語数 1,057 と、ほぼ同数の約 1,000 語となった。以降、ネガポジ辞書には、ポジティブ単語とネガティブ単語のみを使用し、投稿のネガポジ値（以下、投稿ネガポジ値）を計算する。

次に、作成したネガポジ辞書の精度を検証するため、作成したネガポジ辞書を検証データ

[1] http://www.dl.kuis.kyoto-u.ac.jp/slothlib/
[2] http://taku910.github.io/mecab/
[3] https://github.com/neologd/mecab-ipadic-neologd

（2017 年 1 月 1 日から 2017 年 6 月 30 日）の上位 30 銘柄の常連投稿者の投稿を用いて、ポジティブを正例として適合率、再現率、F 値、正解率を測定した。また、比較対象として、高村ら [11] がインターネット上で公開している一般的な単語から作成した辞書 (55,125 語)（以下、一般語ネガポジ辞書）を用い、投稿の単語から投稿がネガティブ感情であるかポジティブ感情であるかを計算し比較した結果を表 4.3 に示す。

表 4.3: 検証データによるネガポジ辞書の評価結果（2017 年 1 月から 2017 年 6 月）。作成したネガポジ辞書は一般語ネガポジ辞書より F 値が高い。[22] では Negative,Positive を分けているが本研究では分けずに作成。

| 辞書 | 適合率 | 再現率 | F 値 | 正解率 |
|---|---|---|---|---|
| 本研究で作成したネガポジ辞書 | 0.507 | 0.697 | 0.587 | 0.609 |
| 一般語ネガポジ辞書 [11] | 0.521 | 0.157 | 0.242 | 0.583 |
| （参考）[22] の結果（Negative) | 0.636 | 0.530 | 0.578 | 0.845 |
| （参考）[22] の結果（Positive) | 0.887 | 0.924 | 0.905 | |

表 4.3 から、作成したネガポジ辞書の F 値は一般的ネガポジ辞書で検証した F 値よりも高い。ここで F 値は辞書の予測性能を表し 0 から 1 の値をとり、値が大きいほど予測性能が高ことを踏まえると、一般語ネガポジ辞書よりも分類の性能が高いことを示している。また、正解率は本研究のネガポジ辞書と一般語ネガポジ辞書ではそれぞれ 0.608 と 0.583 と、若干本研究のネガポジ辞書の方が正解率が高いが、大きく性能の向上は図られなかった。これは、一般語ネガポジ辞書では、ポジティブ単語が 5,036 語に対し、ネガティブ単語が 47,615 語と大きく偏っていることに起因すると考えられる。本研究で利用した一般語ネガポジ辞書のようにポジティブ単語がネガティブ単語に比較し圧倒的に少ない場合には、本来ポジティブ単語と評価しなければいけない単語が、投稿ないから拾えていないために、再現率が低くなってしまったのではないかと推測される。

また参考として、坪内らの作成した辞書の結果と比較すると、ネガティブ単語の F 値は 0.578 と本研究のネガポジ辞書の F 値と大きく変わらないが、ポジティブ単語の F 値が 0.905 と高く、それゆえ正解率が 0.845 と本研究のネガポジ辞書より高くなっている。これは直接比較ができないので性能の差を見ることはできないが、坪内らの辞書の学習データ期間（2012 年 11 月から 2013 年 10 月）や検証データ期間（2013 年 11 月）が異なることが影響していると考えられる。

次に、作成したネガポジ辞書の単語を実際に確認するため、作成したネガポジ辞書の L2 正則化回帰係数の大きいポジティブな単語と、L2 正則化回帰係数の小さいネガティブな単語の一覧を、付録 B に示した。ポジティブな単語を確認すると、「賛成」、「生まれる」というようなポジティブな単語も見られるが、「低迷」、「手遅れ」といったあまりポジティブではなさそうな単語も見受けられ、一貫性があるとは言えない。また、ネガティブな単語を

確認すると、「甲状腺」、「汚染」といった原子力関連の単語が見受けられることや、「嵌め込む」といったネガティブな単語が見受けられる。ネガティブな単語の一覧にはポジティブな単語は見受けられない。このことから、作成したネガポジ辞書は一般的な感情辞書の単語とは違ったものとなっており、掲示板の特有の言葉に左右されていると言える。

以上のように、本研究にて作成したネガポジ辞書は、株式掲示板の特有のものとなってはいるが、ランダムに分類した場合は 0.5 程度の正解率になるのに比較し、正解率が 0.609 と若干高いことから、以降では投稿ネガポジ値の算出には、本章で作成したネガポジ辞書を用いることにする。

### ネガポジ辞書の調整

作成したネガポジ辞書は、L2 正則化回帰の係数によりネガティブ、ポジティブを決定している。これは、係数の大きさに関係なく、係数が正であるか、もしくは負であるかでその単語がポジティブであるかネガティブであるかを決定している。L2 正則化回帰係数が小さく 0 に近い場合と、非常に大きい場合ではその単語を 2 値に分ける際の正解率が変わってくるのではないかと推測される。すなわち、0 に近い係数の単語は中立的な単語も混じっていることが考えられるために、0 に近いあるしきい値以内の単語は無視したほうがいい場合があると推測される。

次に、作成したネガポジ辞書のネガティブかポジティブか判定するための L2 正則化回帰の係数のしきい値を変更させた辞書を複数作成し、その正解率を見ることで、最も正解率のよいしきい値を持ったネガポジ辞書に調整することを目的とする。

最適な L2 正則化回帰の係数のしきい値を求めるためには単語に付与されたネガポジ値の分布を確認し適切な位置でのしきい値を決める必要がある。辞書内の全ての単語のネガポジ値の分布を、ヒストグラムにより確認するために、ネガポジ辞書の単語に付与された L2 正則化回帰の係数のヒストグラムを、図 4.6 に示す。

図 4.6 は、横軸に L2 正則化回帰の係数を示しており、このヒストグラムからは、0 より若干小さい部分に大きな山があることが読み取れる。また、ヒストグラムの両端にも山が作成されており、L2 正則化回帰の係数にはばらつきがあることを示している。ここで、L2 正則化回帰の係数が 0 に近い単語は、ネガティブ感情とポジティブ感情の両方から利用されているということを踏まえると、より 0 から遠い係数を有する単語の方が、単語自体がポジティブ感情もしくはネガティブ感情の投稿からのみ利用されていると言える。そのため、係数が 0 近辺の単語を排除するようなしきい値を設けて辞書を調整し、その精度を評価することで、より精度の高い辞書に調整できるものと言える。

そこで、しきい値ごとのネガポジ辞書の精度を評価するために、ネガポジ辞書を作成した

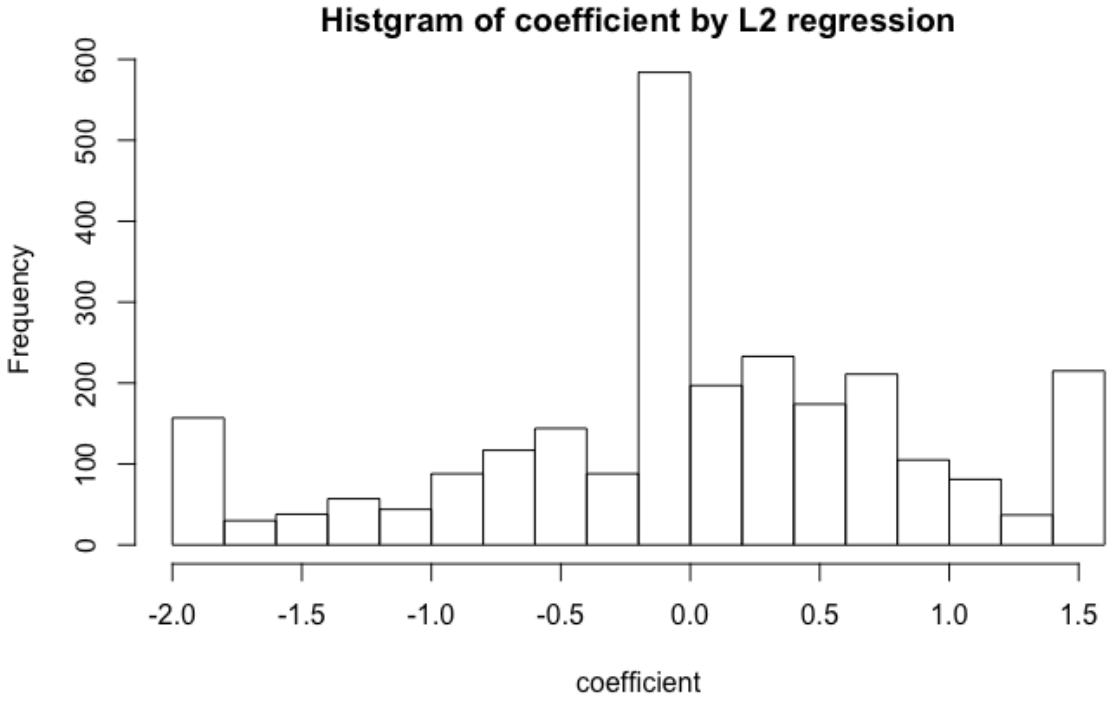


図 4.6: ネガポジ辞書の L2 正則化回帰の係数の分布

学習データ期間（2015 年 1 月から 2016 年 12 月）と同じデータを用いて、しきい値を 0 からポジティブ方向とネガティブ方向にそれぞれ 0.1 刻みで 0.0 から 1.4 、0.0 から-1.8 と変化させてネガポジ辞書を作成し、その辞書の正解率を測定した。正解率の測定には、投稿に付与されている投稿感情値と、投稿の単語からしきい値で調整されたネガポジ辞書を用い式 4.5 で計算されるネガポジ値との平均二乗誤差（MSE）を測定しその誤差の少なさで評価した。

$$Sentiment_k = \frac{1}{n}\sum_{i=1}^{n} Coefficient_{ki} \tag{4.5}$$

$Sentiment_k$ : 投稿 k のセンチメント値
$Coefficient_{ki}$ : 投稿 k の単語 i の L2 正則化回帰の係数
$n$ : 投稿 k に含まれる単語数

ここで、ネガポジ辞書を作成した時と同じ学習データをネガポジ値の予測にも用いるため、k 分割交差検定によりデータを分割して同じ条件で検証し、MSE がもっとも小さくなるしきい値を求めることにした。

まず、k 分割した際の 1 回目学習データで、あるポジティブ方向のしきい値（以下、posi_th）とネガティブ方向のしきい値（以下、nega_th）の組み合わせで作成したネガポジ辞書を用

い、MSE を計算する。この計算を全ての posi_th と nega_th の組み合わせの 252 通りで行い、それぞれのしきい値の組み合わせで得た MSE のうち、最も小さい MSE となったしきい値の組み合わせ（posi_th と nega_th）を 1 回目の学習データのしきい値とする。このしきい値を用い、k 分割した際の 1 回目の検証データを用いて MSE を計算し、これを k 分割交差検定の 1 回目のしきい値とする。この計算を k 回繰り返して行い、しきい値の組み合わせを k 組求める。最後に、k 個のしきい値の組合せの中で、最も出現回数の多いしきい値の組合せを、ネガポジ辞書を調整するためのしきい値とする。

k 分割交差検定に用いる k は k=10 とした。1 回目学習データにおける全てのしきい値の組み合わせでの MSE をヒートマップにした結果を、図 4.7 に示す。

