経済情報に関心をもつSNSユーザの投稿内容に基づく株価騰落予測モデルの構築

Predicting Stock Price Movement based on SNS Articles Posted by Users Concerned about Economic Information

佐藤大吾[1]　大原剛三[1*]　豊田哲也[1]
Daigo Sato[1], Kouzou Ohara[1], Tetsuya Toyota[1]

[1] 青山学院大学理工学部
[1] College of Science and Engineering, Aoyama Gakuin University

Abstract: In this paper, we propose a method of constructing a predictive model of stock price movement that utilizes articles posted to Twitter by users who are likely to be concerned about economic information. There exist some researches that attempt to predict a financial index based on sentiment polarities of tweets. But, they do not take into account who posted those tweets. It is expected that tweets posted by users who are concerned about economic topics can be better indicators to predict a financial index than the ones posted by users who are not. Thus, we extract economic keywords from news articles in the category of economics in advance, and select users who tweet articles including those keywords more frequently among from those who were randomly chosen, and then, utilize the sentiment polarities of their tweets to build a predictive model. In the experiment, we will evaluate the model induced by the proposed method by empirically comparing it with the one based on past stock prices and the one based on them and tweets posted by users randomly chosen to show the effectiveness of our approach.

1　はじめに

近年，株価やその他の金融指標の予測に，インターネット上のテキスト情報を利用する研究が盛んになっている [1]．それらは，ニュース記事や経済レポートなどの文章から特徴を抽出するもの [2] と，Twitter などのソーシャルメディアに投稿された一般ユーザが作成した文章を利用するもの [3, 4, 5, 6] に大別できる．特に後者のアプローチでは，API などを通して大量に取得できる文章から肯定的か否定的かなどの市場の感情傾向を推測し，その傾向を金融指標として利用することで株式指標などを予測する．Giudice は，Apple，Google，Facebook，Microsoft の 4 つの企業の cashtag[1] を含むツイートを感情分析し，株価変動は肯定的な感情をもつツイートに影響することを示している [4]．また，Bollen らは，心的ツイート（“I feel～”，“I’m～” などの書き手の感情が入っているツイート）に絞って感情分析を行い，どのような心理状態がダウ平均株価と相関が強いかを調査している [5]．この研究では，心理学によく用いられる POMS（Profile of Mood States）検定に基づいて極性分析を行い，2～5 日後のダウ平均株価の騰落を高い精度で予測することに成功している．一方，Zhang らは，“hope”，“Fear”，“Happy” などが含まれるツイートに着目し，それらツイートの各単語含有割合，リツイート数，発話ユーザのフォロワー数と Daw，NASDAQ，S&P 500 などの株価指標間の相関を算出し，ツイートから投資家心理を予測可能であることを示している [6]．

このように，既存のアプローチでは感情表現を含むツイートに限定した分析を行っている一方，対象とするツイートを投稿したユーザに関しては，特別な注意は払われていない．ツイートに対する感情極性値を経済指標として用いる場合，無作為に感情表現を含むツイートを抽出するよりも，経済動向に敏感に反応するようなユーザによるツイートを積極的に利用する方が，市場の雰囲気をよりよく反映できることが期待できる．このような考えの下，本研究では，感情極性の分析対象となるツイートを投稿するユーザを無作為に選定するのではなく，経済関連ニュースに対する興味度によって絞り込み，選定されたユーザによるツイートを利用して株価騰落予測の精度を向上させることを試みる．本研究では，オンラインニュース上の経済関連記事から特徴語を抽出し，特定の期間中におけるそれらの特徴語を含むツイートの投稿頻度に基づき，Twitter ユーザ

*連絡先：青山学院大学理工学部情報テクノロジー学科
〒 252-5258 相模原市中央区淵野辺 5-10-1
E-mail: ohara@it.aoyama.ac.jp

