1J5-OS-13b-2in

株価掲示板情報の感情解析と株価との相関の研究

Correlation between Stock Price and Posi / Nega Score from Stock Bulletin Boards Data

坪内　孝太　　山下　達雄
Kota Tsubouchi　Tatsuo Yamashita

Yahoo! JAPAN 研究所
Yahoo! JAPAN Research

This paper researches on correlation between stock price and Posi / Nega score from stock bulletin boards data. Most marketer consider the stock bulletin boards when they judge the stock action such as buy and sell. It means that valid information for estimating stock price are contained in the bulletin board. This paper focused on the Posi / Nega information of stock bulletin board and research the correlation between stock price and the data. The result of the test shows that the correlation are confirmed actually, and concludes that this method are useful for estimating the future stock price.

1. 序論

株価掲示板に投稿された情報を解析し、株の値動きを予測する事を目指す。

株価掲示板情報とは、特定の株式銘柄について議論されている情報である。対象銘柄に対するコメントに加え、「買いたい」や「売りたい」といった投資行動がメタデータとしてひもづいている情報をさし、本稿ではそれらのメタデータを銘柄に対する「感情」と呼称する。坪内ら [坪内 14] は、株価掲示板情報の投稿コメントと、感情スコアがセットになっているデータを解析することで、株価掲示板情報を解析するために基礎となる感情辞書を生成し、感情スコアがついていない記事に対しても感情スコアを予測できる手法の開発に成功した。

本稿は、感情スコアと、実際の株価の動きとの相関を調べる事を目的とする。株価掲示板の情報や株の値動きを時間軸でいくつかのパタンに分け、両者の関連を調べる。

2. Yahoo! 株価掲示板情報

本研究では Yahoo! 株価掲示板情報を対象とする。Yahoo! 株価掲示板は、以下の情報からなる。

対象銘柄 何の銘柄についての情報か

タグ情報 強く買いたい、買いたい、様子見、売りたい、強く売りたいの５種

タイトル 投稿のタイトル

テキスト 投稿の本文

投稿日時 記事を投稿した日時

投稿者 記事を投稿したユーザの ID（解析には匿名化）

対象銘柄についての個別の議論がなされている事と、対象銘柄に対する投資意向を示すタグがセットで投稿されている点が他の一般的な掲示板とは異なっているといえる。

本研究では約 2 年分の投稿データを使用した。

連絡先: 坪内孝太，ヤフー株式会社 Yahoo! JAPAN 研究所，東京都港区赤坂 9-7-1　ミッドタウンタワー，ktsubouc@yahoo-corp.jp

期間 2012-11-21 から 2014-11-17 (727days)

投稿数 14,015,844

以降、5 種のタグ情報「強く買いたい、買いたい、様子見、売りたい、強く売りたい」をそれぞれ「2,1,0,-1,-2」の数値に置き換え「感情スコア」と呼称する。

3. 感情スコアの抽出方法

対象銘柄の特定の１日に付与されている感情スコアと、株価との相関を調査する方法について述べる。本研究では、感情スコアの定義で 4 種類、１日の時間の区切り方の定義で 4 種類を掛け合わせることで、計 16 種類の感情スコアを準備した。

3.1 感情スコアの定義

感情スコアは、投稿毎に付与されているスコアであるが、そのまま使うためには２つの課題がある。まず、ユーザ毎の評価のデータのばらつきである。ユーザによって付与する感情スコアの平均値にばらつきがあるため、それを補正する必要が考えられる。次に、スパムユーザの排除が必要となる。たとえば、宣伝や罵倒、株価操作などを目的としたノイズとなる投稿を行うユーザである。たとえば、人が買いたくなるような嘘や誇張された情報を書き込み、株価をつり上げ、自分は売るという行為をするユーザが少なからずいる。彼らの感情スコアは意図をもってつけられたスコアであり、信頼性に欠けるため排除される必要がある。

3.1.1 平均感情スコア

対象とする銘柄の 1 日における集計期間毎の感情スコアの平均をさす。最も基本的なスコアである。

3.1.2 補正平均感情スコア

平均感情スコアに対してユーザ毎の投稿特性によって補正されたスコアの平均をさす。投稿された記事のスコアをユーザの過去の投稿スコアの平均値から差し引く事で、補正した。

3.1.3 潔白ユーザ平均感情スコア

過去にスパム投稿と判断された投稿が１件でもあれば、そのユーザを除外する。残ったスパム投稿の経験が無いユーザ（潔白ユーザと定義）のみの平均感情スコアを用いる。なお、本投稿におけるスパム投稿か否かの判断は、ユーザによる通報を元に運営側が黙視による判定を行い、判断した結果を採用している。