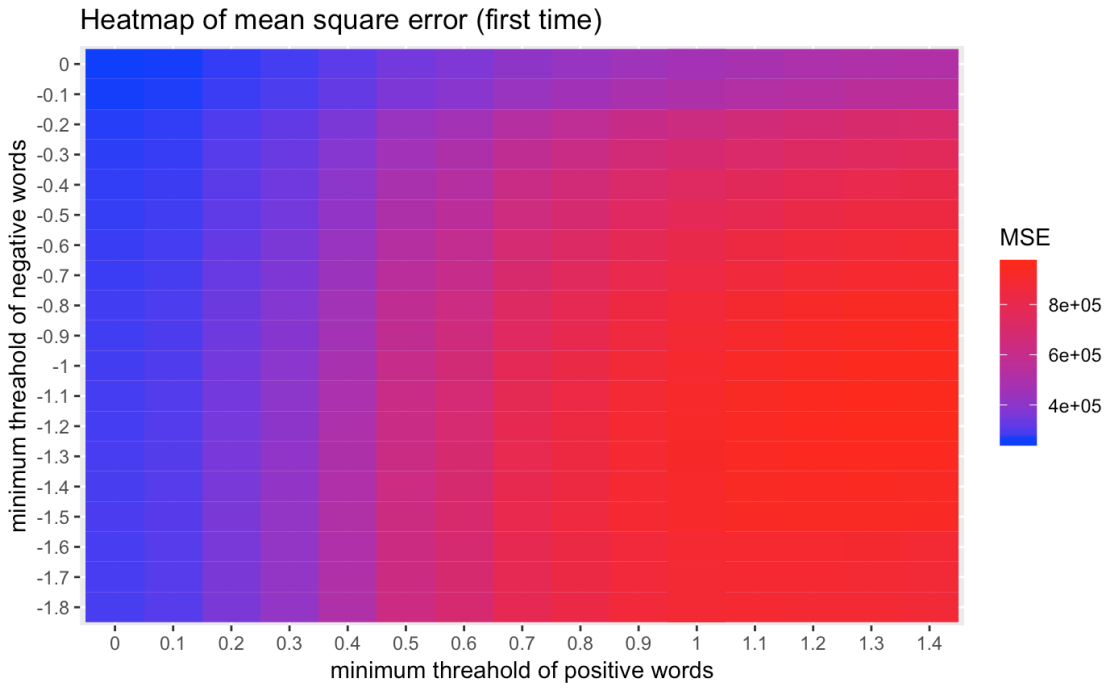


図 4.7: ネガポジ辞書のしきい値別のネガポジ値と投稿感情値の MSE。k 分割交差検定の 1 回目の結果を示している。縦軸は数のしきい値、横軸はポジティブ方向のしきい値を示し縦軸はネガティブ方向のしきい値を示す。ヒートマップ中の色タイルは、そのしきい値の組み合わせで作成したネガポジ辞書から求めたネガポジ値と投稿感情値との MSE を表す。赤は誤差が大きく当てはまりがよく、青は誤差が小さいく当てはまりが悪いことを示す。

図 4.7 は、k 分割交差検定の 1 回目の全てのしきい値の組み合わせでの作成したネガポジ辞書から求めたネガポジ値と、投稿感情値の MSE をヒートマップにして表示している。図からは、ポジティブ方向、ネガティブ方向共に 0 のしきい値の場合に、最も MSE が小さく当てはまりが良いことを示している。これを k 分割交差検定により 10 回繰り返した結果、posi_th と nega_th の組み合わせで最も出現回数が多くなる組み合わせは、posi_th が 0、nega_th が 0 となった。

これは、ネガポジ辞書は全ての単語を使う方が精度よく投稿のネガポジが判定できることを示している。しきい値を決めてネガポジ辞書を調整するということは、ネガポジ辞書を構成する単語量が減るということになる。すなわち、しきい値を大きくし単語量を減らした場合には、投稿内で一致する単語量が減るため、正確なネガポジ判定ができないのではないかと推測される。

以上の結果より、しきい値は特に設けずに作成したネガポジ辞書を使って投稿のセンチメント分析を、次章以降で行う。

### センチメント分析

4.3 章で作成したネガポジ辞書を用いて、投稿の単語から投稿ネガポジ値を計算し、その投稿の投稿感情をネガティブかポジティブであるかの 2 値分類で予測を行なった。検証データを用いた投稿ネガポジ値からの投稿感情の予測は 0.609 の正解率であったが、実際の投稿内容を確認しその投稿に書かれている内容と、投稿感情が一致しているかどうかの検証は行えていない。そこで、投稿ネガポジ値が正しいものとし、この投稿ネガポジ値と投稿評価値の関係を調べ、投稿評価値を予測する際の特徴量として適当かどうかを調査する。

投稿ネガポジ値は、学習データ期間（2015 年 1 月から 2016 年 12 月）の上位 30 銘柄の常連投稿者別に、それぞれの投稿の投稿ネガポジ値を計算した。ここで、表 3.3 から常連投稿者の全投稿数は 584,514 であり、このままではデータ量が多すぎるため分析に適さない。そこで 4.2 章での、投稿評価値は常連投稿者別に決定されるとの結果があることから、常連投稿者別のそれぞれの投稿の投稿ネガポジ値の平均値を計算し、常連投稿者別に投稿評価値との関係を調べることにした。

図 4.8 に、常連投稿者別に投稿ネガポジ値の平均（以下、平均投稿ネガポジ値）と、平均投稿評価値の関係を示す。

図 4.8 は、横軸に常連投稿者別の平均投稿ネガポジ値を表し、縦軸に投稿者別平均投稿評価値を表しており、全常連投稿者 800 名をプロットしている。図から、平均投稿ネガポジ値と平均投稿評価値には弱い正の相関が見られる。これは、常に投稿ネガポジ値が高い投稿する常連投稿者は、平均投稿評価値が高く、常に投稿ネガポジが低い投稿をする常連投稿者は、平均投稿評価値が低いことを意味する。また、調整済み決定係数（Adj-R2) は 0.10967 とあまり高くなく、データにばらつきが見られるが、p-value は非常に小さく、平均投稿評価値は平均投稿ネガポジ値により決定されていると考えられる。すなわち、投稿の単語をネガポジ辞書で定量化し、投稿ネガポジ値を求めることにより、投稿評価値は予測できることを意味していると考えられる。

以上の結果から、ネガポジ辞書を用いて算出することができる投稿ネガポジ値を、投稿評価値を予測するための特徴量 *Sentiment* として、4.4 章で用いることにする。

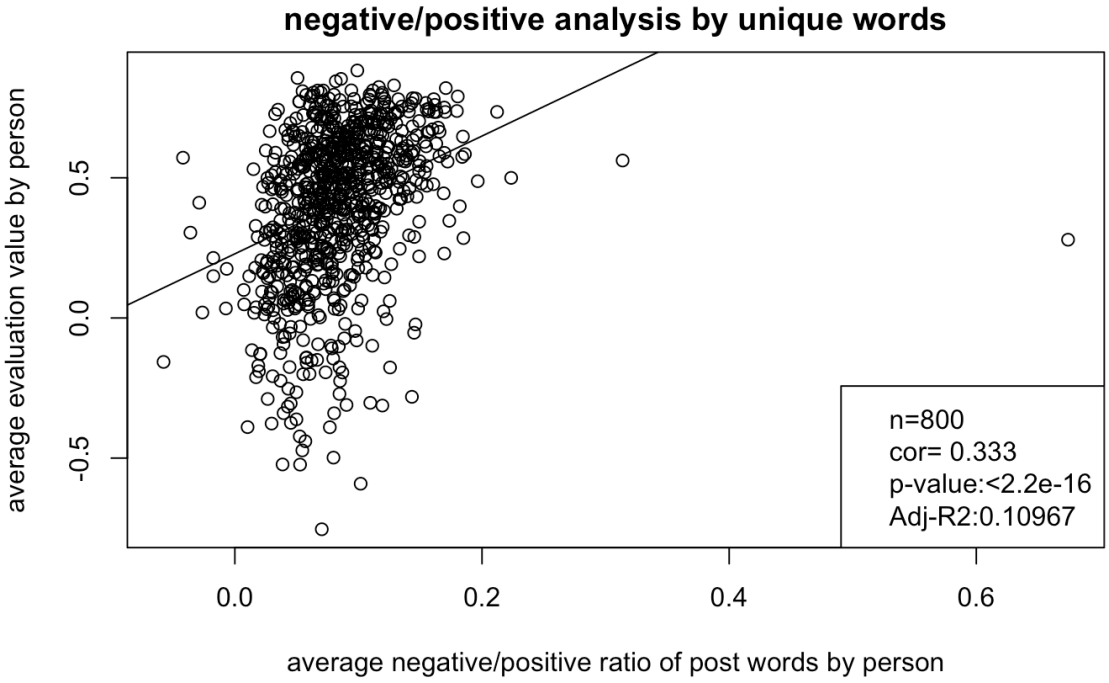


図 4.8: 常連投稿者の平均投稿ネガポジ値と平均投稿評価値の関係（2015 年 1 月から 2016 年 12 月）。常連投稿者数 n は 800、相関係数は 0.333 であり弱い正の相関が見られる。

## 4.4 信頼度予測のモデル検討

### 信頼度を予測するための特徴量

4.1 章では平均投稿評価値と銘柄の関係、4.1 章では平均投稿評価値と平気月次株価収益率の関係、4.1 章では平均投稿評価値と平均月次株価 HV の関係、4.1 章では平均投稿評価値と平均月次株価売買代金、4.2 章では平均投稿評価値と投稿者別平均投稿評価値の関係、4.3 章では、平均投稿評価値と投稿ネガポジ値の関係を調べた。

本章では投稿評価値を予測するために、前章までに、投稿評価値と株価、投稿者、投稿内容との関係から説明変数として適当であると判断した特徴量を用い、投稿評価値を予測するモデルを作成する。表 4.4 に前章までに抽出した特徴量の一覧を示す。

機械学習などを用いてモデルを構築する際には、目的変数の教師データとしての投稿評価値が極端に正や負に偏ったものを用いると、正確なモデルを構築することが不可能であり、教師データの正と負が均一になったものを用いることが望ましい。そのため、一般的にデータの間引きを行なったり、データを人工的に作成して同数にするような手法がとられている。そこで、本研究で使用するデータについて、投稿評価値に偏りがあるかどうかを調査するために、学習データと検証データの投稿評価値と投稿数を集計した結果を表 4.5 に示す。

表 4.4: 投稿評価 $Evaluation_{jk}$ を予測するための特徴量

| 変数 | 項目 |
|---|---|
| $StockReturn_j$ | 銘柄 $j$ の平均月次株価収益率 |
| $StockVolat_j$ | 銘柄 $j$ の平均 12 ヶ月月次ヒストリカル・ボラティリティ |
| $StockTurnover_j$ | 銘柄 $j$ の平均月次売買代金 |
| $StockReliability_j$ | 銘柄 $j$ の平均投稿評価値 |
| $UserReliability_p$ | 投稿 $k$ を投稿した投稿者 $p$ の平均投稿評価値 |
| $Sentiment_k$ | 投稿 $k$ のネガポジ値 |