[1]twitter 上で$GOOGL のようなティッカーシンボルをクリックし，株式や企業に関する検索結果を確認できるもの．

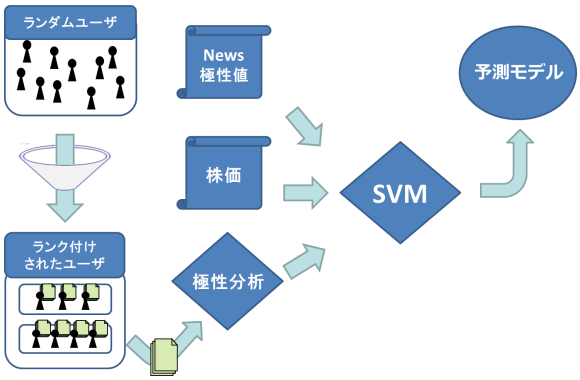

図 1: 提案手法の概要

の経済関連ニュースに対する興味度を推定する．そして，興味度が高いと推定されるユーザのツイートに対する感情極性値を素性として利用し，株価騰落予測モデルを構築する．評価実験では，日経平均株価の終値を対象に，株価のみを用いて構築した予測モデル，および無作為に選定したユーザのツイートも利用して構築した予測モデルと提案手法により構築した予測モデルの正答率を比較し，ユーザを限定することの有効性を評価する．

2 提案手法

本研究で提案する手法の概要を図 1 に示す．まず，Twitter Streaming API[2] を用いて，Twitter ユーザをランダムに収集する．ただし，同じツイートを複数投稿しているユーザはスパムユーザとして排除する．次に，分析期間中の経済カテゴリのニュース記事から特徴語を抽出し，その特徴語を含むツイートの投稿回数でそれらのユーザをランキングする．そして，直前の株価に加え，ランキング上位のユーザのツイート，および経済ニュース記事に対する極性分析結果を素性とし，前日の株価に対する「上昇」「下降」をクラスラベルとして予測するモデルを SVM を用いて構築する．

2.1 Twitter ユーザのランキング

提案手法では，オンラインニュースの経済関連記事から抽出した特徴語を使用し，ある一定期間（以下，ユーザ選定期間）中に特徴語を含むツイートを投稿した Twitter ユーザを分析対象として利用する．特徴語は日ごとに当日のニュース記事から抽出する．具体的には，ユーザ選定期間中の各日の記事集合から名詞を抽出し，1 日分の記事集合を 1 つのドキュメントとして計算した TF-IDF 値の上位 m 件をその日の特徴語として用いる．本稿における実験では，$m = 20$ とした．ニュース記事からの名詞の抽出には，形態素解析器 MeCab [3] を利用し，その辞書としては mecab-ipadic-NEologd [4] を用いる．この辞書は，標準のシステム辞書では正しく分割できない固有表現などの語の表記・フリガナの組を約 208 万組収録しており，これらの語は Web 上の言語資源を活用し，定期的に新しい固有表現が収録される．そのため，常に新しい語が出現するニュースやツイートの形態素解析に適した辞書といえる．

そして，上記のように分析対象として選定した Twitter ユーザを，経済関連ニュースに対する興味度に基づきランキングする．ここでは，ユーザ選定期間中における特徴語を含むツイートの投稿回数をそのユーザの経済関連ニュースに対する興味度と定義し，その投稿回数が多いほど経済関連ニュースにより関心を持っているユーザであると仮定する．なお，複数のユーザが同じ投稿回数となる場合は，それらのユーザは同順位とする．

