大規模ニュース記事からの極性付き評価表現の抽出と株価収益率の予測

前川　浩基，中原　孝信，岡田　克彦，羽室　行信

この 10 年に，インターネットを通じて取得可能なニュース記事や SNS データなどの定性的なデータを株価予測に応用しようとする動きが，ファイナンスとコンピュータサイエンスの両分野で同時に活発化してきている．本研究では，すでに公開情報となった経済に関する大規模ニュース記事テキストから，日経 225 先物価格の近未来の上下変動を予測するモデルを Naïve Bayes 法によって構築した．時系列で一定の window サイズを設定し，その期間をずらしながら out-of-sample による予測実験を行った結果，正答率 52.3%，年平均収益率 11.3%，Sharpe 比 0.689 を達成し，本モデルの有効性が示された．

キーワード：テキストマイニング，日経平均株価，Naïve Bayes

1. はじめに

株価動向はランダム・ウォークであると考えられている．効率的市場においては，株価には企業に関するすべての公開情報が瞬時に反映され，新たな情報の到来によってのみ変動する．したがって，新情報の到来はまったく予期できないため，株価変動も予測不可能だと考えられてきた [6]．しかしながら，プロ・アマを問わず株式市場に関わる多くの投資家が過去の変動パターンに基づいたテクニカル分析に依拠して売買判断し，あるいは，証券アナリストや経営者の発するファンダメンタル情報を解析して投資判断をしている．よって，公開情報が瞬時に株価に反映されると考えるのは非現実的であろう．

本稿の目的は，すでに公開情報となったニュース記事から，近未来の株価（日経 225 先物）を予測する確率モデルを構築し，その有効性を検証することである．

ニュースと株価変動に関する研究はいくつか存在するが，これまでの主たるテーマはニュースに含まれる市場センチメントの株価への影響であった．本稿では実験者が記事のセンチメントを測定し株価予測を行うのではなく，非常にシンプルな bag-of-words アプローチ，すなわち，ニュース記事に含まれる多様な言葉の出現頻度から株価の上下動を予測するモデルを Naïve Bayes 法を用いて構築する．

事前に配信されたニュース記事から株価が予想ができれば，それは既出情報に投資家が反応して株価が形成されている証拠であり，株価のランダム・ウォーク仮説は棄却される．

本稿の構成は以下のとおりである．第 2 章ではテキストマイニングによる株価予測に関する先行研究を概観し，第 3 章では利用するデータについて説明する．第 4 章ではモデル構築の方法論について論じ，第 5 章で実験結果を示し，その考察を行う．

2. テキストマイニングと株価予測

ニュースや掲示板などのテキスト情報と株式市場との関係については，2000 年以降いくつもの研究がなされている．表 1 に主な研究の一覧をその発表順に示す．

ファイナンス分野においては，Antweiler らが Yahoo! Finance の掲示板に投稿されたメッセージを解析し，投稿内容には株価の予測可能性はないものの，投稿数の増加はその後の株価変動率の上昇を予想することを報告している [1]．Tetlock は，Wall Street Journal に日々掲載される市場観測コラム記事を解析し，そのなかに含まれる悲観的な語句の出現度合いが，翌日以降の株価指数リターンに影響を与えていることを明らかにした [11]．また，Tetlock らは個別企業に関する記事を解析し，それらに含まれる悲観語の数が将来の当該企業の株価リターン，業績を予想することを明らかにした [13]．さらに，投資家行動に関する研究において，テキストマイニングの技術を使ってニュースの類似性を計算し，とりわけ個人投資家が新しいニュースとそうでないニュースについて区別できずに行動し

まえがわ　ひろき
（株）Magne–Max Capital Management
なかはら　たかのぶ
関西大学 データマイニング応用研究センター
おかだ　かつひこ，はむろ　ゆきのぶ
関西学院大学大学院 経営戦略研究科
〒 662–8501 兵庫県西宮市上ヶ原一番町 1–155

ていることを報告している [12].

一方，コンピュータサイエンス分野でも，過去のニュース記事を用いた株価リターンの予測が試みられている．Gidófalvi はニュース記事に出現する単語群を特徴ベクトルとする Naïve Bayes モデルを用い，ニュース記事の公表から 20 分後の株価リターンが予測できることを示した [4]．また Mittermayer は，企業のプレスリリースから 60 分後の株価リターンにはランダム・ウォークとは異なる歪みがあり，SVM を用いたモデルによってその方向性が予測できるとした [7]．Fung らは SVM を用いた株価リターンの予測を行うにあたって，ニュースが公表された時点での株価トレンドを教師情報として用いる手法を提案している [3]．Schumaker らも SVM を用いた株価の予測を行っており，予測モデルに投入する特徴ベクトルとして，ニュース記事に出現する名詞句もしくは固有名詞を用いることで予測精度が高まることを示した [9, 10]．Bollen らは，twitter から収集した大量の tweet を用いて世の中全体のムードを測定し，それが株価指数リターンと強い相関を持つことを明らかにした [2].