表 4.5: 学習及び検証データの投稿評価値別の投稿数

| 種類 | 投稿評価値 | 投稿数 |
|---|---|---|
| 学習データ | 投稿評価値 <0 | 288,531 |
| 学習データ | 投稿評価値 >0 | 794,056 |
| 検証データ | 投稿評価値 <0 | 34,075 |
| 検証データ | 投稿評価値 >0 | 34,075 |

表 4.5 から、学習データにおいて負の投稿評価値の投稿数が 288,531 に対し、正の投稿評価値の投稿数が 794,056 と 2 倍以上の差があることがわかる。また、検証データにおいても、負の投稿評価値の投稿数が 34,075、正の投稿評価値の投稿数が 34,075 と、こちらも約 2 倍近い差があるとがわかり、このまま使用すると、うまく学習や評価ができないことが予想される。

そこで、学習データ及び検証データを投稿評価値が正、負で同数になるように、ランダムにそれぞれ抜き出して使用することにした。その結果、学習データは正の投稿評価値を 280,000、負の投稿評価値を 280,000 となるように抜き出し、合計 560,000 の学習データ（以下、調整済み学習データ）を抜き出した。同様に、検証データから正の投稿評価値を 30,000、負の投稿評価値を 30,000 抜き出して、合計 60,000 の検証データ（以下、調整済み検証データ）を作成した。

以降では、この調整済み学習データでモデルの学習を行い、調整済み検証データでモデルの性能を検証することにする。

### 重回帰モデル

投稿評価値は-1 から 1 の連続値で表されることから、投稿評価値を予測するモデルとして重回帰式による予測を行うことにした。その理由として、本研究で予測する投稿評価値は 1 つの変数で表され、それを複数の説明変数で予測するために、重回帰分析は適切であるからである。

そこで、重回帰分析のモデル（以下、重回帰モデル）式を式 4.6 に定義し、この式を調整済み学習データを用いて係数を求め、モデル式を構築する。その後、検証データで検証を行い、重回帰モデルの当てはまり度を検証する。

$$
\begin{aligned}
Evaluation_k &= a_1 \times StockReturn_j + a_2 \times StockVolat_j \\
&\quad + a_3 \times StockTurnover_j + a_4 \times StockReliability_k \\
&\quad + a_5 \times UserReliability_p + a_6 \times Sentiment_k \\
Evaluation_k &: \text{銘柄 j の投稿 k の投稿評価値} \\
StockReturn_j &: \text{銘柄 j の平均月次株価収益率} \\
StockVolat_j &: \text{銘柄 j の平均月次株価 HV} \\
StockTurnover_j &: \text{銘柄 j の平均月次売買代金} \\
StockReliability_j &: \text{銘柄 j の平均投稿評価値} \\
UserReliability_p &: \text{投稿 k を投稿した常連投稿者 p の投稿者別平均投稿評価値} \\
Sentiment_k &: \text{投稿 k の投稿ネガポジ値} \\
a_1, \ldots, a_6 &: \text{係数}
\end{aligned}
\tag{4.6}
$$

式 4.6 は、銘柄を $j$、常連投稿者を $p$、常連投稿者 $p$ が投稿した一つの投稿を $k$ で表現し、目的変数として銘柄 $j$ の投稿 $k$ の投稿評価値、説明変数として、銘柄 $j$ の平均月次株価収益率、平均月次株価 HV、平均月次売買代金、平均投稿評価値と、投稿 $k$ を投稿した常連投稿者 $p$ の平均投稿評価値、投稿 $k$ の投稿ネガポジ値を持つ。

重回帰分析を行う際に、赤池情報量規準 [2] を用いて説明変数量を削減を行なった。赤池情報量規準とは、統計のモデルの良さを評価するための指標である。一般的に、重回帰分析においては説明変数の数を増やすと、目的変数への適合度が高くなることが知られているが、その反面、入力データに無理に合わせた学習結果となってしまうという過学習という現象を招くことが知られている。赤池情報量規準はこの過学習を抑え、適切な説明変数を選択するために用いた。

調整済み学習データを用い、式 4.6 で重回帰分析を行い、赤池情報量規準重による回帰分析で、説明変数の削減を行なった結果、重回帰モデル式は式 4.7 になった。

$$
\begin{aligned}
Evaluation_k &= a_1 \times StockReturn_j + a_2 \times StockVolat_j \\
&\quad + a_3 \times StockTurnover_j \\
&\quad + a_5 \times UserReliability_p + a_6 \times Sentiment_k
\end{aligned}
\tag{4.7}
$$

式 4.7 の結果から、説明変数として、銘柄の平均投稿評価値 $StockReliability$ が削除された。このことより、投稿評価値の予測には銘柄別の平均投稿評価値は関係ないと言える。これは、ある投稿がどの銘柄の掲示板に投稿されたかに関係なく、その投稿として投稿評価値が決定されるとういうことになる。つまり、投稿の投稿評価値は平均月次株価収益率、平均

月次株価 HV、平均月次売買代金、投稿した常連投稿者 $p$ の平均投稿評価値、投稿ネガポジ値が同じ条件で 2 つの銘柄の掲示板があった場合には、銘柄に関係なく同じ投稿評価値が付与されることを意味している。つまり、銘柄の特性、例えば、銘柄の業種や規模といった銘柄固有の情報は関係がないとうことになる。すなわち投稿者が投稿する際には、銘柄固有の情報は関係なく、株価などの情報のみに影響した投稿が評価された投稿評価値、すなわち信頼度になると推測される。

次に、各説明変数の係数や当てはまり度を詳細に調査するため、重回帰分析で出力される各説明変数の係数、t 値、p 値と自由度調整済み係数を表 4.6 に示す。

表 4.6: 重回帰分析で計算された係数。t 値は係数が 0 であるという帰無仮説に対する t 検定により計算される値。p 値は帰無仮説のもとで係数が 0 となる確率。本研究では 5%以下で帰無仮説を棄却する。

| | 係数 | t 値 | p 値 |
|---|---|---|---|
| 切片 | -2.539e-01 | -191.894 | 2e-16 以下 |
| $a_1$ | 9.355e-02 | 9.241 | 2e-16 以下 |
| $a_2$ | 4.159e-02 | 37.466 | 2e-16 以下 |
| $a_3$ | 6.282e-05 | 5.976 | 2.29e-09 |
| $a_5$ | 1.083e+00 | 453.189 | 2e-16 以下 |
| $a_6$ | 6.257e-02 | 18.451 | 2e-16 以下 |

AdjR-2: 0.2976

表 4.6 から、p 値を確認すると全て 10 の-8 乗以下と有意水準を 5%とした際には有効となっている。

次に、係数は全て正の値を取っているが、4.1 章の結果では、平均月次株価収益率、平均月次株価 HV は平均投稿評価値と正の相関、平均月次売買代金は平均投稿評価値と負の相関が見られていたのにも関わらず、重回帰モデルの結果では、$a_3$ で示される平均月次売買代金の係数の値は小さいものの正の値となってしまっている。これは、4.1 章では銘柄別の平均投稿評価値と平均月次売買代金を用いていたが、重回帰モデルでは、個別の投稿を用いているために差が出たものではないかと考えられる。ただし、表 4.6 の $a_3$ は他の係数よりも小さく t 値も 5.967 と最も絶対値が小さいため影響は微小であると考えられる。投稿ネガポジ値は 4.3 章で、投稿ネガポジ値と投稿評価値には正の相関が見られた結果と、投稿ネガポジ値の係数 $a_5$ が正となることと生合成が取れている。投稿者別平均投稿評価値の係数 $a_4$ が最も大きく、この重回帰モデルに影響していると考えられる。

しかし、自由度調整済み決定係数（AdjR-2）が 0.2976 と低く、重回帰分析による投稿評価値の予測モデルは有意であるとは言い難い。また、連続的な投稿評価値をモデルとして予測することは難しいため、投稿評価値が正か負かの 2 値分類でモデルが作成が可能な決定木

を用い、予測モデルの構築を行う。決定木は、分類の処理を可視化可能であり、より詳細に影響の大きい説明係数を確認できることが期待される。

### 決定木モデル

重回帰モデルでは自由度調整済み決定係数（AdjR-2）が低くモデルとしての信頼性があまり良くないために、投稿評価値が 0 より大きいか 0 未満かの 2 値分類による予測モデルを構築し、モデルの当てはまり度を確認する。また、決定木モデルでは、分類の過程を可視化可能であるため、説明変数の分類条件を確認することができるため、可視化による確認を行い、モデルの詳細を確認する。

説明変数は表 4.4 の特徴量を用い、決定木によるモデル（以下、決定木モデル）を作成し、調整済み学習データを用いてモデルを学習する。モデルの検証には調整済み検証データを用いた 2 値分類で、正解率、適合率、再現率、F 値を測定し行う。決定木の学習及び検証にあたっては、数値解析プログラム R[4]の rpart パッケージを用いた。

まず、調整済み学習データを用いて決定木による学習を行う。次に、調整済み検証データでの予測結果の正解率、適合率、再現率、F 値を検証し、調整済み学習データの学習時に得た正解率、適合率、再現率、F 値との比較により、過学習がおこっていないかどうかを確認する。

調整済み学習データを用いて決定木による学習を行い、投稿評価値の正負を予測した結果を表 4.7 に示す。

表 4.7: 決定木による調整済み学習データでの投稿評価値予測結果

| | 予測 | |
|---|---|---|
| | 投稿評価値 >0 | 投稿評価値 <0 |
| 投稿評価値 >0 | 201,757 | 78,243 |
| 投稿評価値 <0 | 75,344 | 204,656 |

Precision：0.721、Recall：0.728、Fmeasure：0.724、Accuracy：0.726

表 4.7 から、F 値（Fmeasure）が 0.724、正解率（Accuracy）が 0.726 の性能を得た。また、適合率（Precision）は 0.707、再現率（Recall）は 0.785 と、適合率と再現率にあまり差がない、つまり、どちらかに偏らず適切な分類を行なっていると言える。

次に、調整済み検証データで検証を行なった結果を表 4.8 に示す。

[4]https://cran.r-project.org/

表 4.8: 決定木による調整済み検証データでの投稿評価値予測結果

| | 予測 | |
|---|---|---|
| | 投稿評価値 >0 | 投稿評価値 <0 |
| 投稿評価値 >0 | 21,201 | 8,799 |
| 投稿評価値 <0 | 5,821 | 24,179 |

Precision：0.707、Recall：0.785、Fmeasure：0.744、Accuracy：0.756

表 4.8 から、調整済み検証データでの F 値が 0.744、正解率が 0.756、適合率が 0.707、再現率が 0.785 と、ほぼ調整済み学習データによる結果と変わらず、過学習が起きていないことがわかる。また、調整済み学習データの結果と同様に、適合率、再現率ともに偏りがなく、適切に学習が行えていることがわかる。正解率は 0.756 と 4 回に 3 回は当たるという形で、十分にモデルとしては有意義なものになったのではないかと思われる。

次に、決定木モデルで、どの説明変数がどのように計算されて分類されたのかを確認するために、調整済み学習データで学習した際の決定木の分類条件を、R の rpart パッケージにて可視化を行った結果を、図 4.9 に示す。

