

# 「大規模ニュースデータと株価収益率の予測可能性について」

- 探索的アプローチによるモデル構築と out of sample の実証テスト -

## “Return Predictability in News : Naïve-Bayes based investment model using Bloomberg News”

前川浩基<sup>1</sup> 中原孝信<sup>1</sup> 岡田克彦<sup>2</sup> 羽室行信<sup>2</sup>

Hiroki Maegawa<sup>1</sup>, Takanobu Nakahara<sup>1</sup>, Katsuhiko Okada<sup>2</sup> and Yukinobu Hamuro<sup>2</sup>

<sup>1</sup> 株式会社 Magne-Max Capital Management

<sup>1</sup> Magne-Max Capital Management

<sup>2</sup> 関西学院大学大学院経営戦略研究科

<sup>2</sup> Kwansai Gakuin University, Institute of Business and Accounting

**Abstract:** We investigate the return predictability in news using a large sample of electronic news article disseminated from the Bloomberg. Using Naïve-Bayes classifier, we create models that predict return of the Nikkei 225 futures. Our test reveals two models that beat the market by substantial margin. Based on the two selected models, we conduct out of sample money management simulation for the period between 2001 to 2011 and achieve 11.3% return per annum with the Sharpe ratio of 0.693. The result is better than the top 5% of the 1000 random models.

### 1. はじめに

本稿の目的は、大量の記事データから投資家が反応している可能性のある「言葉」を探索し、そうした言葉から構成される投資モデルを構築し、言葉の中に将来の株価の予測可能性をもつ情報があるかどうかを検証することである。仮に投資家が、ニュース記事の中にある言語表現に対して反応し、相場見通しを形成し、それによってある一定方向に投資行動が集約されているのであれば、それは非合理的な行動であるかもしれないが、株式市場に与える影響は大きいといえるだろう。記事データの中に含まれる語句の中に株価の予測可能性を示す情報があるのであれば、こうした投資家行動をシステムティックに把握することが可能となり、有効な投資モデルを構築することができる。

効率的市場においては、最も緩い *weak form market efficiency* を仮定したとしても過去の価格情報は全て株価に反映されているはずである。したがって、株価はランダムに到来する新情報によってのみ変動するため、株価の動きはランダム・ウォークであり、どれほど過去の価格パターンを研究したとしても、将来予測は不可能だと考えられている。しかしながら、現実の市場では多くの投資家がテクニカル分析に傾注し、罫線のパターンを見て投資判断している。あるいは、ファンダ

メンタル分析を中心に行う投資家は、既出の事実情報に依拠しながら、真のファンダメンタル価値に対して、現在価格は割高である、割安であるという認識をもって売買判断している。これらの投資家は、効率的市場仮説においてはノイズトレーダーと片付けられており、彼らの投資が資産価格に与える影響はトータルで相殺されると考えられている。過去を研究して意思決定する投資家がノイズであるならば、彼らの投資行動をいかに正確に予想したとしても、将来の株価を、偶然を上回る確率で予測することは不可能である。即ち、どのような予測システムもおしなべてみると50%程度の精度しか持たないということである。しかしノイズトレーダーの動向が実はランダムではなく、ある株式市場の「場味」「雰囲気」という共通のファクターに影響されているとすれば、短期的には株価の予測可能性が生まれよう。

本稿では、トレーニング期間においてニュース記事の中の「言葉」と相場動向の関係性をモデル化し、それを *out of sample data* で検証することで、投資家がシステムティックに「言葉」に影響されながら、ある決まった方向に投資の意思決定を下しているかを明らかにする。

### 2. データと方法論

## 2-1 ニュース記事

本研究はデータ中心科学の方法論を援用して、大量のニュースデータから株価の予測可能性を持つ「言葉」を探索することを特徴としている。そのため、使用するニュースデータは12年間にわたる大規模な日本語による経済関連データである。これらのデータはブルムバーグ社が日々配信するものであり、ファンドマネージャーやトレーダーなどのプロが着目しているニュースデータである。新聞記事と同じ内容が配信されるものも含まれるが、新聞印刷に間に合わないようなタイムリーな内容の記事も多く、市場に与える影響は単独の新聞記事と比較しても大きいと考えられる。各ニュースにはタグと呼ばれる企業コードが付されており、どの記事がどの企業(群)について書かれたものか判別可能となっている。また、同社から配信されるニュースは、経済関連以外にもスポーツや新聞の社会面に載るような記事もあるが、今回は解析の対象外とした。[5]の研究に見られるように、世の中の一般的なニュースが株価に影響を与える可能性もあるが、ブルムバーグニュースの多くが経済関連である事に鑑み、経済ニュースのみを対象とすることにした。