2.2 極性分析

提案手法では，分析対象ユーザのツイート，およびオンラインニュース記事から算出した感情極性値を株価騰落予測モデル構築時の素性として利用する．具体的には，評価表現語の語幹とその極性が対になって登録されている乾らの日本語評価極性辞書（名詞編・用言編）[7, 8] を用い，対象文書内に評価表現語が存在した場合，その極性に応じて肯定であれば 1 を，否定であれば –1 を加算し，その合計値を出現した評価表現語の総数で割った値をその文書の極性値とした．すなわち，1，–1 を肯定，否定それぞれの評価表現語の極性値とし，その平均を文書全体の極性値としているため，文書全体の極性値はその文書量にかかわらず [–1, 1] の範囲の値とする．なお，文書内の表現と評価極性辞書内の表現を比較するために，文書内の単語は MeCab を用いて語幹に変換して利用する．

一方，極性分析の際には，単語の極性を反転させる表現に注意する必要がある．ここで，極性の反転には，肯定から否定と否定から肯定への 2 方向が考えられることに注意する．たとえば，「～が減少する」という表現は，「利益」という評価極性辞書内で肯定に分類される評価表現語と組み合わさると「利益が減少する」となり，その評価を否定に反転する．逆に，「苦情」という評価極性辞書内では否定に分類される評価表現と組み合わさると「苦情が減少する」という肯定の表現とな

[2] https://dev.twitter.com/

[3] http://taku910.github.io/mecab/

[4] https://github.com/neologd/mecab-ipadic-neologd

る．以下，このような極性を反転させる表現を極性反転表現と呼ぶ．このような極性反転表現は多様なものが存在し，機械的にそれらを列挙することは必ずしも容易ではない．そのため，ここでは，本研究が対象とする株価予測は利益を上げることが目的であり，いかに損益を出さないかが重要となることに着目する．そのような観点からは，肯定的な雰囲気を否定的に見積もるよりも，否定的な雰囲気を肯定的に見積もる方がリスクが高いと言える．そのため，本研究では，否定的な表現をより正確に抽出するために，肯定的な評価表現を否定的な表現に変える極性反転表現を事前に収集し，極性分析ではその極性反転表現リストを利用する．

具体的な極性反転表現の収集法としては，まず収集した経済関連のニュース記事から否定に分類される評価表現を機械的に抽出し，その中から経済系の単語のうち肯定に分類される単語に係りやすいものを人手で選別した．一方，極性分析においては，係り受け解析器を利用し，文節中の肯定的単語に係る極性反転表現が存在した場合，その組合せ全体を 1 つの否定的表現として極性値を計算する．提案手法では，係り受け解析器として CaboCha[5] を使用し，そのシステム辞書には MeCab と同様に mecab-ipadic-NEologd を用いた．

2.3 予測モデル構築のための素性の生成

今回，提案手法では騰落予測モデルの生成に SVM を利用する．そのモデル学習時に利用する素性としては，従来の研究では予測対象日の前日，もしくは直近 k 日の株価，ニュース記事やツイートの感情極性値が用いられることが多い．一方，Bollen らは，株価とツイートの感情極性値の時系列変化の相関を両者の z スコアを用いて評価し，$k = 3$ とした場合の z スコアを用いたときの両者の高い相関を示している [5]．この知見と，予備実験の結果から，提案手法では株価，およびツイートとニュース記事の感情極性値をそのままモデル学習における素性として利用するのではなく，それらのデータ値に対する過去 k 日間の z スコアを利用する．具体的には，予測対象日の前日を t，計算に利用するデータの日数を k とした場合，次式により得られる z スコア $\mathbb{Z}_t$ を素性として用いる．

$$\mathbb{Z}_t = \frac{X_t - \bar{X}_{t,k}}{\sigma_{t,k}} \tag{1}$$

$$\sigma_{t,k} = \sqrt{\frac{1}{k}\sum_{i=0}^{k-1}\{X_{t-i} - \bar{X}_{t,k}\}^2} \tag{2}$$

