Twitterの多軸的感情情報を利用した株価の予測

増井 佑亮　　藤野 巖
東海大学大学院情報通信学研究科情報通信学専攻
6bjnm010@mail.u-tokai.ac.jp, fujino@tokai.ac.jp

1 はじめに

近年ではスマートフォンの普及と共に誰もが容易に情報を発信できるようになり、インターネット上のデータが日々肥大化している。そのためそれらデータを分析するビッグデータ解析や機械学習といった分野に注目を集めている。その様なビッグデータのリソースの一つに Twitter がある。Twitter は日本人の利用率は3人に1人以上と広く一般に普及しており、またツイートは140字という短文の為トピックが絞られやすい特性や感情が頻出しやすい傾向にある。その為、大量のテキスト情報を活用しデータに潜むより高次な感情情報の分析が注目されている。

一方、実際に株式投資の世界では会社の業績結果の報告以上に株価が上昇するようなケースがよくある。そのような場合、企業に対する期待や関心といった投資家の感情が大きく関係している。また株価に限らず一般的に、あるグループで関心のあるトピックで意見交換や議論された場合、一つの意見に集約され極端な方向に行きやすい傾向にある。これは心理学でいう情報的影響の「同調」や「集団極性化」と呼ばれる現象が関係してくる。株式市場では、相場の流れに人々の感情が動かされて便乗した結果、株価が異常なほど一方向に動くことがあり、このような株価が急上昇している場合、世間の感情そのものが材料になっているケースが多くある。

これらの背景から、本研究は Twitter の感情情報から特定銘柄の株価を予測することを試みた。具体的には、まず準備としてツイートのテキスト情報を取得し、銘柄と関連のある特定のキーワードが含まれるものを抽出する。感情情報を測定するにあたり既存の感情辞書を元に Twitter のテキスト情報から得られた感情表現を拡張した独自の感情辞書を使う。そして、その感情辞書を用いてツイート分析によって得られた各軸の感情指数値と一つ前の週の株価を説明変数とし、目的変数はその週の株価の値とし重回帰分析を行う。得られた回帰係数をもとに来週の株価の予測し、得られた結果について独自の評価指標を用いて評価した。最後に評価指標の値が最小となるように学習データの時間範囲と感情軸を選択し予測精度を高めた。そのうえで、Twitter の感情情報を加味しない単回帰分析と感情軸を減らした低減感情軸を用いた重回帰分析と比較することで本研究の有用性を示す。

2 先行研究

関連の先行研究としては、Bollen 氏らの Twitter データを利用した株価の予測に関する研究、佐藤らの Web ニュースデータと極性辞書を利用した研究および五島氏らの金融向けの極性辞書を構築した研究を取り上げる。

Bollen 氏らの研究[1]は、英語の感情6軸とネガポジの2軸を利用した株価の予測を試みた研究である。中でも英語に於いて「穏やかさ」を示す軸とは相関が高く86．7%と高い予測結果を弾きだしたと結論づけられている。この結果から Twitter との何らかの因果関係は示唆されたが厳密な関係は不明瞭だった。また、この研究では英語をベースに研究されたが、日本語をベースにした場合、ユーザー層や文法構造などの違いから同じ結果に出せるか、などの疑問が残る。それらの結果を踏まえ検証した研究に佐藤氏らの研究[2]がある。この研究では、ネガティブとポジティブの2軸を利用しニュース記事との相関関係を研究したものだったが、明瞭な結果を得ることができなかった。当該論文ではニュース記事に比べ SNS は感情が頻出しやすいこと、感情を多軸化して、より細やかな感情分析が必要なことが示唆されていた。しかし多軸の日本語を利用して予測する研究はあまりない。また、最近では五島氏らが、株価の情報をもとに金融に適した極性辞典を作成する[3]という報告もある。この研究では、金融分野に関連のあるニュースを用いて教師あり学習によって金融に特化した極性辞書を作成した。その結果、予測に有効であることを示しニュース記事が何らかの形で株式に影響を与えていることが示唆された。そこから日本語に於いても株価の予測が可能であると考えた。

3 提案手法

3.1 研究概要

本研究のシステムの全体像を図1に示す。今回作成したシステムは、Twitter の情報を取得し、その情報をデータベースに保存する段階、取得した感情情報について感情解析を行なう段階とその結果について重回帰分析を行う段階に分かれる。更にそこから得られた結果と実測値を比較する指標を作り、その株価の銘柄に適した時刻の範囲と感情軸を選定するようにした。