本研究も，これらの一連の研究の延長線上に位置づけられる．本研究の新規性は，11 年間にわたる 40 万件以上の大規模な経済関連ニュース記事を解析対象としていることにある．またテキストの解析においても，多くの bag-of-words アプローチに見られるように，単に単語の出現をみるのではなく，評価表現辞書を独自に作成することで，テキストの意味をより正確にとらえた解析となっている．

表 1　テキスト解析と株価変動に関する先行研究一覧

文献	年	ソース	期間	記事数	目的変数	モデル
[4]	2001	Yahoo! Finance	1999/11 –2000/2	5,500	stock return	Naïve Bayes
[7]	2004	PRNews Wire	2002	6,602	stock return	SVM
[1]*	2004	Yahoo! Finance, Raging Bull	2000/1 – 2000/12	1.5mil.	stock return, turn-over	Naïve Bayes
[3]	2005	Reuters	2003/1 –2003/6	N/A	stock return	SVM
[9]	2006	Yahoo! Finance	2005/10 – 2005/11	9,211	stock price	SVM
[11]*	2007	WSJ	1984/1 –1999/9	3,709	DJIA return	PCA VAR
[13]*	2008	DJNS, WSJ	1980 –2004	350,000	stock return	OLS
[10]	2009	Yahoo! Finance	2005/10 – 2005/11	9,211	stock price	SVM
[2]	2011	twitter	2008/2 – 2008/12	9.8mil.	DJIA return	VAR
[12]*	2011	DJNW	1996/11 – 2008/10	30mil.	stock return, turn-over	OLS

*はファイナンス分野，それ以外はコンピュータサイエンス分野

表 2　年別ニュース記事の件数

Year	件数	Year	件数	Year	件数
2000	28,038	2004	41,594	2008	64,961
2001	35,917	2005	54,577	2009	36,131
2002	33,624	2006	44,391	2010	22,974
2003	28,440	2007	39,954	2011	18,285

3. データ

3.1 ニュース記事

本研究では，資金運用に携わるファンドマネージャーや証券アナリストがほぼ例外なく使用している Bloomberg 社の日本語のニュース記事を利用する．英語の記事も配信されているが，今回は対象外である．また，同社が提供する記事の多くは経済に関するものではあるが，スポーツや事件のような経済とは関係の薄い内容も配信される．[2] の研究に見られるように，理由がわからなくとも，世の中の一般的なムードと株価の関係を明らかにするアプローチも有望ではあるが，Bloomberg ニュースが主に経済活動に携わる人間を対象としていることを鑑み，あえて，株価に直接影響を与えやすいと考えられる記事を対象とすることとした．

同社が配信する記事には，その記事が関係する上場企業の銘柄コードが人間の手によってタグづけされている．今回の実験では，これらタグづけされた記事のみを対象とした．対象期間は 2000 年 1 月 1 日から 2011 年 5 月 31 日までの 11 年強で，対象とした記事の件数は表 2 に示すとおりである．

本研究では，記事の内容によって日経 225 先物価格を日単位で予測しようとするものであるが，そこでは，記事が配信された「翌日」の始値で購入することを想定している．ゆえに，市場がオープンするまでに配信された記事内容は，取引への影響が想定される．そこで記事の日付変更時刻を 8:45 とすることにした．すなわち，午前 0:00 から 8:45 までに配信された記事の日付は，前日に配信されたものとして扱う．

3.2 日経 225 先物データ

株価データとしては大阪証券取引所に上場する日経 225 先物価格（以下，「株価」とも呼称する）を用いる．日経 225 先物とは，一般的に馴染みのある日経 225 平

均株価指数を原資産とし，ある将来時点（満期）において現在約束した価格で原資産を引き渡す契約である．日経 225 先物を買い持ち (Long) している主体は満期において株価指数を契約した価格で受け取り，売り持ち (Short) している主体は契約した価格で引き渡す．日経 225 平均株価指数とは，日本経済を代表する 225 銘柄の平均価格であるから，ある将来時点において株式市場全体の価格が上昇するだろうと予想する投資家は現時点で Long し，下落するだろうと予想する投資家は現時点で Short する．したがって，先物価格の売買は原資産の売買と実質的に同じである．ただ，日経 225 先物取引を通じて売買することで，原資産の 225 銘柄全部を取引するよりも流動性が高く，取引コストが安価である．また，先物価格と原資産価格の間では，価格裁定（どちらか一方が割安，割高であれば是正する投資家行動）が働くため，先物価格が株価指数に忠実に追随するという特徴を持つ．したがって，本稿で示された結果は見せかけなものではなく，実務的実現可能性が高いと言える．