ここで，$\bar{X}_{t,k}$ は t から $t-k-1$ までの予測日前 k 日間のデータ値 $X_t, \cdots, X_{t-k-1}$ の平均を表す．たとえば，$k = 3$ とした場合，12 月 24 日の株価騰落予測に使用する z スコアは，12 月 21 日から 12 月 23 日までの 3 日間のデータ値を用いて式（1）に基づき計算する．なお，ツイートの極性値に対する z スコアの計算では，X_t は興味度上位 n 位までのユーザが日付 t に投稿したツイートの極性値の平均とする．ニュース記事の極性値に関しては，日付 t の記事に対する極性値の平均を X_t として用いる．ただし，これらの平均は肯定・否定それぞれの極性ごとに計算し，感情極性値が 0 となるツイート/ニュース記事は対象としない．

一方，Giudice らは，極性が肯定であるツイートの極性値がその数日後の株価と強い相関をもつことを示している [4]．この結果に基づき，提案手法では，経済関連ニュースに対する興味度上位 n 位までのユーザが k 日間に投稿した極性が肯定であるツイート（PT: Positive Tweet），およびその期間中の極性が肯定である経済関連ニュース記事（PN: Positive News）それぞれに対する極性値の z スコア $\mathbb{Z}_{PT}$，$\mathbb{Z}_{PN}$，およびその期間中の株価（SP: Stock Price）に対する z スコア $\mathbb{Z}_{SP}$ を予測モデル学習時の素性として用いる．

3 評価実験

3.1 分析対象データ

本実験では，2015 年 11 月 1 日から 1 月 31 日までの 3 か月間における 63,015,587 件のツイートを使用した．まず，Twitter が提供するストリーミング API を用いて 50,956 件のユーザをランダムに取得した．そして，それらのユーザが 10 月に投稿したツイートと，10 月の経済関連ニュース記事から抽出した 521 個の特徴語（TF-IDF 値上位 20 まで）を用い，それらの特徴語を含むニュースツイートの投稿頻度に基づきユーザをランク付けした．経済関連ニュース記事は，日毎の記事数が比較的豊富である産経ニュース [6] の経済カテゴリから収集し，その数は 10 月 1 日から 1 月 31 日の間で 4,215 記事となった．ニュースツイートの投稿頻度に対するユーザ数の分布を表 1 に示す．

一方，予測対象は日経平均株価の終値とし，日本経済新聞社が運営する日経経済プロフィル [7] から 2015 年 11 月から 2016 年 1 月までの平日 59 日分の終値を取得して利用した．図 2 に当該期間の日経平均株価終値の推移を示す．図 2 に示すように，11 月，12 月は比較的穏やかな騰落の傾向がみられる．一方，1 月は原油安や，中国株の暴落などに影響を受け，急激な株価の下落がみられた．今回は株価の予測ではなく，株価の騰落を予測対象とするため，前日の株価に対する上昇・

[5] https://taku910.github.io/cabocha/

[6] http://www.sankei.com/

[7] http://indexes.nikkei.co.jp/nkave/index/profile

表 1: ユーザのニュース興味度（経済関係ニュース記事の特徴語発話頻度）の分布

ニュースツイート投稿頻度（回）	ユーザ数（人）	ニュースツイート投稿頻度（回）	ユーザ数（人）
171	1	80～89	3
151	2	70～79	7
145	1	60～69	16
129	2	50～59	22
123	1	40～49	43
102	1	30 39	116
93	1	20～29	454
92	1	0	31,492

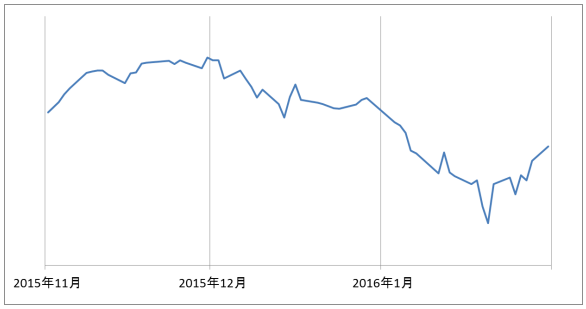

図 2: 日経平均株価終値

表 2: 極性反転表現リスト（一部抜粋）

極性反転単語			
負担	懸念	下落	悪化
低下	低迷	落ち込む	難航
後退	転落	停滞	高騰
衰える	迷走	暴落	頓挫
降格	隠蔽	横領	消失
減価	超過	捏造	窮迫