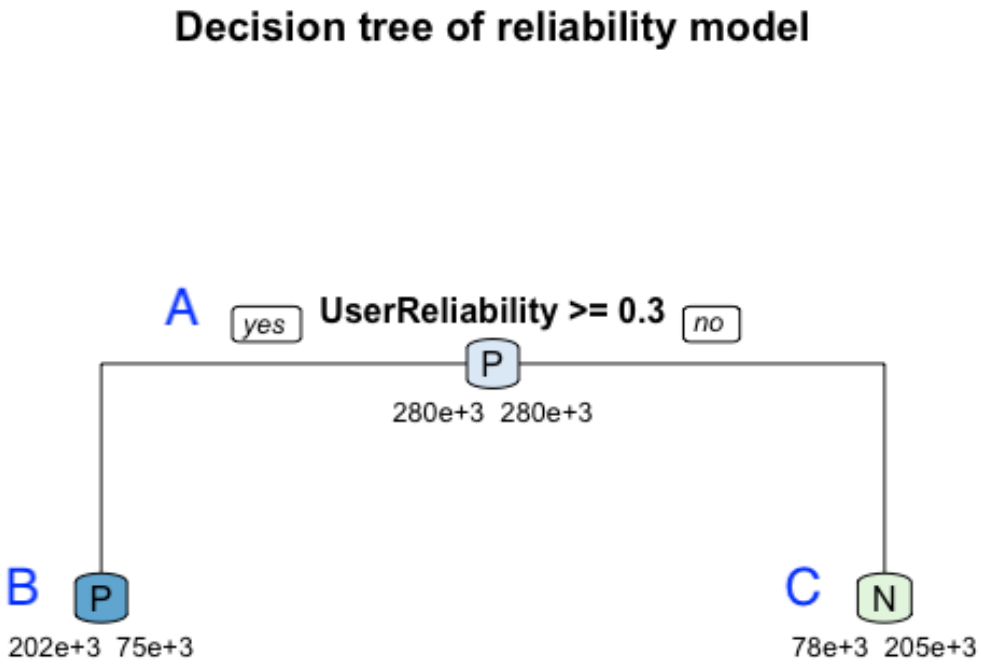


図 4.9: 決定木モデルの可視化。常連投稿者の平均投稿評価値のみが説明変数として存在し、投稿評価値の予測は投稿者の平均投稿評価値で決定されることを示している。

図 4.9 は決定木の 2 分木を表しており、2 分木の根において投稿者平均投稿評価値（User-Reliability）が 0.3 以上か未満かで 2 分している。この時、投稿評価値を予測し、投稿評価値が正であると予測した条件、すなわち投稿者平均投稿評価値が 0.3 以上に振り分けられた投稿数が 280,000、投稿評価値が負であると予測した条件、すなわち 0.3 未満に振り分けられた投稿数が 280,000 である（図中の A）。投稿評価値が正であると予測された投稿のうち、実際の投稿評価値が正であったものが 202,000 投稿、負であったものが 75,000 である（図中の B）。同様に、投稿評価値が負であると予測された投稿のうち、実際の投稿評価値が正であったものが 78,000、負であったものが 205,000 である（図中の C）。

以上を踏まえ、図 4.9 から、表 4.4 に示した説明変数の内、常連投稿者の平均投稿評価値のみが投稿評価値の予測に影響していることを意味していることがわかった。すなわち、決定木モデルによる分類は、どの投稿においても、その投稿を投稿した常連投稿者の、投稿者平均投稿評価値によって、投稿される投稿評価値が決定される。つまり、その投稿の投稿評価値が正か負、すなわち信頼できるかできないかの判定は、その投稿を行なった常連投稿者の学習データから計算された投稿者平均投稿評価値が 0.3 以上の時には投稿評価値が正、すなわち信頼できると予測され、投稿者平均投稿評価値が 0.3 未満の時には投稿評価値が負、すなわち信頼できない投稿と予測される。

以上の結果より、決定木モデルを用いて常連投稿者の投稿の投稿評価値が正であるか負であるかの 2 値分類の予測は、過去に投稿した常連投稿者の投稿者平均投稿評価値により予測が可能である。これは、重回帰モデルでの投稿者平均投稿評価値の係数が最も重回帰モデルに影響していたことと整合している。また、決定木モデルによる投稿評価値が正か負かであるかの 2 値分類の予測の正解確率は 0.756 であることがわかった。

# 第5章　株価収益率の予測

4.4 章では、投稿者の平均投稿評価値により、投稿の投稿評価値が予測できるということがわかった。すなわち、投稿の信頼度が予測できることを意味している。

本章では、株価掲示板の投稿において投稿評価値の高い投稿と投稿評価値の低い情報の 2 種類、すなわち信頼度の高い情報と低い情報の 2 種類の情報から、それぞれが株価収益率が予測できるかどうかの検証を行う。信頼度の高い情報は、平均投稿評価値の高い常連投稿者の投稿を用い、信頼度の低い情報は平均投稿評価値の低い常連投稿者の投稿を用いることにし、それぞれの投稿と翌日株価収益率の関係を調べることで、株価の予測性能があるかどうかを検証する。

## 5.1　方法

投稿の投稿評価値は常連投稿者の平均投稿評価値により決定されることから、常連投稿者の 800 名のうち平均投稿評価値の低い下位 5 %の常連投稿者（以下、上位投稿者）と、平均投稿評価値の高い上位 5 %の常連投稿者（以下、上位投稿者）をそれぞれ 40 名抽出した。

上位投稿者が投稿者である投稿は信頼度が高く、下位投稿者が投稿者である投稿は信頼度が低いと定義し、これらの投稿が株価収益率を予測できるかどうかを、次の方法で求める。

投稿には投稿感情が付与されているものがあるが、その投稿のうち「強く買いたい」、「買いたい」の投稿感情（ポジティブ感情）が付与されている投稿は翌営業日株価収益率が上昇すると予測しているとし、「売りたい」、「強く売りたい」の投稿感情（ネガティブ感情）が付与されている投稿は翌営業日株価収益率が下落すると予測しているものとして、この 2 値分類による予測の正解率から予測性能があるかどうかを検討する。ここで、予測においては、ランダムに上昇下落を予測した場合には 0.5 の確率で正解することから、0.5 に対する正解率を比較し予測性能を評価する。

なお、投稿は営業時間中にも投稿できるが、営業時間中は株価が動いているため、その動きに合わせた投稿感情になることが考えられる。そのため、営業中の株価の動きによる投稿感情の予測に対する影響を排除するため、営業日 15 時から翌営業日 8 時 59 分までの投稿を用い、翌営業日の株価収益率との関係を調べた。また、株価予測におけるデータは投稿評価値の予測に関係がないため、検証データに比較しデータ量が多い学習データを用いた。

## 5.2 予測結果

学習データ期間中に上位投稿者の買い感情または売り感情が付与された投稿は 936 投稿、下位投稿者の買い感情または売り感情が付与された投稿は 6,728 投稿であった。これらの投稿と、翌営業日株価収益率の関係を調べた結果を表 5.1、表 5.2 に示す。

表 5.1: 上位投稿者の翌営業日株価収益率の予測結果（2015 年 1 月 1 日から 2016 年 12 月 31 日）

| 投稿感情 | 予測 | |
|---|---|---|
| | 翌営業日株価収益率 >0 | 翌営業日株価収益率 <0 |
| 買いたい、強く買いたい | 520 | 384 |
| 売りたい、強く売りたい | 22 | 10 |

Precision：0.575、Recall：0.959、Fmeasure：0.719、Accuracy：0.566

表 5.1 は上位投稿者の投稿感情と翌営業日の株価収益率の予測結果を示しており、適合率（Precision）が 0.575、再現率（Recall）が 0.959 となり、適合率に比較し再現率が高い。これは、買い感情の投稿数が売り感情の投稿数に比較し非常に大きいためであり、偏ったデータで 2 値分類を行うと偏ったデータの予測数が増えるため、必然的に再現率が高くなると推測される。また、正解率（Accurary）は 0.566 となった。

表 5.2: 投稿評価値の低い常連投稿者の翌営業日株価収益率の予測結果（2015 年 1 月 1 日から 2016 年 12 月 31 日）

| 投稿感情 | 予測 | |
|---|---|---|
| | 翌営業日株価収益率 >0 | 翌営業日株価収益率 <0 |
| 買いたい、強く買いたい | 2,566 | 3,069 |
| 売りたい、強く売りたい | 452 | 641 |

Precision：0.455、Recall：0.850、Fmeasure：0.593、Accuracy：0.477

表 5.2 は下位投稿者の投稿感情と翌営業日の株価収益率の予測結果を表しており、適合率が 0.455、再現率が 0.850 となり、適合率に比較し再現率が高い。これも、買い感情の投稿数が売り感情の投稿数に比較し非常に大きいためであり、偏ったデータで 2 値分類を行うと偏ったデータの予測数が増えるため必然的に再現率が高くなると推測される。また、正解率は 0.477 となった。

ランダムに予測した場合には、正解率は 0.5 であることを踏まえ、投稿評価値の高い常連投稿者の予測と投稿評価値の低い常連投稿者の正解率を比較したところ、投稿評価値の高い常連投稿者の正解率は 0.566 と 0.5 を上回っていた。一方、投稿評価値の低い常連投稿者の

正解率は 0.477 と 0.5 を下回っていた。この正解率の差が偶然のものであるのか、予測に用いたモデルによる差なのかを、カイ二乗検定で検定する。検定にあたり、次の仮説を立て、有意水準 5%で検定を行う。

帰無仮説 $H_0$：常連投稿者の株価収益率の予測正解率とランダムに行なった予測の予測正解率に差はない。

対立仮説 $H_1$：常連投稿者の株価収益率の予測正解率とランダムに行なった予測の予測正解率に差はある。

ランダムに予測した際に 0.5 の正解率となるとき、投稿評価値の高い常連投稿者の予測および、投稿評価値の低い常連投稿者の予測をそれぞれ複数回に予測を行った際のカイ二乗値を、福井ら [31] の手法を参考に、式 5.1 で求めた。

$$\chi^2 = \frac{(\text{正解数} - \text{正解期待度数})^2}{\text{正解期待度数}} + \frac{(\text{不正解数} - \text{不正解期待度数})^2}{\text{不正解期待度数}} \tag{5.1}$$

正解数は翌営業日株価が上がるもしくは下がると予測したデータのうち、実際にそうであったデータ数のことを示し、不正解数は翌営業日株価が上がるもしくは下がると予測したデータのうち、実際にそうでなかったデータ数のことを示している。正解期待度数および不正解期待度数は、ランダムに予測した際には 0.5 の正解率であることから、予測データ数の 0.5 である。なお、式 5.1 で得られた $\chi^2$ 値は、自由度 1 の $\chi^2$ 分布に従う。$\chi^2$ 値を、表 5.1 および表 5.2 から求め、有意水準 5%にて検定を行なった結果を、表 5.3 に示す。

表 5.3: 翌営業日株価収益率予測の正解率の有意水準 5%におけるカイ二乗検定結果（2015 年 1 月 1 日から 2016 年 12 月 31 日）

| 投稿者 | $\chi^2$ 値 | p 値 | 検定結果 |
|---|---|---|---|
| 投稿評価値の高い常連投稿者 | 16.42735 | 0.00005055 | $< 0.05$ |
| 投稿評価値の低い常連投稿者 | 14.65458 | 0.000129121 | $< 0.05$ |