4. 手法

本研究は，ニュース記事に出現する評価表現から日経 225 先物価格の上下変動の予測を試みる．そこで，以下では，極性付き評価表現辞書の構築手法，そして予測モデルとしての Naïve Bayes モデルについて解説する．また，モデルの評価指標として平均収益率とリスクを同時に評価する Sharpe 比の求め方についても言及する．

4.1 極性付き評価表現辞書の構築

極性付き評価表現とは，例えば，「回復する」「株価が上昇する」（好評表現），「下落する」「業績が落ち込む」（不評表現）など，事物に対する評価について好評と不評の軸を持った表現のことである．

評価表現の抽出方法はいくつか提案されている．例えば，[5] では，シソーラスに登録された語彙の類義語関係からネットワークを構成し，事前に定義した極性語（例えば，「良い」と「悪い」）との距離の近さによって語彙の極性を測定する．この方法は，言語学的見地から構築された辞書に基づいており，極性評価の妥当性は高いものの，一般的なシソーラスを用いるため，業界特有の表現についての評価が困難となる．例えば，株価に関する用語で「底固い」とは，株価が大幅に下落しない状態が続くという意味で，肯定的な極性を持つが，一般的なシソーラスを用いてはそのような判定は困難であろう．また，[14] では，巨大なコーパスを用いて，極性ごとの共起関係を見ることによって評価の極性を判定する方法が提案されている．例えば Web の全文書をコーパスとして用いるのであれば，Google を代表とする検索エンジンを用い，極性語と対象の言葉を and 条件で検索したときのヒット件数に基づいてそれらの言葉の類似度を定義できる．この方法は，コーパスさえ用意できれば簡単に実現できるという手軽さはある一方で，単純な共起関係に基づいているため，得られる類似度の妥当性に問題が残る．そこで本研究では，那須川・金山が開発した，周辺文脈の一貫性を利用した極性付き評価表現辞書の構築手法 [15] を用いることにした．この手法では，上記に示したいずれの問題もクリアすることが可能となる．

この手法は，好評と不評のラベルを伴った少数の評価表現を教師値として与え（種表現と呼ぶ），コーパスから評価表現を次々と取得していく半教師あり学習の手法を用いている．既知の極性付き評価表現が存在すると，その周辺文脈もその評価表現と同じ極性で一貫しているという仮定をおく．そして，それらの種表現が含まれる文章の周辺文脈を調べることで，次々と新たな評価表現を獲得していく．

この手法で対象とする評価表現は単純エントリと複合エントリに大別でき，単純エントリとは，「上昇する」「回復する」といった用言句からのみ構成される表現で，複合エントリとは「株価が上昇する」「景気が回復する」といった格助詞句と用言句のペアである．

評価表現辞書の構築アルゴリズム genLexicon の概略を図 1 に示す．getCandidates（手順 8）は，記事集合 $\mathcal{D}$ から評価表現辞書 L に含まれる評価表現を検索し，表現を含む周辺文脈の評価表現を候補表現 C として抽出する．そして evalCandidates（手順 9）において抽出した候補表現 C から以下に示す三つの条件を満たす評価表現を選択し，それらの表現 L' を新たに評価表現辞書に加える．

1. **最小サポート**(σ)：候補表現の出現度数が σ 以上である．
2. **最小極性割合**(ρ)：候補表現の好評もしくは不評文脈での出現割合が ρ 以上である．
3. **確信度**(α)：評価表現の平均誤用確率 π（評価表現を誤った極性として使う確率で，本研究では 0.01 に固定した）が一定で，記事集合 $\mathcal{D}$ における出現総数 N をベルヌーイ試行における試行回数と考えると，評価表現の誤用数 x は二項分布 $B(N,\pi)$ に従う．ここで候補表現の一方の極性の出現数 q が誤用とは言えないほど統計的に十

1: **procedure** genLexicon($\mathcal{D}, L, \sigma, \rho, \alpha$)
2: $\mathcal{D}$：記事集合
3: L：評価表現集合
4: σ：最小サポート
5: ρ：最小極性割合
6: α：確信度
7: **loop**
8: $C = \text{getCandidates}(\mathcal{D}, L)$
9: $L' = \text{evalCandidates}(C, \sigma, \rho, \alpha)$
10: **break if** $L' = \emptyset$
11: $L = L \cup L'$
12: **end**
13: **return** L