下降に対してそれぞれ 1 と 0 のラベルを付与して利用した．

また，収集したニュース記事からは，極性分析に用いる極性反転表現リストを構築した．具体的には，まず 4,215 件のニュース記事に出現した単語のうち，評価極性辞書に否定の極性として登録されている名詞・動詞 2,012 個を抽出し，その中から，肯定の極性をもつ表現に係ることで否定の意味を成すような単語を人手で選別し，最終的に全 89 個の単語から成る極性反転リストを構築した．表 2 に極性反転表現リストの一部を示す．

3.2 実験設定

株価騰落予測モデルを学習する SVM の実装には，Python の機械学習ライブラリである sckit.learn[8] を用いた．SVM の正則化係数は予備実験に基づき 100 を設定し，カーネル関数には RBF（Radial Basis Function）カーネルを用いた．学習のための訓練期間は 1ヵ月とし，その直後の 1ヵ月を予測期間とした．具体的には，以下の 2 期間を対象とした評価実験を行った．

期間 1： 訓練期間 11 月，予測期間 12 月

期間 2： 訓練期間 12 月，予測期間 1 月

[8]http://scikit-learn.org/stable/index.html

評価指標としては，予測日の日経平均株価終値の前日終値に対する「上昇」「下降」2 値の正答率を用いた．また，経済関連ニュースに対する興味度上位 n 位までのユーザが予測日前 k 日間に投稿した Positive Tweet，その期間中の Positive News，およびその期間中の日経平均株価それぞれに対する z スコア $\mathbb{Z}_{PT}$，$\mathbb{Z}_{PN}$，および $\mathbb{Z}_{SP}$ を予測モデル学習時の素性として用いる提案手法に対して，$\mathbb{Z}_{SP}$ のみを用いて予測モデルを学習する手法（ベースライン 1）と $\mathbb{Z}_{SP}$ と興味度に関係なく全ユーザが当該 k 日間中に投稿した Positive Tweet に対する z スコアを利用して予測モデルを学習する手法（ベースライン 2）を比較対象として用いた．z スコアを計算する日数 k に関しては，Bollen らの研究と同様に，$k = 3, 4, 5, 6, 7$ の 5 つの値を採用した．

また，極性分析に用いる極性反転表現リストは経済関連ニュースから抽出したものであるため，ツイート本文にそのまま利用することが妥当であるとは限らない．そのため，提案手法に関しては，極性反転表現リストの有用性を評価するために，極性反転表現リストを利用しない場合（設定 1），経済関連ニュース記事とツイート双方に利用する場合（設定 2），ツイートのみに利用する場合（設定 3），および経済関連ニュース記事のみに利用する場合（設定 4）のそれぞれの場合における正答率を求めて比較した．

表 3: $k = 3$ における各手法による騰落正答率（%）

予測対象月	12 月	1 月
ベースライン 1	61.9	47.4
ベースライン 2	47.6	52.6
提案手法（設定 1）	78.9	66.7
提案手法（設定 2）	76.2	63.2
提案手法（設定 3）	61.1	42.1
提案手法（設定 4）	76.2	63.2

3.3 実験結果

ベースライン 1，2 の双方が 12 月，1 月の予測対象月いずれにおいても $k = 3$ 日間の z スコアで正答率が最高となったため，$k = 3$ にける各手法の結果を表 3 に示す．なお，提案手法の結果はいずれも $n = 10$ の場合のものである．これらの結果から，12 月，1 月のいずれにおいても設定 1（極性反転表現リスト未使用）の下での提案手法の正答率が最も高いことがわかる．ベースライン手法に関しては，12 月と 1 月で 2 つの手法の優劣が逆転している．これは，株価の z スコアのみを用いるベースライン 1 は，値動きの大きくない期間ではランダムなユーザの Positive Tweet のみを用いる場合より正答率が高くなるものの，値動きが激しくなると，付加情報なしではその動きに追従することが難しくなるためであると考えられる．