表 5.3 から、上位投稿者および下位投稿者の予測の正解率の p 値は 0.05 以下となり、帰無仮説は棄却され、本章で行なった常連投稿者の株価収益率の予測の正解率と、ランダムに行なった予測の予測正解率に差はあるといえる。

本章の結果から、上位投稿者の翌営業日株価収益率の予測は、ランダムに予測したものより高く、下位投稿者の翌営業日株価収益率の予測は、ランダムに予測したものよりも低いといえる。また、上位投稿者は下位投稿者に比較して、翌日株価収益率の予測性能が高いと言える。よって、投稿者平均投稿評価値が高い投稿を行う上位投稿者、すなわち常に信頼度の

高い投稿を行う投稿者の予測性能は、投稿者平均投稿評価値が低い投稿を行う下位投稿者、すなわち常に信頼度の低い投稿を行う投稿者の予測性能よりも高いことがわかった。

# 第6章　考察およびまとめ

## 6.1　考察

### 株価と信頼度の関係

4.1 章にて平均月次株価収益率、平均月次株価 HV、平均月次売買代金と、平均投稿評価値の相関関係を調べた。平均月次株価収益率及び平均月次株価 HV と平均投稿評価値は正の相関、平均月次売買代金と平均投稿評価値は負の相関が見られた。ここで、投稿者＝投資家という仮定を置いて考察してみる。

まず、平均月次株価収益率及び平均月次株価 HV と平均投稿評価値は正の相関という理由を考察する。投資家はおもに、株が安い時に買って高い時に売ることにより、利益を出すという行動をとろうとする。利益を出すためには、同じ投資期間でも株価収益率が高いほど、より高い利益が得られるということであることを考えると、平均月次株価収益率が高いほど、投資家にとってはその銘柄は良いということになる。同様に、平均月次株価 HV が高いということは、同じ期間においても株価の動きが大きくなる、すなわち、株価収益率の絶対値が大きくなりやすいため、同じ期間でもより高い利益を得ることができるということになる。よって、高い利益の得やすい状況の時には、掲示板の信頼度が上がるということになるのではないかと言える。また、投資家の利益が増えるということは、心理的に非常にポジティブな状態になると考えられ、その結果、ポジティブな投稿が増えたとすると、図 4.8 の投稿ネガポジ値が高い投稿ほど平均投稿評価値が高くなるという相関関係とも一致する。よって、投資家の利益が増えるような株価の状況の時には、掲示板の投稿評価値が高くなると考えられる。

次に、平均月次売買代金と平均投稿評価値が負の相関関係という理由を考えてみる。通常、平均月次売買代金が大きくなるということは、取引が活発になるので株価が上昇しやすく、すなわち、平均投稿評価値が高くなるのではないかと考えられる。しかし、図 4.4 では、逆の関係となっている。これを詳細に確認するために、上位 30 銘柄の平均月次売買代金と平均月次株価 HV の関係を調べた結果を、図 6.1 に示す。

図 6.1 は、横軸に銘柄ごとの平均月次売買代金、縦軸に平均月次株価 HV を示している。なお図は、上位 30 銘柄のうち 25 銘柄のものをプロットしているが、これは、12 ヶ月の期間で月次ヒストリカル・ボラティリティを算出するためには、学習データ期間が 2015 年 1 月から 2016 年 12 月のときには、3 年分 2014 年 1 月から 2016 年 12 月までの株価データが必要になるため、この期間全てにデータが存在する 25 銘柄について計算した結果となって

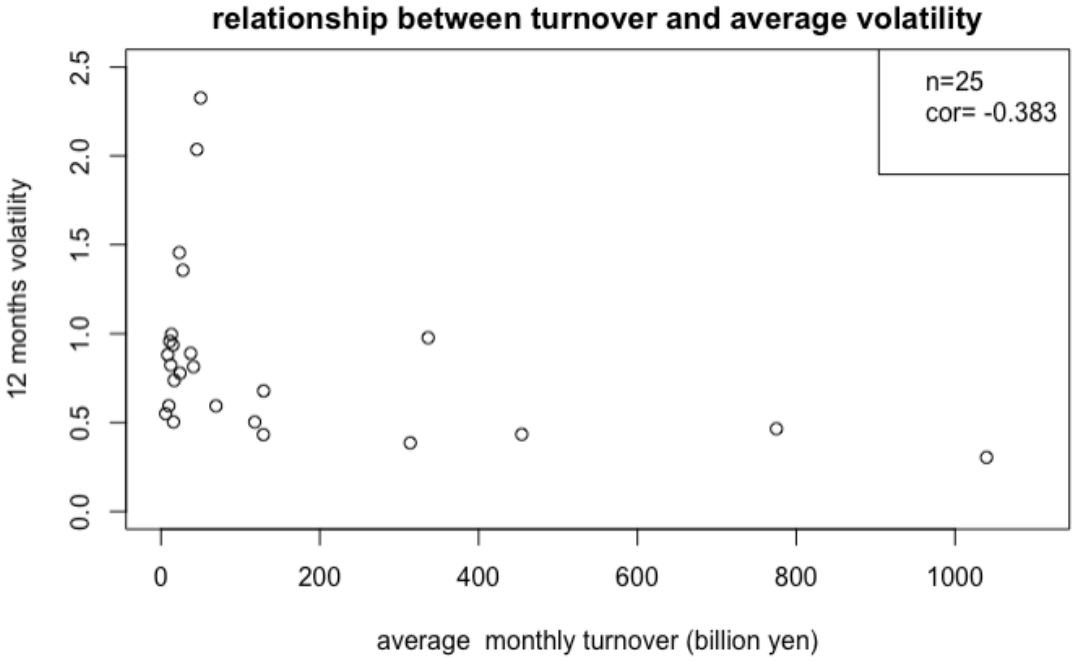


図 6.1: 平均月次売買代金と平均月次株価 HV の関係（2015 年 1 月から 2016 年 12 月）

いる。

図 6.1 から、平均月次売買代金が多くなると平均月次株価 HV が小さくなるという関係が読み取れる。これは、平均月次株価 HV が大きくなると平均投稿評価値は高くなり、平均月次株価 HV が小さくなると平均投稿評価値は小さくなるという正の相関があることを踏まえると、次の説明通りとなる。平均月次売買代金が多くなるということは、平均月次株価 HV が小さくなる、すなわち平投稿評価値が低くなるという関係と同じである。逆に、平均月次売買代金が増えると、平均月次株価 HV が高くなり、それゆえ平均投稿評価値が高くなるということになる。

以上の結果から、株価収益率、株価ヒストリカル・ボラティリティから算出される株価の指標は、投資家の利益が増えるような状態、すなわち、株価収益率が高い、もしくは株価ヒストリカル・ボラティリティが高い株価の状況の時には、掲示板の投稿評価値が高くなると言える。

### 信頼度による投稿者の分類

投稿者別に信頼度が大きく異なるという理由について、次に考えてみる。投稿者の信頼度は、図 4.5 から中央値は 0.46 と、全体的に信頼度が高い投稿が多いことがわかった。しかしながら、投稿評価値が 0 以下となる投稿者も存在し、投稿者による違いが見られた。

この関係を詳細に調べるために、実際の投稿内容を目視にして確認し理由を推測するために、

平均投稿評価値の最も高い常連投稿者である投稿者 ID が”pOkHo5h_sS8CKupGDa3ySs5oQLI-”の投稿者の投稿を 5 つ抜き出した結果を、表 6.1 に示す。

表 6.1: 投稿評価値の最も高い常連投稿者の投稿の例。平均投稿評価値:0.8829

| 投稿評価値 | 投稿内容 |
|---|---|
| 0.821 | ちとわからないのですが水素って地球上で一番小さい原子だと思うんですが、その一番小さい原子をタンクで貯めるってほんまに大丈夫なんですかね？イメージではざるに水を入れてるような感じだと思うし事故にあったらほんまに怖くて乗れない。と自分が思いつくぐらいなので日本中で結構思ってる人がいるよって売れない、普及しない。やっぱり EV の方が利点はあるし今はコンビニの駐車場にも充電場所も設置されてるし水素はこないかな…早くバテナイスの組み込まれた製品を使ってみたい今日この頃まだ石の上にも 10 年で気絶しときます |
| 0.747 | 糸満の漁師様に以前電池展見に行ってきたほうがいいよと言われあの時の興奮がもう少しで始まると思うととてもわくわくしてきます。いよいよ序章の始まりまでのカウントダウンが切ってくれれば最高です♪石の上にも只今 2 年目ですのであと 8 年は気絶しておきます。 |
| 0.935 | いよいよですね♪ |
| 0.922 | あんたまだ居たのか？ずっと前からネガティブ発言ばっかりしてるけど結局のところマイクロニクスの株が欲しいのか？どうしたいん？ |
| 0.867 | 第三者から見る限り多分持ってないんぢゃないですかね。載せれば一発ですむ話がダラダラと続いてますし (笑) 私は友人からブライトリングのお金があっても要件満たした人しか買えないエマージェンシーっていうやつを買いに行ったらデザインが微妙で…(;´д｀) 止めましたわ結局ウブロの方がほしくなりました。あとはオーバースペックのディープシーもいいなって思った次第です。まぁ時計と車は自己満の世界だと思ってます♪ |

表 6.1 は、平均投稿評価値が 0.8829 と最も高い常連投稿者の投稿内容であり、投稿の投稿評価値と投稿内容を表にしたものである。投稿は銘柄は指定せずに投稿日付の若い順に 5 つ抜き出したものである。表 6.1 から、平均投稿評価値の高いこの常連投稿者の投稿は、全て投稿評価値が 0.8 以上と評価が高い。また、内容を目視で確認したところ、言葉遣いも丁寧な文章であり、”♪”のような記号を用いておりポジティブな内容であるように読み取れる。上から 4 つ目の投稿は反論するような内容ではあるが、それほど嫌味を感じるような書き方でもないように見える。投稿自体の書き方も他者を煽るような書き方ではなく、個人の主張が主な投稿内容となっているように読み取れる。

このように、平均投稿評価値の高い常連投稿者の投稿は、目視で確認してもポジティブな内容の投稿に読み取れることから、投稿評価値が高くなったものではなかと推測される。

次に、投稿評価値が最も低い常連投稿者である、投稿者 ID が”3KhnilVgsTFsbfgSZX4-”の常連投稿者の投稿を 5 つ抜き出したものを表 6.2 に示す。