図 1　評価表現辞書獲得のメイン手順

分大きい，すなわち，$p(x < q) \geq \alpha$ を満たす．

以上の過程を，新たな評価表現 L' が獲得できなくなるまで繰り返す．ただし，本研究では手順 9 における候補表現の判定に人間の判断を加え，明らかに極性を持っていない表現については除外することにした．より詳細については原著 [15] を参照されたい．

また種表現は，好評表現として「増益となる」「黒字となる」「株価が急騰する」「株価が反発する」，不評表現として「減益となる」「赤字となる」「株価が急落する」「株価が反落する」の計 8 つを用いた．評価表現の選択パラメータは，$\sigma = 10$，$\rho = 0.9$，$\alpha = 0.9$ とし，29 回の繰り返しで収束し，最終的に合計 3,053 の評価表現を獲得することができた．その内訳を表 3 に，得られた評価表現の例を表 4 に示す．また，今回の実験では，評価表現の意味をより限定するために，複合エントリのみを用いることにした．

表 3　獲得した評価表現の数

	単純エントリ	複合エントリ	合計
好評	195	1,282	1,477
不評	185	1,391	1,576
合計	380	2,573	3,053

表 4　獲得した評価表現

	極性	評価表現
単純	好評	上回る，好調だ，伸びる，回復，良好だ など
	不評	下回る，落ちる，赤字，警戒，減額する など
複合	好評	（過去最高を，更新），（業績が，回復する） （計画を，上回る），（期待が，高まる）など
	不評	（赤字と，なる），（株価が，続落） （計画を，下回る），（ことが，嫌気する） など

4.2　株価予測モデル

記事テキストを説明変数として株価予測モデルを構築するアプローチは大きく二つある．一つは，複数の単語の出現回数データから PCA に代表される次元縮約の手法を用いて，少数の説明変数を作成し，それらの説明変数から株価変動を目的変数としたモデルを構築する方法である．この方法の利点は，回帰モデルを初めとする多様なモデリング手法を利用できることである．しかし一方で，次元縮約において得られたクラスタ（もしくは潜在変数）が何を意味しているか不明確になる可能性があり，因果関係の意味解析が難しい．他方の方法は bag-of-words に代表される方法で，多数の単語の出現をそのまま説明変数として用いる方法である．この方法に適用できるモデリング手法は，Naïve Bayes 法などに限定される反面，単語の出現をそのまま扱うために，得られたルールの解釈が比較的容易であるという利点がある．またこのアプローチでは，多数の表現から，予測に貢献するであろう表現を事前に選択することが多い．その方法として，独立性の検定で用いるカイ二乗値を用いる方法から，近年注目を浴びている lasso まで多くの手法が提案されている．本研究では，よりシンプルな方法を用いることにした（後述）．

前節の手法によって得られた評価表現の出現が将来の株価にどのように影響するかについてのモデルを構築する．ここでは，1 日をサンプルに，個々の評価表現の出現を説明変数として，将来の株価の上昇，下降を予測するモデルを構築する．以下に詳細な定義を与える．

日 t における $\tau = 0, 1, 2, \ldots$ 日後の収益率 r_τ を次式で定義する．

$$r_t^\tau = \frac{cl_{t+\tau}}{op_t} - 1.0 \tag{1}$$

ここで，op_t，cl_t は，それぞれ日 t の日経 225 先物価格の始値と終値を表す．

また日 t における $\tau = 0, 1, 2, \ldots$ 日後の株価の変動クラス c_t^τ（up：上昇，down：下降）を次式で定義する．

$$c_t^\tau = \begin{cases} \text{down}, & \text{if } r_t^\tau < 0.0, \\ \text{up}, & \text{else.} \end{cases} \tag{2}$$

次に，説明変数について定義を与える．日 t におけるニュース記事観測期間 ν（区間 $[t-\nu, t]$）に評価表現 $i = 1, 2, \ldots, n$ が出現する記事の件数を $f_i^{t,\nu}$ で表す．一つの記事の中で同じ評価表現が複数回出現していても，それは 1 回とカウントする．日 t における観測期間 ν の評価表現特徴ベクトル $\mathbf{f}^{t,\nu} = (f_1^{t,\nu}, f_2^{t,\nu}, \ldots, f_n^{t,\nu})^\top$