図表1に示すのは、対象となった記事の件数、記事の中の文字数、単語数、評価表現数(後述)である。対象記事は企業名のタグがついているものに限定しているため、株式投資ブームがあった2001年、2004年から2006年までと、リーマン・ショック時の2008年に記事数が多くなっている。

図表1 2000年1月から2011年12月までの記事数推移

年	記事数	文字数	単語数	評価表現数
2000	22,921	7,885,629	1,569,865	60,865
2001	26,834	9,997,082	2,010,493	97,701
2002	21,887	8,474,244	1,692,046	99,628
2003	17,430	7,973,477	1,611,431	101,186
2004	23,858	10,337,200	2,086,540	102,733
2005	28,958	12,262,796	2,498,105	133,854
2006	24,868	11,137,319	2,271,411	97,546
2007	21,170	7,943,778	1,621,329	70,270
2008	34,664	9,652,851	1,974,293	107,959
2009	23,428	5,339,554	1,080,999	57,439
2010	20,564	3,799,933	763,383	30,812
2011	27,129	4,323,732	871,360	26,814
計	293,711	99,127,595	20,051,255	986,807

## 2-2 方法論

### 2-2-1 評価表現辞書の構築

本稿では、ニュース記事に含まれる語句を機械的に処理し、その中のキーワードと株価の関連性

を探る。したがって、ニュース記事に含まれる様々な表現を認識させる辞書を構築する必要がある。これを評価表現辞書と呼ぶ。評価表現とは、「回復する」「株価が上昇する」(好評表現)「下落する」「上値が重い」(不評表現)など事象を評価する表現のことを指す。筆者らはこれらの評価表現に好評・不評の二つの極性を付した極性付き評価表現辞書を構築した。通常、辞書の構築方法にはいくつか選択肢がある。日本語の場合、既に構築された辞書が公開されているためそれを活用することもできる。しかし、株式市場についての表現には業界特有の言い回しが多く、既存の辞書では対応が難しい。そこで、[2]が提唱する、周辺文脈の一貫性を応用した辞書の構築方法を採用し、自ら評価表現辞書を作成することにした。

この方法では、当初に筆者らがいくつかの評価表現を種表現として準備しておき、それを緒にブートストラップ的に評価表現を増やす。例えば前述の「上値が重い」は不評の種表現とした。この表現を含む文章は不評な事象を記述していると認識する。また前後の文脈は反対の意味を示す「しかし」等の表現が出現しないかぎり一貫するという特徴を持つことから、前後の文章も不評表現で構成されていると仮定する。こうして新たな表現を獲得し、種表現から語彙を増やすことで評価表現辞書を構築するのである。詳しくは[2]を参照されたい。

こうして作成された株式市場特有の言い回しを含む評価表現辞書は3053の表現から構成され、好評・不評とも概ね半々の割合である。本稿で構築する投資モデルの説明変数は全てこの辞書にある評価表現である。

### 2-2-2 株価予測モデルの構築

本稿の目的は記事の中のテキスト情報に含まれる株価の予測可能性を探索的に求め、それをout of sampleのデータで検証することである。そのため、記事テキストに含まれる評価表現の出現と株価の将来リターンの関係を表すモデルを構築する。このような記事テキストを説明変数として株価予測を行うアプローチには大きく二つ存在する。一つは、複数の出現回数データから主成分分析(PCA)と呼ば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(例えば[10])。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、次元縮約において得られたクラスター(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。

もう一つの方法は Bag-of-words アプローチ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。後述するナイーブ・ベイズ (Naïve Bayes) モデルなど限られたモデリング手法に限定される反面、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だという利点がある。本稿ではニュース記事に出現する評価表現をそのまま説明変数とする後者のアプローチを採用する。