3.1.4 潔白ユーザ補正感情スコア

潔白ユーザの、さらにユーザ毎の投稿特性のばらつきで補正された感情スコアの平均値を用いる。

3.2 1日の時間の区切り方の定義

今回は日本の株式市場を対象とするが、当該市場は、9 時から 15 時まで市場が開催している。そこで、株式市場の開催時間に応じたデータの区切り方を目指す。

3.2.1 終日の反応時間枠：0 時ー 24 時

単に同日の終日の反応時間枠をとる。取引中、事前、事後、すべての反応がスコアに含まれている。

3.2.2 取引中の反応時間枠：9 時ー 15 時

本時間枠によって求められれるスコアは、取引中のみにしぼった反応をスコア化したものである。

3.2.3 事前の反応時間枠：前日 15 時ー 9 時

前日の市場がクローズしてから、当日の市場が開くまでの期間の反応のみを集めて、スコア化したものである。

3.2.4 事後の反応時間枠： 15 時ー翌日 9 時

当日の市場がクローズしてから翌日に取引が始まるまでの期間の反応をスコア化したものである。

3.3 株価データの加工

前節には感情スコアの抽出方法について述べたが、このスコアと突合させる株価データには正規化処理を施した。具体的には前日終値に対する当日の終値の比率を株価データとして用いる。銘柄毎に平均株価が異なるため、株価の値をそのまま使うと、銘柄毎の平均的な株価の値に結果が大きく左右される可能性があるためである。なお、突合される株価データには、オープン時の株価、クローズ時の株価、当日の最高値、最安値の 4 種類を用いた。

4. 実証実験

実証実験には、株価掲示板に対する感情ありの投稿が 1 件以上ある日が 500 件以上ある銘柄、13 銘柄を対象とした。すなわち、13 銘柄×スコア 4 種×時間の区切り方 4 種×株価 4 種＝ 832 通りの相関を調べる。なお、回帰は線形回帰を前提とし、得られた回帰直線の平均二乗誤差の値をスコアとして用いた。この値は小さければ小さいほどモデルの予測誤差が小さい事を示している。

性能の良かったモデルの例を図 1 に示す。これは、取引中の感情スコアと株価（終値）の相関を示したものであるが、両者には相関が見られる事が確認できる。

表 1 に、時刻の区切り方 4 種と感情スコアの計算方法 4 通りにおける平均二乗誤差を示す。表の平均二乗誤差は、各項目において 13 銘柄および 4 種類の株価データそれぞれのケースにおける相関をとったものである。

表より、ユーザ毎の補正をかける方が高精度に予測できている事がわかる。ユーザの投稿の傾向は様々で、それをユーザ毎に補正する事は高精度な回帰モデルの構築に役立っている事がわかる。

しかし、潔白ユーザを考慮する事によりモデルの精度は低下する事がわかる。これは、スパム投稿の判定をユーザの報告と黙視に頼っており、不十分であることが原因と推察できる。また、活発でまじめなユーザであっても一時的に熱くなってスパム認定されてしまう投稿をしてしまうことがあり、今回の基準では排除されてしまう。つまり、正しいユーザを省く、本来スパムを省けないなどのようになる、データの数や精度が不十分となり、このような結果になる。したがって、煽り等のスパム投稿を正しく省く事は今後の課題と言える。

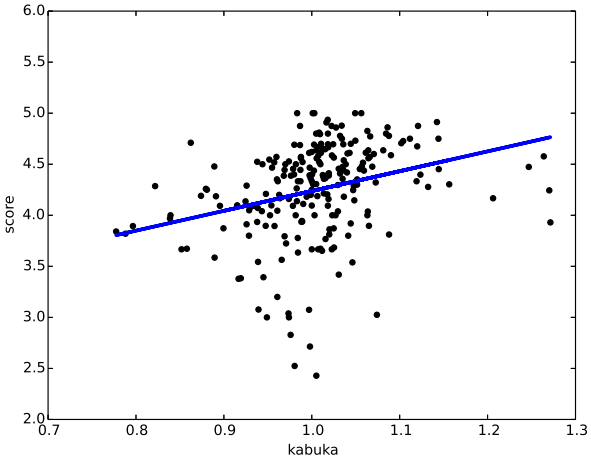

図 1: 性能の良かったモデルの例

終日の反応や取引時間帯の反応をインプットとするとモデルは株式市場の値動きを直接表現しているモデルであるため、当然モデルの精度は高くなるはずである。興味深いのは、事前の情報だけでも平均二乗誤差は 10 %ー 20 %程度の精度低下となっていることである。この結果は、市場が開く前の掲示板情報から当日の相場観を示す要素が含まれており、その結果、当日の株価の値動きと相関している事を示唆している。

なお、株価 4 種類（オープン時、クローズ時、高値、安値）についての解析も行ったが、各モデルにおいて若干の際はあるものの、共通して考察できるような有為な特徴の差はみられなかった。

表 1: スコア導出手法毎の性能比較 (MAE)

	終日	取引中	事前	事後
平均	0.4019	0.4621	0.5471	0.5232
補正	0.1024	0.1252	0.1387	0.1291
潔白平均	0.7757	0.8270	0.9733	0.9619
潔白補正	0.2306	0.2501	0.2910	0.2918

5. 結論

株価掲示板の情報と株価の連動について相関を調べ、感情スコアが市場の動きを表現したり、翌日の相場観を示すのに有効である事を確認した。

特に翌日の相場観との相関の示唆は、今後株価を予測する際に掲示板情報から得られた感情スコアが有用である事を示唆しており、非常に興味深い知見と言える。

今後の課題としては、ユーザ毎の補正の仕方や、スパムユーザ判定の高精度化などがあげられる。

参考文献

[坪内 14] 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” , 2014 年度人工知能学会全国大会講演集, 3L4-OS-26b-1in.