とする．

ここで，日 t について，期間 $[t-\nu, t-1]$ に特徴ベクトル $\mathbf{f}^{t-1,\nu}$ が観測されたとき，τ 日後の株価変動クラス c_t^τ の確率 $p(c_t^\tau|\mathbf{f}^{t-1,\nu})$ を推定する．すなわち，日 t において 1 日前までの ν 日間に出現した評価表現から，日 t の始値を起点として τ 日後の収益率を予測する．$p(c_t^\tau|\mathbf{f}^{t-1,\nu})$ の推定には Naïve Bayes モデルを用いる．

4.3 Naïve Bayes モデル

まずは，簡単のために，評価表現の出現を $w_i = 0, 1$，すなわち，ある日にその評価表現を伴った記事が 1 回でも出現したかどうかを表す評価表現特徴ベクトル $\mathbf{w} = (w_1, w_2, \ldots, w_n)^\top$ を考える．$\mathbf{w}$ の出現を条件とした株価の上下変動クラス c の確率 $p(c|\mathbf{w})$ は，ベイズの定理により式 (3) で表される．

$$p(c|\mathbf{w}) = \frac{p(\mathbf{w}|c)p(c)}{p(\mathbf{w})} \tag{3}$$

分母の $p(\mathbf{w})$ は，c によらず一定のため，分子について見ると，$p(c)$ はクラス c の事前確率で，この確率が評価表現 $\mathbf{w}$ の出現という evidence を得ることで尤度 $p(\mathbf{w}|c)$ によって事後確率 $p(c|\mathbf{w})$ へと更新される．これがベイズの定理が意味することである．

しかしながら，$\mathbf{w}$ の次元が高くなると，単語の同時確率 $p(\mathbf{w}|c)$ の推定が困難となる．そこで，式 (4) に示されるように，すべての単語の出現は独立であるという，ナイーブな仮定をおくことで，容易に $p(\mathbf{w}|c)$ を計算できるようになる．これが Naïve Bayes 法である．

$$p(\mathbf{w}|c) = \prod p(w_i|c) \tag{4}$$

またベイズ推定においては，事前確率 $p(c)$ の設定に恣意性が入ることでしばしば論争の種になるが，株価の変動はランダム・ウォークであることを考えると，$p(c = down) = p(c = up) = 0.5$ を仮定できる．

以上のことを踏まえると $p(c|\mathbf{w})$ は式 (5) で表され，株価の上下変動の推定クラス $\hat{c}$ は，式 (6) によって求めることができる．

$$p(c|\mathbf{w}) \propto \sum_i \ln p(w_i|c) \tag{5}$$

$$\hat{c} = \mathrm{argmax}_c \sum_i \ln p(w_i|c) \tag{6}$$

ここで，本研究で扱う特徴ベクトル $\mathbf{f}$ の要素 f_i は 1 日における評価表現 i の出現頻度である．頻度情報を扱うためには，Multinomial Naïve Bayes モデルを利用する必要がある．詳細は [8] に詳しいが，結論だけを言えば，本研究の文脈では，株価の上下変動の推定クラス $\hat{c}$ は，式 (7) に示されるように，尤度に頻度 f_i を乗じるだけでよい．

$$\hat{c} = \mathrm{argmax}_c \sum_i f_i \ln p(w_i|c) \tag{7}$$

さらに，Naïve Bayes モデルの精度を高める目的で，目的変数 c への影響力の強い評価表現を事前に選択する．式 (8)，(9) について，ユーザが与えた閾値 $minRate$（クラス別出現確率最小値），$minSupp$（最小出現頻度）の両条件を同時に満たす評価表現 i を事前に選択する．ここで，$f_{i,u}(f_{i,d})$ は，株価が上昇（下落）した日に評価表現 i が出現した総件数である．本研究では，$minRate = 0.55, minSupp = 20$ の設定にて実験を行っている．

$$\frac{f_{i,d}}{f_{i,d}+f_{i,u}} \geq minRate \ \mathbf{or} \ \frac{f_{i,u}}{f_{i,d}+f_{i,u}} \geq minRate \tag{8}$$

$$f_i \geq minSupp \tag{9}$$

Naïve Bayes を用いるメリットの一つは，w_i の $\hat{c}$ 推定への貢献度合いを $p(w_i|c)$ すなわち，各クラスにおける評価表現の出現確率によって測ることができ，株価の上下変動の予測への貢献度の高い評価表現の意味的評価が可能となる．