本研究のシステムを構築するにあたり、Twitter Streaming API[4]を利用し無作為に抽出されたサンプルツイートのアカウント名、ツイート内容、日付や位置情報などの情報をデータベース Couch DB[5]に保存した。そのデータベースに保存されたツイートから銘柄と関連のある特定のキーワードを含むもののみを抽出した。そのデータを週ごとの時系列にまとめ、MeCab[6]を用いて文章中の単語をすべて原形に戻して記録した。その時系列情報を基に重回帰分析を行い、特定の時間範囲を学習データとしそこで求めた回帰係数と一つ先の時刻のツイートの感情情報を説明変数とし、目的変数を株価として予測を試みた。その後、テストデータの結果を独自の予測精度の評価指標で分析した。その上で、より高い精度で予測ができるように学習データの時間範囲の調整を行い、より高い精度に予測できるようにした。

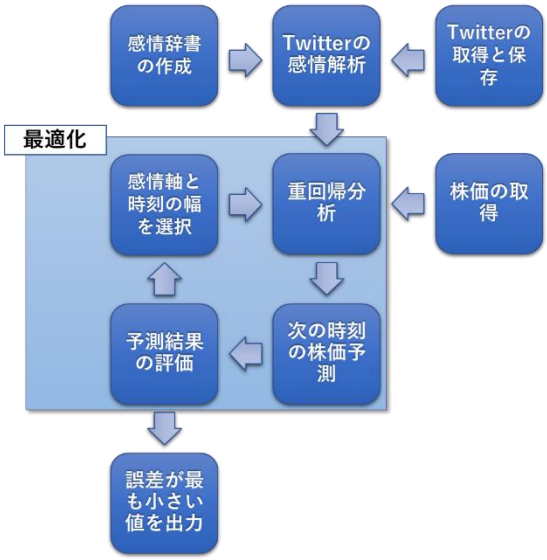

図1 システムのブロック図

3.2 感情辞書の拡張

本研究では、当初、中村氏編感情表現辞典[7]を参考に、その辞書内に記載されている単語や慣用句を基に喜・怒・哀・怖・恥・好・厭・昂・安・驚の10 軸に感情表現を分類した。しかし、1993年に出版された文学作品を基に作成された当該辞書では、流行の絵文字、顔文字や単語など最先端のトレンドを常に吸収し発信する傾向にある Twitter のテキストデータでは10分に対応していないように思われた。そこで本研究では既存の感情辞書にインターネットに用いられる固有表現を拡張することで Twitter 上の感情表現に対応できるように、より実用的な感情辞書を拡張した。

具体的な作業としては、被験者4人にランダム抽出された同じ範囲内のツイートを読んでもらう。各々がそのツイート内に10軸の内のいずれかの感情表現が含まれていると感じた絵文字を含む文字表現をそれぞれの各軸に書き出してもらう。そのデータを基に、被験者の書き出した文字表現を一つに集計し、各軸に2つ以上同じ文字表現があった場合は、各軸で重複が起きないように考慮し感情辞書に追加登録した。

3.3 重回帰分析による株価予測

社会現象を捉えるためには、結果となる現象とそれを引き起こす要因となる感情を捉え、その結果を引き起こすと考えられる複数の要因を結びつけるモデルの構築が必要となる。そこで本研究では重回帰モデルが最適なモデルと考え、株価の予測を試みた。重回帰分析は、予測部分の目的変数 Y と、その目的変数に影響を与える各説明変数 X とその説明変数ごとの回帰係数 β で構成された線形モデルである。本研究では、目的変数を株価の調整後終値とし説明変数を各感情にした。この調整後終値とは、株式分割で所持している実質的な株の価値に変化がなくとも一株あたりの株価が大きく変動することがあり、そのような株式分割を考慮した値段である。また、時系列ごとのツイート収集数に偏りがあり、それをなくすために式(1)のようにツイート数で正規化したものを感情軸x_i (i=1,2…10)とする。

$$\text{感情指数}x_i = \frac{\text{各時刻に於ける感情軸 i の感情語の数}}{\text{各時刻に於けるツイート数}} \quad (i=1,2\cdots10) \qquad (1)$$

ここでx_1は喜、x_2は怒、x_3は哀、x_4は怖、x_5は好、x_6は恥、x_7は昂、x_8は安、x_9は厭、x_{10}は驚のようにそれぞれの感情軸に於ける感情指数値を表す。

重回帰分析を行うにあたり各時刻に於けるツイートから算出した10軸の感情指数値を説明変数とした。感情指数値に前週の株価の平均値x11を加えた重回帰式を式(2)に示す。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} \quad (2)$$