以下、モデルの被説明変数、説明変数の順に詳細な定義を与える。

日  $t$  における  $r_t = 0, 1, 2 \dots$  日後の収益率  $r_t$  を次式で定義する。

$$r_t = \frac{c_{t+1} - c_t}{c_t} \quad (1)$$

ここで、 $op_t$ ,  $cl_t$  は日経225先物価格の日  $t$  における寄付値と引け値を意味する。また、日  $t$  における  $r=0, 1, 2 \dots$  日後の株価変動のクラス  $c_t^r$  は  $r_t < 0$  であれば下落、それ以外であれば上昇と定義する。この株価変動のクラスが被説明変数である。次に説明変数についての定義を与える。日  $t$  におけるニュース記事観測期間  $v$  (区間  $(t-v, t)$ ) に評価表現  $s=1, 2, \dots, n$  が出現する記事の件数を  $f_s^{t,v}$  で表す。一つの記事の中で同一の評価表現が複数回出現していても1回のカウントとする。日  $t$  における観測期間  $v$  の評価表現特徴ベクトル  $\mathbf{f}^{t,v} = (f_1^{t,v}, f_2^{t,v}, \dots, f_n^{t,v})^T$  とする。ここで、日  $t$  について、期間  $(t-v-1, t-1)$  に特徴ベクトル  $\mathbf{f}^{t-1,v}$  が観測された時、 $r$  日後の株価変動クラス  $c_t^r$  の確率  $p(c_t^r | \mathbf{f}^{t-1,v})$  を推定する。即ち、日  $t$  において、1日前までの  $v$  日間に出現した評価表現から、日  $t$  の寄付値を起点として  $r$  日後までの収益率を予測するのである。確率  $p(c_t^r | \mathbf{f}^{t-1,v})$  の推定にはナイーブ・ベイズ法を用いる。

### 2-2-3 ナイーブ・ベイズモデル

まずはナイーブ・ベイズモデルを簡単に説明するために、評価表現の出現を  $\omega_c = 0, 1$ 、即ち出たか出なかったかで表現する。ある日にその評価表現を伴った記事が1回でも出現した場合、必ず評価表現特徴ベクトルを  $\mathbf{w} = (\omega_1, \omega_2, \dots, \omega_n)^T$  とする。 $\mathbf{w}$  の出現を条件としたある日の日経225先物の上昇・下落のクラス  $c$  の確率はベイズの定理より (2) 式であらわされる。

$$p(c|\mathbf{w}) = \frac{p(\mathbf{w}|c)p(c)}{p(\mathbf{w})} \quad (2)$$

分母の  $p(\mathbf{w})$  は、 $c$  によらず一定であるため分子だけを考えると、 $p(c)$  は上昇するか下落するかの事前確率であり、この確率が評価表現  $\mathbf{w}$  の出現という事象が起こることで尤度  $p(\mathbf{w}|c)$  によって事後確率  $p(c|\mathbf{w})$  に更新される。これがベイズの定理の意味するところである。しかし、 $\mathbf{w}$  の次元が高くなると、評価表現の同時出現確率  $p(\mathbf{w}|c)$  の推定が困難になるので、全ての評価表現は独立に出現するというナイーブな前提を置くことによって  $p(\mathbf{w}|c)$  を推定するのである。この手法はスパムメールか否かを分類するために活用されており、前提条件がすべての言葉の出現パターンがすべて独立であるというナイーブなものであるにも関わらず、非常に頑健な分類結果が得られることが一般に知られている。この前提のもとに、 $p(c|\mathbf{w})$  は(3)式で求められる。