5. 実験

5.1 モデル検証方法

モデル精度の検証は，実際の運用での適用を考慮し以下に示す方法を用いる．日 $t = 1, 2, \ldots, T$ を任意の window サイズ $winSize$ で分割し $m(= \mathrm{ceil}(T/winSize))$ 個の window $b_1, b_2, \ldots, b_m$ を用意する．訓練データとして用いる連続した window の個数 $trainSize$ を定め，b_j のテストデータを $b_{j-trainSize}$ から b_{j-1} の訓練データで構築したモデルによって予測する．$m - trainSize (j = trainSize+1, trainSize+2, \ldots, m)$ 回のモデル構築および out-of-sample によるテストを行う．その結果によってモデルを評価する．$winSize$ と $trainSize$ の大きさによって，モデルの精度は大きく変化することが予想される．パラメータが定常状態にある環境においては，比較的長期間のデータに基づいてモデルを構築し，逆に変化の激しい非定常状態の環境においては短期間のデータに基づき，遠い過去のルールは忘却したほうがよいであろう．これは定常状態と非定常状態の変化の検知が必要となり，本研究の範囲を超えるため，ここでは，$winSize = 25$, $trainSize = 10$ として実験を行った．これは，過去約

1 年のデータで学習し，次の 1 カ月の株価変動を予測することに対応している．

モデル評価は，正答率と Sharpe 比を用いる．また補助的に，年間の収益率，標準偏差も示す．正答率とは，モデルが全テストサンプルに対して正しく予測したサンプルの割合である．Sharpe 比とは，William Sharpe によって考案された，資産運用の効率性を評価する指標で式 (10) で定義される．

$$\text{Sharpe 比} = \frac{E[R_a - R_b]}{\sqrt{V[R_a - R_b]}} \tag{10}$$

ここで R_a は評価対象の資産運用の年間収益率で，R_b は無リスク資産（短期国債）の年間収益率である．この式が意味することは，評価対象の資産運用 a について，資産運用の期待リスク（標準偏差）を一定にした場合の収益率を表したものである．収益率が高ければ高いほど，そしてリスクが小さければ小さいほど，Sharpe 比は大きくなる．一般的に，収益率は年利で計算されることが多く，1.0 を超えると安定した資産運用であると言われている．また，本実験では，無リスク資産の年間収益率がほぼ 0 である現状に鑑みて $R_b = 0$ として計算した．

本稿で提案したモデルで資産運用した場合の，Sharpe 比は式 (11) で計算される．

$$\text{Sharpe 比} = \frac{\frac{1}{T}\sum_{t=1}^{T} a_t \cdot 250}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(a_t - \bar{a})^2 \cdot 250}} \tag{11}$$

ここで，a_t は日 t における手持ち資産の運用収益率平均で，$\bar{a}$ は対象期間における a_i の日平均である．

この計算式の意味するところは，日々の株価の値動きを反映させた資産価値の計算を行うことにある．τ 日後の日経 225 先物価格を予測する場合に，もし τ 日後の収益率のみを評価したとすると，ポジション途中の株価の変動が評価されず，τ が大きくなれば大きくなるほど，Sharpe 比の分母である変動を少なく見積もることになり，結果として Sharpe 比が過大評価されてしまう．

5.2 結果と考察

過去のニュース記事を観察する日数 $\nu = 1, 2, 3, 4, 5$，そして予測日数 $\tau = 0, 1, 2, 3, 4, 5$ についての全組み合せ 30 通りについて実験を行った．結果を表 5 に示す．

驚いたことに，正答率，Sharpe 比ともに，トップの成績を収めたのは過去 1 日のニュース記事を見て 5 日後の収益率を予測するモデルであった．比較的良いモデルは記事の観測日数は短く，予測期間の長いモデルとなっている．例外的に $\nu = 3$，$\tau = 0$ のモデルの成績が良いことがわかる．この中で，最も良いモデルで 52.3%の正答率であり，株価予測に不慣れな人はこの数字を低いと考えるかもしれないが，ランダム・ウォーク仮説から見れば，十分に高い数字であると言える[1]．