前段階として10軸の感情を説明変数に利用し予測を行ったが、ツイッターの感情だけでは説明できない部分も多く結果が不安定だった。そこで一つ前の時刻を一つの説明変数として組み込むことで予測精度を安定化させた。この結果をもとに特定の学習データ N 回の範囲で重回帰分析を行い得られた回帰係数を基に次の時刻の感情指数値を代入することで目的変数の株価を予測した。

3．4　予測精度の評価指標

予測精度を評価する計算式を以下に示す。

$$m^2 = \frac{1}{M-1}\sum_{i=1}^{M}\left(\frac{\tilde{y}_i - y_i}{y_i}\right)^2 \quad (3)$$

ここでは、y_iは株価の実測値で、$\tilde{y}_i$は提案手法による予測値である。この実測値と予測値の差を計算し相対誤差、つまり変化の割合を求めた。相対誤差に正と負の値が出てくるので2乗し、各値の平均を取ることで評価指標を作成した。図2は N 個のデータから次の時刻 N+1 の株価予測に用いられるデータ範囲を示す簡略図である。

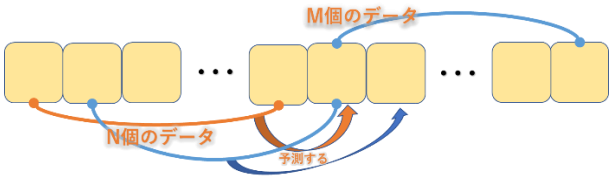

図2　予測に用いるデータの簡略図

3．5　ステップワイズ法を応用した感情軸の低減

先行研究で述べた Bollen 氏らの研究から特定の感情が株式に影響を与えているという結果がでているように、感情軸の数を減らすことで予測結果に優位な影響を与えることが分かった。そこで独自の評価指標とステップワイズ法を応用し説明変数の取捨選択をした。ここでは変数減少法を採用する。10軸から説明変数を一つずつ減らし、その結果からm^2の値が最も大きかったものを減らす。この手順を繰り返し、徐々に軸を減らした。その結果を時刻の範囲ごとに指定し最小となるm^2の値を求めた。実験の結果、優位な結果が確認された感情の組み合わせは、喜の軸と哀の軸を削除した感情軸であった。この残った x_2(怒)、x_4(怖)、x_5(好)、x_6(恥)、x_7(昂)、x_8(安)、x_9(厭)、x_{10}(驚)を説明変数として用いたものだった。

$$y = \beta_0 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + b_{11}\beta x_{11} \quad (4)$$

4　検証実験と考察

4．1　感情辞書の拡張

辞書拡張によって登録された感情表現は頻出度が高く多様な文章から感情を抽出できるようになった。紙面の都合上、全ての感情表現を載せることはできないので一部を抜粋して以下に記載する。

喜:
＼(^o^)／,面白い,わろた,ハッピー,グッチョブ,イイネッ,(笑),(≧▽≦),だるい,♪,♡

好:
イケメン、かわゆい、いいね、これはいい、愛してる、おめでとう、懐かしい

4．2　株価の予測実験

今回の実験では2014年時の「モンスト」のキーワードを含むツイートから算出した感情指数値、目的変数を Mixi に於ける2014年時の一週間ごとの調整後終値とした。先行研究の踏まえ、ツイッターのクラスタ層が関心を持ちやすく企業の株価にも少なからず影響がある「モンスト」というトソーシャルゲームを選んだ。尚、「モンスト」とは Mixi が 2013 年 10 月に提供を開始したソーシャルゲームのタイトル、「モンスターストライク」の略称である。このキーワードを選んだ理由として以下が挙げられる。一つに、企業名より商品名の方がツイートに感情が含まれる可能性が高いことや、モンスターストライクというソーシャルゲームが爆発的にヒットし当時の収益の90%以上がモンストによる収益だったこと、企業名より商品名の方が大量にツイート情報を取得できた点などの理由によりこのキーワードを選択した。その上で評価指標 m^2の値が最も低くなる範囲の回帰係数と翌週の説明変数を用いその範囲内の株価の予測値を求め、予測精度を評価した。また、予測値を介して感情がどのように影響を与えるかを検証する為に予測時の一つ前の週の株価の調整終値を説明変数として単回帰分析で予測を行った。尚、単回帰分析と低減感情軸で実施した場合は37週の時、評価指標 m^2の値が最小値となり、重回帰分析の場合は38週が最小値になった。以下に実験結果を記載する。