表 6.2: 平均投稿評価値の最も低い常連投稿者の投稿の例。平均投稿評価値: -0.7538

| 投稿評価値 | 投稿内容 |
|---|---|
| -0.622 | 最近、正直下値模索の確信があったので投稿する気にならなかったのです・・・でも、いくらなんでも反転する（して欲しい！！）気が致します！！ホルダーの方々それぞれのスタンスがありますから私のスタンスを主張する気はございませんが東京電力が復活しない日本の将来なんて希望のかけらも無いと私は確信しております！！ちょっと停滞すれば失敗だ！倒産だ！・・・と全体が見えていない連中が多くって反吐がでますよね！！！ |
| -0.722 | 上場廃止だとか・・・こいつら 120 円の時からほざいていますｗｗ |
| -0.619 | お友達の皆様！！ありがとうございました！！東電株価がテイタラクの中なんだか気持ちがスッキリ致しました！！皆様・・・くれぐれも資産を大切にして信義を自信持って主張してくださいませ！！！ m(_ _)m |
| -0.590 | TPP を頑張った甘利氏は復活で OK カス知事のマスゾエは辞任で OK マスゾエがバッシングされるのはそもそも親韓政策を取る事が全ての原因と言い切っても良いマスゴミは問題の確信をいつもはぐらかして嫌韓感情から日本国民を逸らそうとする成田の大韓航空の故障の原因はどうなったんだよ！ |
| -0.571 | 年初来最安値更新 472 円・・・”120 円の大底””318 円の中底” に続く ”小底” の値として相応しくはないかね？？もちろん東電株価 4 桁奪還の最後の発射台としてのね！！（＾_=）-※いい加減、無意味な機関のカラ売り攻勢も買い戻しに向かうんじゃないかなｗｗ |

表 6.2 は、平均投稿評価値が-0.7538 と最も低い常連投稿者の投稿内容であり、投稿の投稿評価値と投稿内容を表にしたものである。投稿は銘柄は指定せずに投稿日付の若い順に 5 つ抜き出したものである。表 6.2 から、平均投稿評価値の低いこの常連投稿者の投稿は、全て投稿評価値が-0.5 以下と評価が低い。また、内容を目視で確認したところ、言葉遣いはそれほど悪いような印象は得ない文章ではあるが、”！”や”w”のような記号を多用しており、激しい印象を受ける。また、”マスゴミ”や”反吐”のような汚い単語を用いており、読む側の印象としてはあまりいいものではないように見受けられる。投稿自体の書き方も他者を煽るような書き方が多く、個人の主張も激しい印象を受けるような稿内容となっているように読み取れる。

このように、平均投稿評価値の低い常連投稿者の投稿は、目視で確認するとあまりいい印象を受ける内容ではなく、ネガティブな印象を受ける内容の投稿に読み取れることから、投稿評価値が低くなったものではなかと推測される。

さらに、表 6.1 と表 6.2 の比較からもわかるように、平均投稿評価値が最も高い投稿者の投稿は、読んだ時に同意を得やすい文章であるのに対し、平均投稿評価値の最も低い投稿者の投稿は、読み手に嫌悪感を与えるような文章である。このことから、ネガポジ辞書の単語から予測された投稿ネガポジ値からだけではなく、文章の表現方法によっても投稿評価値が変わってくるのではないかと推測される。

以上の結果より、丁寧であり、ポジティブな印象を受ける文章を投稿する投稿者は投稿評価値が高くなり、攻撃的な書き方であり、ネガティブな印象を受ける文章を投稿する投稿者は投稿評価値が低くなるという結果になった。なお、Potthast ら [8] は、フェイクニュースの発見手法を、内容によるもの（Knowledge-based）、伝達方法によるもの（Context-based）、文章スタイルによるもの（Style-based）の 3 種類にカテゴライズしており、本研究においては、Style-based に関連した投稿者の信頼度は文章のスタイルである書き方により信頼度が異なるという結果が得られた。

### 投稿者と掲示板の関係

投稿者と投稿評価値の関係を示した図 4.5 から、常連投稿者別に投稿者別平均投稿評価値が異なることがわかった。また、銘柄別の平均投稿評価値を調べた図 4.1 から、銘柄別にも投稿評価値が異なることがわかった。もし、常連投稿者が全ての銘柄に均等に投稿しているとするならば、銘柄別の投稿評価値はほぼ等しいものになるのではないかと考えられる。しかし異なっていることということは、常連投稿者はある特定の掲示板に対してのみ投稿しており、しかも投稿者別平均評価値の低い人が集まる銘柄は、銘柄の投稿評価値が低く、逆に、投稿者別平均投稿評価値が高い人が集まる銘柄は、銘柄の投稿評価値が高くなるのではないかと推測される。

そこで、常連投稿者の投稿の偏りを調べるため、常連投稿者が一人当たり幾つの銘柄に投稿しているのかを調べた結果を図 6.2 に示す。

図 6.2 は常連投稿者が学習データ期間における、一人当たりの投稿した銘柄数である。1 投稿でもした場合には 1 銘柄と数えている。図から、投稿者一人当たりの投稿数は常連投稿者は平均的に全ての銘柄に投稿しているのではなく、ある特定の銘柄に集中的に投稿しているということが読み取れる。これを、銘柄別の投稿評価値が異なることを合わせて考えると、投稿者別平均投稿評価値が高い投稿者は平均投稿評価値が高い銘柄に、投稿者別平均投稿評価値が低い投稿者は平均投稿評価値が低い銘柄に集まっているのではないかと推測される。この関係を調べるために、常連投稿者と銘柄の平均投稿評価値の近さを調べることにする。そのために、常連投稿者で投稿者別平均投稿評価値の近いもの同士を同じグループとして 8 つのグループとし、また、銘柄別に、銘柄の平均投稿評価値の近い銘柄同士を同じ 5 つのグループとし、それぞれのグループのコレスポンデンス分析で、どのように分布しているかを調べた図を、図 6.3 に示す。

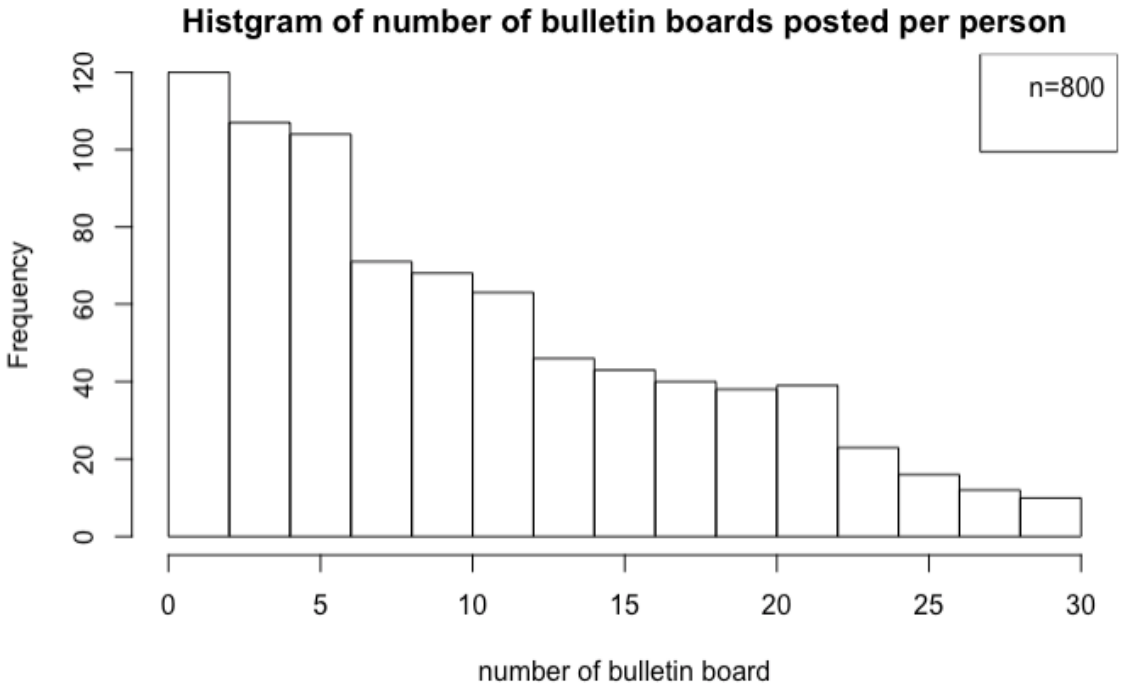


図 6.2: 30 銘柄中、常連投稿者一人当たりの投稿した掲示板銘柄数（2015 年 1 月から 2016 年 12 月）。中央値:8.0

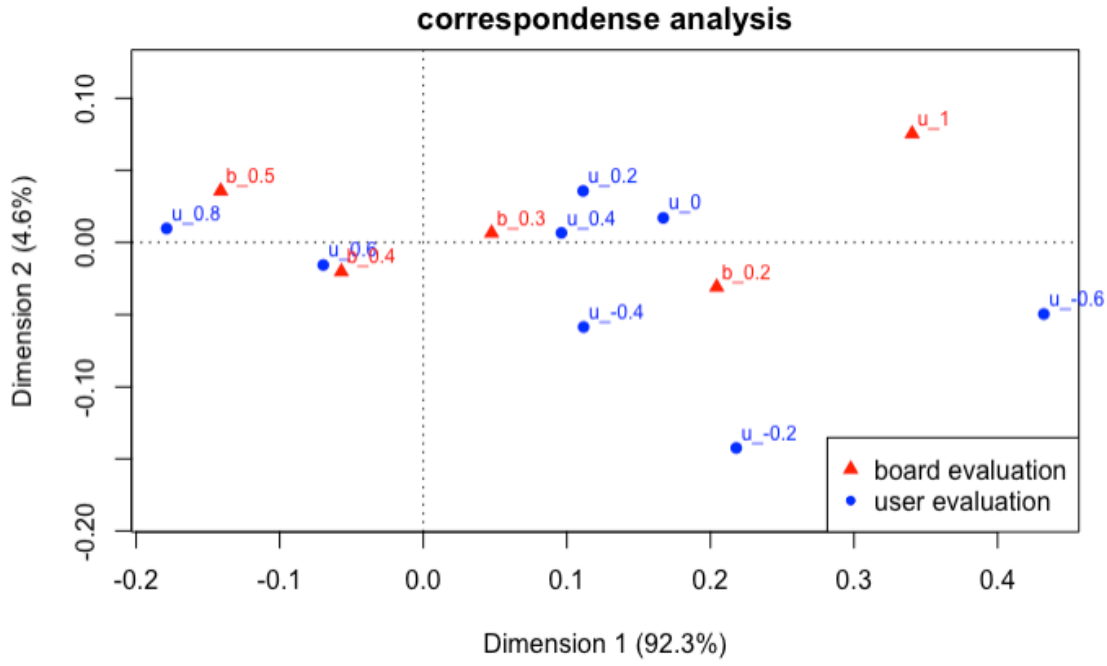


図 6.3: 常連投稿者と掲示板の投稿評価値のコレスポンデンス分析。それぞれ、投稿評価値別に 8 グループ、5 グループにグループ化を行い、コレスポンデンス分析による対応分析を行った。プロットの接頭辞 b_は掲示板、u_は常連投稿者を示し、接尾辞の数値は投稿評価値の下限値である。

図 6.3 は、それぞれのグループの近さを 2 次元平面上に表した図であり、三角のプロットが銘柄のグループ、丸のプロットが常連投稿者のグループを示している。平均投稿評価値は常連投稿者は 0.2 刻みで、銘柄は 0.1 刻みで 1.0 から-1.0 までのグループに分けてグループ化した。その結果、常連投稿者を 8 つのグループ、銘柄別掲示板を 5 のグループに分けた。ここで平均投稿評価値が 0 となるグループは除外した。

プロットの添字はグルーピングした際の投稿評価値を識別する値を示しており、例えば▲b_0.5 は、銘柄の平均投稿評価値が 0.5 以上の銘柄のグループ、●_0.2 は投稿者別平均投稿評価値が 0.4 未満かつ 0.2 以上の常連投稿者のグループである。数字が大きいほど、そのグループの平均投稿評価値は高く、数字が低いほどそのグループの平均投稿評価値は低いことを示している。