表 5 モデルの正答率と Sharpe 比

ν	τ=0	τ=1	τ=2	τ=3	τ=4	τ=5
1	50.4	49.6	49.3	**51.6***	50.1	**52.3****
	−10.7	−2.6	−1.2	6.3	4.2	11.3
	24.6	19.2	18.9	17.4	17.7	16.3
	−0.433	−0.135	−0.064	**0.364**	0.236	**0.689**
2	**51.9****	50.1	50.2	49.8	50.5	**51.4***
	2.9	1.0	−0.9	4.6	7.2	6.0
	24.6	21.0	21.3	19.8	20.3	19.4
	0.120	0.047	−0.042	0.233	**0.354**	**0.307**
3	**52.0****	50.2	48.1	50.4	50.2	49.3
	8.8	0.3	−5.7	1.0	5.0	0.8
	24.5	22.9	21.7	21.2	21.3	21.1
	0.359	0.012	−0.261	0.046	0.236	0.039
4	50.1	49.5	48.3	49.0	50.3	48.2
	−9.1	−5.9	−4.3	−2.2	0.4	−6.1
	24.5	23.2	22.6	22.6	22.4	21.1
	−0.372	−0.254	−0.192	−0.096	0.020	−0.288
5	49.1	49.4	50.6	48.7	50.2	48.7
	−4.3	−4.4	−1.7	−7.9	−2.3	−4.6
	24.4	23.1	23.0	23.4	22.6	21.2
	−0.175	−0.190	−0.074	−0.337	−0.103	−0.217

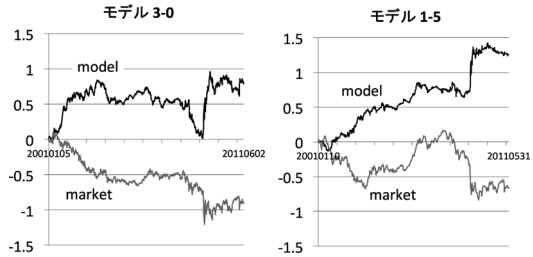
各行は，モデル構築時における過去の参照データ日数を，列は将来の予測日数を表す．各セル内，上から正答率（単位は%），年収益率（同%），標準偏差（同%），Sharpe 比を示す．例えば，過去 1 日のデータで，将来 5 日後を予測するモデルの正答率は 52.3%で年率 11.3%の収益率を達成し，標準偏差は 16.3%で，Sharpe 比は 0.689 である．Sharpe 比の太字は Sharpe 比が 0.3 以上を表す．正答率の太字は，**は 5%有意，*は 10%有意を示す．

次に，上述の実験結果から成績の良かった二つのモデル $\nu = 3$，$\tau = 0$（モデル 3-0 と呼ぶ）および $\nu = 1, \tau = 5$（モデル 1-5 と呼ぶ）について，式 (11) に示された日別運用収益平均 a_t の累積推移を図 2 に示す．これらの二つのモデルを使って資産運用を行うと，約 10 年間で，モデル 3-0 で約 1.8 倍，モデル 1-5 で約 2.3 倍になることがわかる．グレーの実線で表されるモデルを用いない運用に比べると，比較的良い成績であると言えよう．

両モデルに共通して言えることは，2000〜2002 年までは安定的な成績を残し，また 2008 年後半のリーマンショックで大幅なゲインを得ていることである．しかしながら一方で，リーマンショック直前の半年の成績で明暗が分かれ，モデル 3-0 では大幅に下落し，モデル 1-5 ではイーブンな成績をキープしている．

これは，モデル 3-0 が過去 3 日の情報を基に当日の始値と終値で取引するモデルであり，株式市場の足下の動向とその見方についてのズレが大きい場合は，投資モデルとしては適していない可能性が考えられる．

[1] H0：正答率=0.5，H1：正答率 >0.5 における，正答率 =0.523 の有意確率は 0.016.



黒の実線はモデルに従った資産運用で，グレーの実線は，モデルを用いずに日経 225 先物を毎日購入した場合の資産運用を示す．

図 2　日別の運用収益率平均の累積推移

リーマン・ショック前の時期には，米国で深刻化していたサブプライムローン問題が，米国内の問題であり日本市場は影響を受けないという楽観的な見方と，大きな不安を抱く見方が混在していた．それらを反映して，株式市場も下落トレンドにありながらも時折反発するという動きを繰り返しており，将来 1 日しかみないモデル 3-0 では，変動の激しい目先の利益にフィットしたモデルが構築され，より長期的で緩やかな下落に対応できなかった．一方で 5 日後を目的変数としたモデル 1-5 では，より長期的な下落傾向を学習できたために，モデル 3-0 に比べ，資産の減少が少なく済んだと考えられる．

リーマン・ショックを世界的な金融危機だと，市場関係者が気づきはじめた 2008 年 8 月後半からは，株式市場のトレンドと市場関係者の認識が一致し，両モデルとも成績が急回復している．