4．3　実験結果

表1にそれぞれの回帰分析で得られた決定係数の平均値を示す。

表1　決定係数の平均値の計算結果

単回帰分析	重回帰分析	低減感情軸
0.960753	0.967820	0.966615

表1より、どれも高い値だが10軸全てを利用した重回帰分析が最も学習データに対してフィットした結果になった。

そして、表2にその株価の調節終値と各手法による株価の予測値を示す。

表2 各手法の予測結果

週番号	調節終値	単回帰分析	重回帰分析	低減感情軸
39	5396.0	5629.67	5557.20	5560.32
40	5418.0	5518.71	5478.75	5491.06
41	5467.5	5522.77	5509.47	5588.15
42	5886.0	5565.16	5524.20	5576.95
43	6006.0	5974.70	5925.15	5934.60
44	6400.0	6114.41	6303.58	6336.89
45	5868.0	6504.58	6492.56	6493.29
46	5304.0	6000.21	5923.02	5924.95
47	5425.0	5394.71	5391.10	5394.23
48	5342.0	5477.05	5449.41	5440.23
49	5242.0	5399.31	5350.41	5341.10

表2より単回帰分析、重回帰分析共に株価が上昇したときは比較的予測ができている。実測値が両方上昇し、予測値も同様に上昇した確率は、71.43%となった。逆に下降した場合は、40.0%という結果になった。これは株価が上昇をしているときは、感情や口コミ、評判に左右されている場合が多い。逆に上昇し続けた株価が突然下落した場合、会社の不祥事が発覚や大多数の株主の予測に反して適正開示で売り上げが芳しくなかったことが開示されるなどセンチメント分析だけでは、説明できない部分がある為だと考えた。例えば会社が不祥事を発表してから株価が下落するなどの場合は、いち早く情報をキャッチし投資を行う投資家に対して少なからず問題に社会感情が後追いになってしまう為、SNS の情報だけでは問題が認知された瞬間の予測が難しいと考える。

表3に3．4に定義した予測精度の評価指標の計算結果を示す。

表3 評価指標 m^2の計算結果

単回帰分析	重回帰分析	低減感情軸
0.004485	0.004402	0.004050

この結果から分かるように感情指数値を加味した重回帰分析の方がより誤差が小さく、精度が高い結果になった。

5 まとめ

本研究では、Twitter 上の感情の動きに着目し、社会感情との関連が強い株価のデータを利用し線形回帰の手法を用いて特定銘柄の予測を試みた。従来の研究では、ネガティブかポジティブの2軸で行われる研究が殆どであった。しかし、人間の感情はネガティブかポジティブの2値に割り切れるものではなく、多様な感情に分類することで多様な情報を得ることができ、線形回帰分析に於いても2値の情報だけでなく複数の情報から分析することでより高い精度の予測ができた。そこでネガティブかポジティブの2軸でなく10軸に分類された感情表現辞典を参考に作成、拡張し現代に即した感情表現を追加し感情辞書を構築した。その上でツイートのテキストデータから感情情報を数値化し出力するシステムを作成した。更にツイートの感情指数値を求め、そこから重回帰分析を用いて株価の予測をする手法を提案した。一つ前の時刻の株価と現在時刻のツイート内容にある感情情報を加味することで、従来の上昇、下降の予測モデルに囚われることなく、より精確な数値の予測を実現した。

参考文献

[1] Johan Bollen,Huina Mao,Xiao-Jun Zeng：Twitter mood predicts the stock market. arXiv:1010.3003v1 [cs.CE]（14 Oct 2010）

[2] 佐藤謙太,高知宏,黒岩丈介,白井治彦 :ネガポジ解析 による Web デ ー タと株価変動の相関関係評価, 福井大学 大学院工学研究科 研究報告 第 63 巻（2015 年 3 月）

[3]五島圭一,高橋大志：株式価格情報を用いた金融極性辞書の作成，自然言語処理 Vol.24 No.4（2017年9月）

[4] Twitter Developer Platform — Twitter Developers : https://developer.twitter.com/en/docs（2018/1/17 アクセス）

[5] Apache CouchDB: http://couchdb.apache.org/（2017/12/22 アクセス）

[6]MeCab-日本語形態素解析システム https://www.mlab.im.dendai.ac.jp/~yamada/ir/MorphologicalAnalyzer/MeCab.html（2017/12/22 アクセス）

[7] 中村明:感情表現辞典,東京堂出版,1993 年 5 月