$$p(c|\mathbf{w}) \propto \sum_i \ln p(\omega_i|c) \quad (3)$$

また日経225先物の推定変動クラス  $c$  は(4)式で求められる。

$$c = \operatorname{argmax}_c \sum_i \ln p(\omega_i|c) \quad (4)$$

(4)式の直感的説明としては、ニュース記事に出現する評価表現からニュース発生後に株式市場が上昇、あるいは下落の確率が高い評価表現の組み合わせを考えるということである。ただ、この方法であれば記事の中に出現したすべての評価表現を説明変数とすることになる。株価動向にシステムティックに影響を与えうる評価表現からのみモデルを構築するため、本稿では、より選択的に説明変数を選択するため、出現頻度で評価表現の重み付けをすることにする。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が、本稿では詳細な説明は割愛する。詳しくは[8]を参照されたい。一日における評価表現  $i$  の出現頻度の特徴ベクトル  $\mathbf{f}$  の要素を  $f_i$  とすると、最終的に本研究で推定に用いるモデルは(5)式で表される。また、この精度を高める目的で、モデルで用いる評価表現についても最低20回の出現頻度を持つもので且つ、上昇か下落のどちらかに0.55以上の偏りをもって出現する評価表現に限定した。

$$c = \operatorname{argmax}_c \sum_i f_i \ln p(\omega_i|c) \quad (5)$$

### 2-3 投資モデルの検証

株価予測モデルの制度の検証は、実際の運用での適用を考慮し、次に示す方法を用いる。日  $t = 1, 2, \dots, T$  を任意の  $winSize$  (window size) で分割し  $m$  個の window  $b_1, b_2, \dots, b_m$  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、 $b_1$  のテストデータを  $b_1 - trainSize$  から  $b_1 - 1$  の訓練データで構築したモデルによって予測する。 $m - trainSize$  ( $i = trainSize + 1, trainSize + 2, \dots, m$ ) 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。株式市場の動向が比較的ボックスレンジで取引されるような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、逆に変化の激しく、トレンドを形成しながら強気相場が継続するような非定常状態の環境においては短期間のデータにもとづき、遠い過去のルールは忘却した方が良くであろう。定常と非定常を繰り返す株式市場に対して何らかの指標を作成し、それを検知しながら可変モデルを作成することも考えられるが、それは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結果を報告する。これは過去約1年のデータで学習し、次の1ヶ月の株価変動を予測することに対応している。

モデルの評価は、正答率と Sharpe 比を用いる。また、補助的に、投資モデルにしたがって投資した場合の年間収益率、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。また Sharpe 比を計算する過程で必要となる短期の無リスク利率は、ほぼゼロ金利である現状に鑑みてゼロとする。したがって、本稿で提案したモデルで資産運用した場合の Sharpe 比は (6) 式であらわされる。

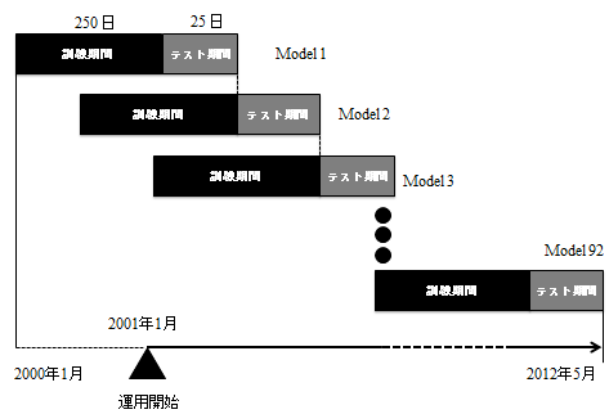
$$S = \frac{\frac{1}{T} \sum_{t=1}^T a_t \times 250}{\sqrt{\frac{1}{T} \sum_{t=1}^T (a_t - \bar{a})^2 \times 250}} \quad (6)$$

ここで、 $a_t$  は日  $t$  における手持ち資産の平均リターンで、 $\bar{a}$  は対象期間における  $a_t$  の日平均である。分子は手持ち資産の平均リターンを年率換算表示したものであり、分母は年率換算した標準偏差である。単位当たりの不確実性に対する報酬を最大化できる投資モデルが良いモデルだと評価される。