さらに，両モデルの予測性能の違いは何かを探るために，特徴選択で選ばれた評価表現の内容を分析した．

5.3　内容分析

モデル 3-0 とモデル 1-5 の両モデルに採用された評価表現の内容を概観することによって両モデルの特徴を考察してみたい．とりわけ両モデルのパフォーマンスに大きな差異が生まれるのはリーマン・ショック前から危機発生までの期間であるため，2006 年 10 月 10 日から 2008 年 8 月 15 日までの間に両モデルの特徴として出現した 1055 の評価表現内容を概観した．モデル 3-0 に選択されている評価表現は，「○○ 円まで急落」「大幅安となる」「株価の反発が大幅だ」「株価が急騰」等の株価の急激な動向を示すものが多い．一方，モデル 1-5 に選択されている評価表現には，「減額修正となる」「○○ 億円に下方修正する」「投資判断を格上げする」「収益拡大期待が広がる」などの業績に関連する評価表現が多い．これらの特徴は，両モデルの性格を表していると言えよう．3-0 モデルは 1 日の動向を的中させるために学習したものであるから，直近の株価動向を表す評価表現が選択され，1-5 モデルには，投資家が咀嚼し株価に反映させるまでに数日を要する業績関連の評価表現が多く選択されたと推察される．

6.　終わりに

本稿では，ニュース記事から極性付き評価表現辞書を構築し，評価表現を利用した Naïve Bayes モデルで日経 225 先物の収益率を予測した．過去 1 日のニュース記事を観察し，5 日後の収益率を予測するモデルの精度が最も良く，正答率 52.3%，年平均収益率 11.3%，Sharpe 比 0.689 を達成し，10 年間の資産運用で約 2.3 倍に資産が増加することがわかった．また，このモデルの内容としても，業績に関する評価表現が多く出現しており，ニュースが出てから株価に反映されるまでの 5 日間のタイムラグをうまくとらえることで収益率を増加させていると推察できた．本研究で構築した株価予測モデルの有効性が示されたと言えよう．今後は，価格変動をより正確にとらえるために，評価表現をクラスタリングすることで汎化性能を向上させるなど，さらなる精度の向上を目指したい．

謝辞　本研究の一部は，ERATO 湊離散構造処理系プロジェクト，公益財団法人石井記念証券研究振興財団，および科学研究費補助金基盤研究 (B) の研究助成を受けている．

参考文献

[1] W. Antweiler and M. Z. Frank, Is all that talk just noise? the information content of internet stock message boards, *Journal of Finance*, **59**(3), 1259–1294, June 2004.

[2] J. Bollen, H. Mao and X. Zeng, Twitter mood predicts the stock market, *Journal of Computational Science*, **2**(1), 1–8, March 2011.

[3] G. P. C. Fung, J. X. Yu and H. Lu, The predicting power of textual information on financial markets, *IEEE Intelligent Informatics Bulletin*, **5**(1), June 2005.

[4] G. Gidófalvi, Using news articles to predict stock price movements, 2001, Department of Computer Science and Engineering, University of California, San Diego.

[5] J. Kamps, M. Marx, R. Mokken and M. De Rijke, Using wordnet to measure semantic orientation of adjectives, *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 1115–1118, 2004.

[6] B. G. Malkiel, *A Random Walk Down Wall Street, 1st edition*, New York, W. W. Norton & Co., 1973.

[7] M. Mittermayer, Forecasting intraday stock price trends with text mining techniques, In *The Proceed-*

ings of the Hawaii's International Conference on System Sciences, 2004.
[8] J. D. M. Rennie, L. Shih, J. Teevan and D. R. Karger, Tackling the poor assumptions of naive bayes text classifiers, In *Proceedings of the Twentieth International Conference on Machine Learning*, 616–623, 2003.
[9] R. P. Schumaker and H. Chen, Textual analysis of stock market prediction using financial news articles, In *12th Americas Conference on Information Systems (AMCIS-2006)*, 2006.
[10] R. P. Schumaker and H. Chen, Textual analysis of stock market prediction using breaking financial news: The azfintext system, *ACM Transactions on Information Systems*, **27**(2), February 2009.
[11] P. C. Tetlock, Giving content to investor sentiment: The role of media in the stock market, *Journal of Finance*, **62**(3), 1139–1168, June 2007.
[12] P. C. Tetlock, All the news that's fit to reprint: Do investors react to stale information?, *Review of Financial Studies*, **24**(5), 1481–1512, May 2011.
[13] P. C. Tetlock, M. Saar-Tsechansky and S. Macskassy, More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance*, **63**(3), 1437–1467, June 2008.
[14] P. D. Turney, Thumbs up or thumbs down semantic orientation applied to unsupervised classification of reviews, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 417–424, 2002.
[15] 那須川哲哉，金山博，文脈一貫性を利用した極性付評価表現の語彙獲得，情報処理学会自然言語処理研究会 (NL-162-16), 109–116, 2004.