図 6.3 の結果を見ると、横軸が負の方向には、それぞれ、銘柄、常連投稿者のグループとも、平均投稿評価値が高いグループが集まっている。逆に、横軸が正の方向には、それぞれ、銘柄、常連投稿者のグループの平均投稿評価値が低いグループが集まっている。コレスポンデンス分析では、同じ傾向を持つものは近い位置にプロットされるということを踏まえると、投稿評価値が低い銘柄には、投稿者別平均投稿評価値が低い投稿者が集まり、投稿評価値が高い銘柄には、投稿者別平均投稿評価値が高い投稿者が集まることが読み取れる。

以上のことから、投稿評価値が高い掲示板とは、株価収益率が高いか株価ヒストリカル・ボラティリティが高い銘柄であり、この銘柄は利益が得やすい銘柄となっている。それゆえ、投稿内容がポジティブな内容が増え、結果的に投稿評価値が高い投稿者が集まることがわかった。

## 6.2　リサーチクエスチョンへの回答

本研究を進めるにあたり、設定したリサーチクエスチョンについて以下に回答する。

**SRQ1:「株価は投稿にどのように影響を与えるか？」**

株価の指標である平均月次株価収益率および平均月次株価 HV が高くなると掲示板の投稿の信頼度は上昇し、平均月次売買代金が多くなると掲示板の投稿の信頼度は減少することがわかった。これは、株価が投資家の利益の得やすい状況になると、信頼度が高い投稿が増えるということになるのではなかと推測される。

**SRQ2:「投稿の信頼度はどのように分類されるか？」**

投稿者は信頼度の高いグループと低いグループに分類される。さらに、掲示板の信頼度と投稿者の信頼度には正の相関関係が見られる。すなわち、信頼度の高い掲示板には信頼度の高い投稿者が集まることがわかった。また、自然言語処理による投稿のネガポジ分析から、投稿の信頼度と投稿ネガポジ値には正の相関がみられ、ポジティブ

な感情の投稿ほど信頼度が高いことがわかった。

**SRQ3:「投稿の信頼度はどのようにモデル化されるか？」**

信頼度を予測するために、信頼度が正か負であるかの 2 値分類モデルを決定木により作成した結果、信頼度は投稿者の信頼度で決定されることがわかった。すなわち信頼度の高い投稿者の投稿は、常に信頼度が高いというモデルとなった。

**MRQ:「投稿の信頼度はどのようにモデル化され、そのモデルから株価は予測できるか？」**

投稿の信頼度予測は、信頼度が正か負であるかの 2 値分類を決定木によりモデル化され、そのモデルの説明変数は投稿者の信頼度のみである。信頼度の高いグループと低いグループの投稿者による翌営業日株価収益率の予測結果は、投稿者の高いグループの予測の正解率が、信頼度の低いグループの予測の正解率に比較し高いことがわかった。

## 6.3 まとめ

本研究では、textream 掲示板を対象に、2015 年 1 月から 2016 年 12 月までの掲示板への投稿データを用い、投稿数の多い上位 30 銘柄の掲示板の 1,000 投稿以上行なっている 800 名の常連投稿者の投稿を分析し、投稿の信頼度をモデル化した。

投稿の信頼度は、投稿に付与された投稿評価値で定量化し、これを目的変数とし、説明変数を株価の指標である、株価収益率、株価ヒストリカル・ボラティリティ、売買代金と、常連投稿者の投稿評価値、掲示板の投稿内容のネガポジ値を説明変数として決定木による予測モデルの構築を、2015 年 1 月から 2016 年 12 月までの掲示板データと株価データを用いて行った。このモデルを 2017 年 1 月から 2017 年 6 月までのデータによって検証した結果、投稿評価値の予測は常連投稿者の投稿評価値、すなわち、投稿者自身の信頼度のみによって決定されることがわかった。つまり、信頼度の高い投稿者の投稿は信頼度が高いと予測され、逆に信頼度の低い投稿者の投稿は低いと予測される結果となった。

このモデルを用い、常連投稿者のうち、信頼度の高い 40 名と信頼度の低い 40 名で、翌日株価収益率の予測性能を検証したところ、信頼度の高い常連投稿者の予測性能が高いことがわかった。すなわち、信頼度の高い情報は信頼度の高い人から得ることができ、その信頼度の高い情報の株価の予測性能は高いということが言える。

## 6.4 今後の展望

本研究では、投稿数上位 30 銘柄の株式掲示板の投稿について、株価、投稿者、銘柄、投稿内容の観点から、投稿の信頼度の予測についての検討を行った。30 銘柄に限った投稿の信頼度は、常連投稿者の信頼度に依存するという結果が得られたが、他の銘柄や常連投稿者

以外の投稿者の信頼度の予測については考慮外であった。そのため、本研究でのモデルが一般的なものであるかどうかの検証ができていない。

今後、全銘柄の予測や全投稿者を含めた予測などを行い本研究でのモデルが一般的なものであるかどうかの検証を行う必要がある。さらには、株価掲示板以外の投稿内容について、信頼度を定義できるような予測モデルの構築が期待される。

昨今、フェイクニュースの存在が指摘され、定性データの信頼性が求められることも多い。そのためにも、一般的な定性データの信頼度予測のモデル構築が期待される。

# 謝辞

本論文は、筆者が北陸先端科学技術大学院大学先端科学技術研究科前期博士課程在学中の研究成果をまとめたものである。本研究を進めるに当たり、ご指導いただいた主指導教員である Dam Hieu Chi 准教授に感謝いたします。並びに、内平直志教授、神田陽治教授、伊藤泰信准教授、白肌邦生准教授には東京サテライトにてご指導いただき、大変ありがとうございました。

また、在学中に東京サテライトの同窓生の方々には有用な助言等いただくとともに、学生生活においても非常に楽しく過ごさせていただきました。感謝いたします。

最後に、学生生活中、生活を支えていただいた妻千絵子に感謝の意を表します。

# 参考文献


[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau, *Sentiment analysis of twitter data*, In Proceedings of the Workshop on Languages in Social Media,LSM ' 11 (2011).

[2] Hirotogu Akaike, *Information theory and an extension of the maximum likelihood principle*, pp. 199–213, Springer New York, New York, NY, 1973.

[3] J. Bollen, H. Mao, and X. Zeng, *Twitter mood predicts the stock market*, Journal of Computational Science **2** (2011), no. 1.

[4] Damian Jimenez, *Towards Building an Automated Fact-Checking System*, SIGMOD' 17 Student Research Competition (2017).

[5] Rubin V. L, *Deception detection and rumor debunking for social media*, Sloan L. and Quan-Haase A. (2017).

[6] Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang, *Fake news detection through multi-perspective speaker profiles*, Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (Taipei, Taiwan), Asian Federation of Natural Language Processing, November 2017, pp. 252–256.

[7] Li M. Long Y. Lu, Q. Xiang R. and Huang C.R., *Fake news detection through multi-perspective speaker profiles*, 2017.

[8] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein, *A stylometric inquiry into hyperpartisan and fake news*, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Melbourne, Australia), Association for Computational Linguistics, July 2018, pp. 231–240.

[9] Manuel Gomez-Rodriguez Arpit Merchant Sebastian Tschiatschek, Adish Singla and Andreas Krause, *Detecting fake news in social networks via crowdsourcing*, CoRR, abs/1711.09025 (2017).

[10] Jin Yea Jang Svitlana Volkova, Kyle Shaffer and Nathan Oken Hodas, *Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter.*, (2017).

[11] 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 (ACL2005) (2005).

[12] et al Tsubasa Tagami, Hiroki Ouchi, *Suspicious news detection using micro blog text*, arXiv preprint, arXiv:1810.11663 (2018).

[13] Peter D. Turney and Michael L. Littman, *Measuring praise and criticism: Inference of semantic orientation from association*, ACM Trans. Inf. Syst. **21** (2003), no. 4, 315–346.

[14] Yimin Chen Victoria L Rubin and Niall J Conroy, *Deception detection for news: three types of fakes*, Proceedings of the Association for Information Science and Technology 52(1):1-4 (2015).

[15] Andreas Vlachos and Sebastian Riedel, *Fact checking: Task definition and dataset construction*, LTCSS@ACL. (2014).

[16] You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, and Cong Yu, *Computational Fact Checking through Query Perturbations*, ACM Transactions on Database Systems **42** (2017).

[17] アレックス・ペトランド，小林啓倫訳, **ソーシャル物理学**, 2015.

[18] 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 (2015).

[19] 加藤 恒昭，三木 光範, **自然言語処理（情報工学テキストシリーズ５）**, 共立出版, 2014.

[20] 兵庫県立大学，応報情報科学研究科講義資料, `https://www.ai.u-hyogo.ac.jp/~arima/lectures/JT-13.pdf`, 2018/11/27 取得.

[21] 和泉潔，後藤卓，松井藤五郎, **テキスト情報を用いた金融市場分析の試み**, 人工知能学会全国大会論文集 (2008).

[22] 坪内 孝太，山下 達雄, **株価掲示板データを用いたファイナンス用ポジネガ辞書の生成**, 人工知能学会全国大会（第 28 回） (2014).

[23] 宮崎邦洋，松尾豊, **株価掲示板におけるユーザ行動異常検知を用いた相場操縦発見手法に関する研究**, 第 15 回 人工知能学会 金融情報学研究会 (2015).

[24] 山下 達雄，坪内 孝太, **個人の予測信頼度を加味した株価掲示板情報からの株価予測**, 人工知能学会全国大会（第 30 回） (2016).

[25] 日本経済新聞社, **日経平均株価はどう計算しているのですか**, https://support.nikkei.com/app/answers/detail/a_id/451/~/日経平均株価はどう計算しているのですか, 2019/05/07 取得.

[26] 日本証券協会, **金融・証券用語集**, http://www.jsda.or.jp/manabu/word/word10.html, 2018/11/27 取得.

[27] 日経 BP, **判例で理解する *it* 関連法律**, https://tech.nikkeibp.co.jp/it/article/COLUMN/20101115/354134/, 2019/05/07 取得.

[28] 中村 明, **感情表現辞典**, 東京堂出版, 1993.

[29] 東京大学教養学部統計学教室, **統計学入門（基礎統計学Ⅰ）**, 東京大学出版会, 1991.

[30] 元田 浩，山口 高平，津本 周作，沼尾 正行, **データマイニングの基礎** *(it text)*, オーム社, 2006.

[31] 福井正康, **経営統計学基礎資料**, http://www.heisei-u.ac.jp/ba/fukui/pdf/stattext09.pdf, 2018/11/27 取得.

[32] 豊田秀, **購買心理を読み解く統計学**, 東京図書, 2006.

[33] 高橋宏圭，関和広，上原邦昭, **株価回帰と *web* ニュース記事分析を組み合わせた株価動向推定**, 電子情報通信学会技術研究報告 **111** (2012).