## 2-4 12年間に作成されたモデル

この検証を行うにあたって、先物取引の売買注文は寄付値で発注することから、当日 8 時 45 分までに配信された記事を探検対象とする。また、午前 0 時から 8 時 45 分までのタイムスタンプが押されているニュースは前日に配信されたものとして扱う。記事データは 2000 年 1 月から 2011 年 5 月までの約 12 年間に発信された約 29 万記事を活用する。過去 1 年間の記事と先物価格の動向から売買指示を出すため、売買指示が発せられるのは 2001 年 1 月からである。モデルの構築のために必要な記事データは過去 1 年である。推定したモデルを 1 ヶ月 (25 日) は使用し、1 ヶ月が過ぎるともう一度その時点から遡る 1 年間のデータで再度モデルを構築する。理想的には日々更新するのが良いと考えられるが、計算時間の都合上 25 日間隔でモデル再構築することにした<sup>1</sup>。従って、全検証期間の 12 年間に作成したモデルの数は全部で 92 個である。図表 2 にイメージ図を記す。

図表2 投資モデル構築、検証のイメージ図



## 3. 実験結果

### 3-1 正答率、Sharpe比

<sup>1</sup> 評価表現を説明変数とするモデルであるため変数が数百に上る。したがって、一つのモデルを推定するためかなりの計算時間を要する。一日毎にウィンドウをずらしながら大量のモデルを推定するためには 25 倍の計算時間がかかるが、それほど大きな改善効果が見込めるとは考えにくいので今回は断念した。

これら92個のモデルに依拠して、過去のニュース記事を評価し、 $\tau = 0, 1, 2, 3, 4, 5$ 、そして予測日数 $\tau = 0, 1, 2, 3, 4, 5$ についての全組み合わせ30通りについて実験を行った。図表3に示すのはその結果である。例えば $\nu=1$ 、 $\tau=0$ の場合、前日のニュース記事を評価し、当日のintra-dayの取引で手仕舞うという意味である。即ち、一日前のニュース記事に現れる評価表現を(5)式から得られたモデルで評価し、当日の寄り付き値で日経225先物をLong(推定 $\hat{c}_t$ が上昇を示している場合)あるいは、Short(推定 $\hat{c}_t$ が下落を示している場合)し、引け値でスクエアに戻すという取引をすることである。このモデルで運用した場合、上昇・下落の正答率は50.4%でデタラメに売買を繰り返すのとほぼ同等であった。また、年率リターンは-10.7%となっている。この投資モデルに従えば約11年間で元本を毀損することになる<sup>2</sup>。

ニュース記事の評価表現に何らの将来の予測可能性が包含されていないのであれば、正答率はランダム・ウォークの期待値である0.5を有意には上回らないはずである。しかし、最も良い成績を収めた過去1日のニュース記事を構築したモデルで評価して、5日後を予測する $\nu=1$ 、 $\tau = 5$  (投資モデル1-5) に依拠して12年間トレードすれば、52.3%の正答率と11.3%の年率リターンを達成できた。前日のニュース記事の評価表現の出現パターンをモデルで評価し、5日間ポジションを持って5日後の引け値で手仕舞うという手法である。達成された正答率は、ラスベガスのカジノルーレットが「00」の目が主催者側の勝ちという設定にあることで37/36だけ有利なことに鑑みると、非常に大きな正答率だと言えよう<sup>3</sup>。正答率=0.5 であるという帰無仮説 $H_0$ に対する対立仮説 $H_1$ 正答率 $>0.5$ における、0.523の正答率の有意確率は0.016である。投資モデル1-5のシャープ・レシオは0.693であった。

過去3日のニュース記事に出現する評価表現をモデルで評価し、当日の寄り付きでポジションを取り、引け値でスクエアにするという投資モデル3-0も有望である。正答率52.0%、年率リターン8.8%、シャープ比で0.359 である。また、正答率からだけ判断すれば、投資モデル1-3、 2-0、 2-5も有意