次に，提案手法における極性反転表現リストの利用に関しては，残念ながら今回の実験結果では，利用した場合に正答率を下げてしまう結果となっている．特に，ツイートにのみ適用する設定 3 では，正答率の大きな減少が見て取れる．これは，ニュース記事から抽出した極性反転表現の係り受けをツイート特有の崩れた表現の中で正しく抽出することが困難であったためと考えられる．一方，ニュース記事に適用した場合も正答率を下げていることから，今後，実験結果のより詳細な検証と，極性反転表現リストの洗練が必要であると言える．

以上の結果から，極性反転表現リストの有用性は確認できなかったものの，経済関連ニュースに一定の興味をもつであろうと推測されるユーザに限定してそのツイートの極性情報を用いることは，株価騰落予測に一定の効果があることが確認できたと言える．

4 まとめ

本研究では，Twitter に投稿されたツイートに対する感情極性値を利用して株価騰落予測モデルを構築する際に，対象ツイートを無作為に選定するのではなく，経済関連ニュースに興味をもっていると推測されるユーザが投稿したものに限定する手法を提案した．提案手法では，経済関連ニュースから事前に特徴語を抽出し，その特徴語を含むツイートの投稿頻度に基づきユーザの経済関連ニュースに対する興味度を定義した．日経平均株価の騰落予測を対象とした評価実験では，過去の株価のみ，もしくは無作為に抽出したユーザによるツイート内容を用いるよりも，提案手法により経済関連ニュースに対する興味度が高いと推定したユーザのツイートを用いる方が，株価の騰落予測に対する正答率が向上することを確認した．一方，感情分析における否定表現の利用はうまく機能しなかったため，今後，さらなる洗練と検証が必要である．また，今回の実験では，経済関連ニュースに対する興味度上位 10 位までのユーザしか利用しておらず，評価期間も 2ヶ月のみであったため，今後，より広範な評価実験を通して，提案アプローチの有効性を検証する必要がある．

参考文献

[1] 和泉 潔，松井 藤五郎 : 金融テキストマイニング研究の紹介，情報処理，Vol.53, No.9, pp.932-937 (2012).

[2] 和泉 潔，後藤 卓，松井 藤五郎 : 経済テキスト情報を用いた長期的な市場動向推定，情報処理，Vol.52, No.12, pp.3309-3315 (2011).

[3] Ruiz, J.E., Hristidis, V., Castillo, C., Gionis, A., and Jaimes, A. : Correlating Financial Time Series with Micro-Blogging Activity，Proceedings of the fifth ACM International Conference on Web Search and Data Mining，pp.513-522 (2012).

[4] Lo Giudice, M. : The predictive character of the social sentiment on the stock market:Twitter and the stock trend，http://essay.utwente.nl/67238/ (2015).

[5] Bollen, J., Mao, H., and Zeng, X-J. : Twitter mood predicts the stock market，Journal of Computational Science，Vol.2, No.1, pp.1-8 (2011).

[6] Zhang, X., Fuehres, H., and Gloor, P : Predicting Stock Market Indicators Through Twitter “I hope it is not as bad as I fear”，Procedia-Social and Behavioral Sciences，Vol.26, pp.55-62 (2011).

[7] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一 : 意見抽出のための評価表現の収集，自然言語処理，Vol.12, No.3, pp.203-222(2005).

[8] 東山昌彦，乾健太郎，松本裕治 : 述語の選択選好性に着目した名詞評価極性の獲得，言語処理学会第 14 回年次大会論文集，pp.584-587(2008).