**発表論文**

鞠勝彦、DamHieu Chi, 2018, ”株式掲示板における信頼度予測”, 情報処理学会第 81 回全国大会, 2018/3

# 付 録A　分析対象の銘柄一覧

本研究の本調査で使用した、銘柄一覧及び投稿数、投稿者数の一覧を示す。

表 A.1: 学習データの一覧。2015 年 1 月 1 日から 2016 年 12 月 31 日までの投稿数が多い上位 30 銘柄の掲示板の、投稿の投稿数及び投稿者数を表示。

| 銘柄コード | 銘柄名 | 投稿数 | 投稿者数 |
|---|---|---|---|
| 2121 | (株) ミクシィ | 96,995 | 4,561 |
| 2138 | クルーズ (株) | 96,651 | 3,363 |
| 2315 | (株) カイカ | 105,267 | 4,378 |
| 2321 | (株) ソフトフロントホールディングス | 117,556 | 3,690 |
| 3664 | (株) モブキャストホールディングス | 168,102 | 6,079 |
| 3692 | (株) ＦＦＲＩ | 112,493 | 3,919 |
| 3753 | (株) フライトホールディングス | 78,027 | 3,113 |
| 3782 | (株) ディー・ディー・エス | 96,244 | 2,978 |
| 3823 | (株) アクロディア | 97,182 | 4,336 |
| 3903 | (株) ｇｕｍｉ | 118,120 | 4,703 |
| 3914 | ＪＩＧ－ＳＡＷ (株) | 112,360 | 4,480 |
| 4080 | (株) 田中化学研究所 | 95,027 | 2,984 |
| 4347 | ブロードメディア (株) | 85,651 | 2,729 |
| 4563 | アンジェス (株) | 146,212 | 5,113 |
| 4564 | オンコセラピー・サイエンス (株) | 118,609 | 4,648 |
| 4565 | そーせいグループ (株) | 29,813 | 7,726 |
| 4571 | ナノキャリア (株) | 86,060 | 2,864 |
| 4572 | カルナバイオサイエンス (株) | 97,877 | 3,432 |
| 4777 | (株) ガーラ | 223,720 | 8,427 |
| 6079 | (株) エナリス | 134,062 | 4,584 |
| 6176 | (株) ブランジスタ | 77,719 | 3,525 |
| 6502 | (株) 東芝 | 131,860 | 6,992 |
| 6753 | シャープ (株) | 196,026 | 7,912 |
| 6871 | (株) 日本マイクロニクス | 81,373 | 2,436 |
| 7211 | 三菱自動車 (株) | 79,755 | 5,717 |

| | | | |
|---|---|---|---|
| 7974 | 任天堂 (株) | 216,505 | 9,451 |
| 8462 | フューチャーベンチャーキャピタル (株) | 390,341 | 8,025 |
| 8789 | フィンテック | 146,837 | 4,196 |
| 9501 | 東京電力ホールディングス (株) | 197,049 | 3,781 |
| 9984 | ソフトバンクグループ (株) | 99,237 | 4,642 |

表 A.2: 検証データの一覧。2015 年 1 月 1 日から 2016 年 12 月 31 日までの投稿数が多い上位 30 銘柄の掲示板の、2017 年 1 月 1 日から 2017 年 6 月 30 日までの投稿の投稿数及び投稿者数を表示。

| 銘柄コード | 銘柄名 | データ数 | 投稿者数 |
|---|---|---|---|
| 2121 | (株) ミクシィ | 12,652 | 697 |
| 2138 | クルーズ (株) | 2,707 | 318 |
| 2315 | (株) カイカ | 15,959 | 1,349 |
| 2321 | (株) ソフトフロントホールディングス | 11,016 | 649 |
| 3664 | (株) モブキャストホールディングス | 97,717 | 3,449 |
| 3692 | (株) ＦＦＲＩ | 3,801 | 401 |
| 3753 | (株) フライトホールディングス | 30,959 | 1,534 |
| 3782 | (株) ディー・ディー・エス | 85,993 | 3,563 |
| 3823 | (株) アクロディア | 6,691 | 373 |
| 3903 | (株) ｇｕｍｉ | 27,438 | 1,337 |
| 3914 | ＪＩＧ－ＳＡＷ (株) | 9,391 | 603 |
| 4080 | (株) 田中化学研究所 | 2,987 | 244 |
| 4347 | ブロードメディア (株) | 9,865 | 360 |
| 4563 | アンジェス (株) | 53,437 | 3,564 |
| 4564 | オンコセラピー・サイエンス (株) | 7,962 | 648 |
| 4565 | そーせいグループ (株) | 61,896 | 2,473 |
| 4571 | ナノキャリア (株) | 15,886 | 699 |
| 4572 | カルナバイオサイエンス (株) | 8,935 | 390 |
| 4777 | (株) ガーラ | 9,561 | 1,071 |
| 6079 | (株) エナリス | 17,378 | 757 |
| 6176 | (株) ブランジスタ | 12,946 | 899 |
| 6502 | (株) 東芝 | 355,249 | 9,235 |
| 6753 | シャープ (株) | 74,522 | 3,088 |
| 6871 | (株) 日本マイクロニクス | 13,345 | 332 |

| | | | |
|---|---|---|---|
| 7211 | 三菱自動車 (株) | 5,832 | 587 |
| 7974 | 任天堂 (株) | 76,583 | 3,161 |
| 8462 | フューチャーベンチャーキャピタル (株) | 18,714 | 1,307 |
| 8789 | フィンテック | 36,466 | 972 |
| 9501 | 東京電力ホールディングス (株) | 26,967 | 696 |
| 9984 | ソフトバンクグループ (株) | 42,354 | 1,873 |

# 付 録B　ネガポジ辞書

本研究で作成したネガポジ辞書の、ポジティブ及びネガティブのＬ２正則化回帰係数を有する上位 25 単語を示す。

表 B.1: ポジティブ単語一覧（上位 25 単語）

| 単語 | L2 正則化回帰係数 |
|---|---|
| 北斗 | 1.4075114524355 |
| (● | 1.40746452489239 |
| 三菱財閥 | 1.4074634914318 |
| 瓜 | 1.40743000713768 |
| 安倍自民党 | 1.40741494889633 |
| 朝日 | 1.40741030572554 |
| (｀_´)ゞ | 1.4074020443935 |
| 低迷 | 1.40740010651698 |
| クソミンス | 1.40739570529269 |
| 予約 | 1.4073949914834 |
| 師 | 1.4073878549775 |
| 電 | 1.4073866721168 |
| アジア | 1.40738650239727 |
| ～！ | 1.40738637187994 |
| (￣▽￣ ;) | 1.4073858502497 |
| 賛成 | 1.40738527914612 |
| チーム | 1.40738426967946 |
| 波 | 1.40738323177155 |
| 手遅れ | 1.40738286396322 |
| 流れる | 1.4073816236482 |
| ゼロ | 1.40737947332302 |
| 乗せる | 1.40737941491371 |
| ｸﾞｯ | 1.40737902459245 |
| 生まれる | 1.40737900474613 |
| ( 👆 | 1.40737897548248 |

表 B.2: ネガティブ単語一覧（上位 25 単語）

| 単語 | L2 正則化回帰係数 |
|---|---|
| 甲状腺 | -1.80043940017999 |
| インチキ | -1.80027325389129 |
| 公衆便所 | -1.80024957344196 |
| 乞食 | -1.80019677467269 |
| ダニ | -1.80019220521428 |
| 汚染 | -1.80019180226207 |
| ＼ | -1.80014743477515 |
| 柏崎刈羽 | -1.80014029351909 |
| [ | -1.80013736837402 |
| (´・ ‿・‘) | -1.80013664804856 |
| クレーンゲーム | -1.80013541106201 |
| 新潟県 | -1.80013106119371 |
| ガハ | -1.80012871576391 |
| 妖怪 | -1.80011773411134 |
| 液晶 | -1.80011740974874 |
| ｊｄｉ | -1.80011660824669 |
| (≧∇≦) | -1.80011297053162 |
| 放電 | -1.800109146835 |
| ｑ | -1.80010850216667 |
| ^ | -1.80010762439514 |
| 嵌め込む | -1.8001066537244 |
| 吊り上げる | -1.80010650374257 |
| 株式会社 GABA | -1.8001060122949 |
| 試作 | -1.80009977751529 |
| すぎる | -1.80009943219816 |

# 付 録C　K-分割交差検証

図 C.1 に k=5 としたときの交差検定の概念図を示す。

1 回めでは 5 分割したデータの最初の分割データを検証データとし、残りの 4 つのデータで学習を行い、学習結果の検証を検証データで行う。これを 5 回繰り返し、学習したパラメタの平均値を用いる。

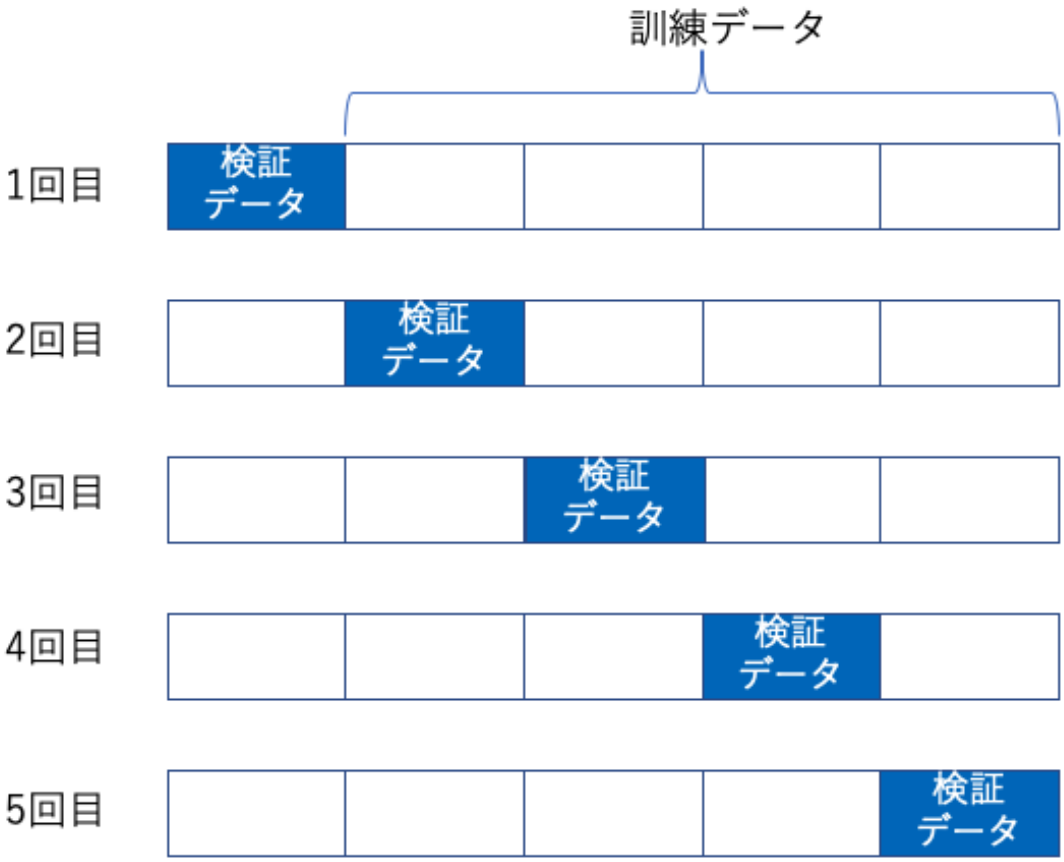


図 C.1: 交差検定の概念図、図は k=5 のときの例