な結果となっている。

図表3: 2001年1月5日から2012年5月31日までの評価表現から構成される投資モデルの正答率, リターン, リスク, シャープ比

$\nu$		$\tau=0$	$\tau=1$	$\tau=2$	$\tau=3$	$\tau=4$	$\tau=5$
1	正答率(%)	50.4	49.6	49.3	51.6*	50.1	52.3**
	年率リターン(%)	-10.7	-2.6	-1.2	6.3	4.2	11.3
	標準偏差(%)	24.6	19.2	18.9	17.4	17.7	16.3
	Sharpe比	-0.435	-0.135	-0.063	0.362	0.237	0.693
2	正答率(%)	51.9**	50.1	50.2	49.8	50.5	51.4**
	年率リターン(%)	2.9	1.0	-0.9	4.6	7.2	6.0
	標準偏差(%)	24.6	21.0	21.3	19.8	20.3	19.4
	Sharpe比	0.118	0.048	-0.042	0.232	0.355	0.309
3	正答率(%)	52.0**	50.2	48.1	50.4	50.2	49.3
	年率リターン(%)	8.8	0.3	-5.7	1.0	5.0	0.8
	標準偏差(%)	24.5	22.9	21.7	21.2	21.3	21.1
	Sharpe比	0.359	0.013	-0.263	0.047	0.235	0.038
4	正答率(%)	50.1	49.5	48.3	49	50.3	48.2
	年率リターン(%)	-9.1	-5.9	-4.3	-2.2	0.4	-6.1
	標準偏差(%)	24.5	23.2	22.6	22.6	22.4	21.1
	Sharpe比	-0.371	-0.254	-0.190	-0.097	0.018	-0.289
5	正答率(%)	49.1	49.4	50.6	48.7	50.2	48.7
	年率リターン(%)	-4.3	-4.4	-1.7	-7.9	-2.3	-4.6
	標準偏差(%)	24.4	23.1	23	23.4	22.6	21.2
	Sharpe比	-0.176	-0.190	-0.074	-0.338	-0.102	-0.217

(注) \*,\*\*はモデルの正答率=0.5であるという帰無仮説をそれぞれ10%,5%の信頼区間で棄却することを示す。

### 3-2 有効な投資モデルの時系列検証

これまでの結果から有効な投資モデルである1-5と3-0について現実の運用に即した視点で再評価して行くことにする。取引の対象は日経225株価指数先物であり、現物の株式を取引するのに比較するとマーケットインパクトも小さく手数料も安価であるため、示されたリターンは概ね実現可能性の高いものと言える。

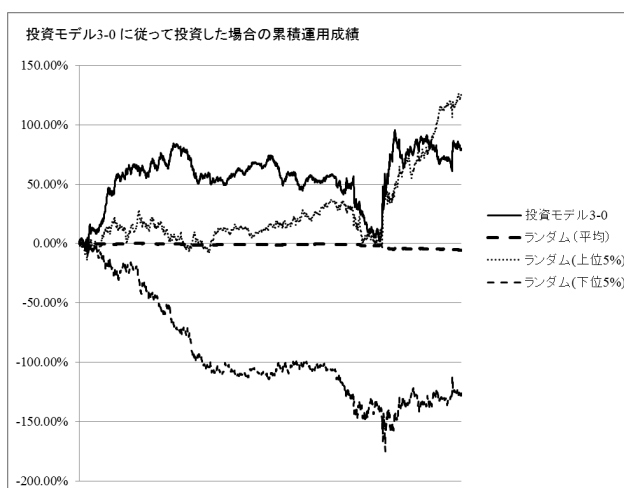
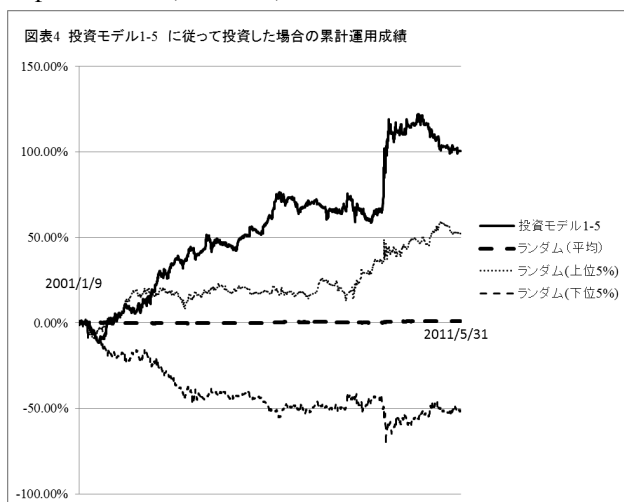
図表4に示すのは、2000年1月からニュース記事データで1年間トレーニングしたモデルを使ってout-of-sampleの投資を行い、その収益率を日毎に累積したものと、ランダムに毎日売買を繰り返したモデルの成績を比較したものである。ランダムに売買を繰り返すモデルは、投資モデルの期間に対応した取引を繰り返す。例えば、投資モデル1-5との比較においては、毎日コインを投げ表であれば日経225先物をLongして5日後に売り戻す、裏であればShortして5日後に買い戻すという具合に毎日売買し11年間の取引を行うモデルのことである。こうした手法で11年間の取引を行うのが1試行であり、これを1000回繰り返し、その平均をとったのが太い点線で示しているランダム(平均)である。平均的には10年間のリターンはほぼ0であり、株価動向がランダム・ウォークであるという仮説を裏付けている。平均では0であるが、偶然高リターンを生む戦略や低いリターンとなる戦略は存在する。それらがランダム(上位5%)、ランダム(下位5%)である。本稿では良いモデルを単純に期末においてリターンの高いものではなく、Sharpe比の高いものと定義しているの

<sup>2</sup> 正答率が50%でも小さく正解し、大きく誤れば平均リターンは負となる。

<sup>3</sup> ルーレットでは顧客の1ドルの期待収益が36/37ドルである。

1000回の試行の中で上位5%（50番目）と下位5%（950番目）のSharpe比のモデルの日毎の累積リターンを点線で表示している。この上位5%と下位5%の軌跡が経験的分布(empirical distribution)の5%の信頼区間に対応している。

投資モデル1-5の成績はこれらの経験的p値(empirical p-value)で0.007(7番目)であり、評価表現の出現頻度の中に何らかの将来情報が含まれている可能性を示している。一方、投資モデル3-0については、最終的なリターンは上位5%のランダムモデルを超えることができず、Sharpe比の経験的p値は0.136(136番目)であった。



#### 4. 結論

本稿では、データ中心科学の手法を援用して、大規模ニュース記事の中に株価の将来予測可能性が包含されているかを検証した。使用したデータは、ブルムバーグにより配信された12年分のニュース記事約29万件と日経225の先物価格である。

ニュース記事については、既に構築している辞書を用いて、評価表現を抽出し、評価表現の出現と株価指数の上下動を過去1年間にわたって学習させた。学習については、ナイーブ・ベイズ法を用いて、どういった評価表現が上昇（下落）に確率的に結びつきやすいかを認識させた。投資モデルは、直近のニュース記事を学習したモデルで評価し、売買指示を出す。直近どの範囲までの記事を対象とし、どこまでの予測をするのかについて全60通りの組み合わせを用意し、日々将来の日経225先物の上下動を予測しながら11年間運用した。

実験の結果、いくつかのモデルの成績が偶然では得られない結果を残すことが明らかとなった。投資モデルの評価基準には、どの程度正しく上下動を当てることができたかという正答率と、その投資モデルで運用した場合のSharpe比を用いた。正答率で評価した場合、1-3、1-5、2-0、2-5、3-0の5つのモデル（1-3は前日のニュースを評価して3日後を予測するモデル）が有意な結果を出すことがわかった。これらの5つのモデルの中から、Sharpe比が高かった2つのモデルについて、運用シミュレーションを行った結果、前日の記事の評価し、翌日の寄り付きでポジションを取り、5日後に手仕舞うというモデルが非常に良い成績を収めた。ランダムに日々ポジションを取り、5日後に手仕舞うという投資実験を1000回行った実験結果と比較しても、Sharpe比で5%の有意水準をはるかに超える7番目の成績であった。

ニュース記事における評価表現の出現パターンは将来リターンに影響を与えている可能性が高い。今回の検証では、評価表現を機械的に組み合わせることでニュース記事を評価した結果を投資モデルに応用している。評価表現の組み合わせの中にも、より意味的妥当性があり、投資家の反応度合いも強いものが存在する可能性がある。そういったものに重み付けをする等の工夫をすることにより、より投資家行動を正確に予測するモデルの構築が可能になると思われる。今後の課題としたい。

#### (謝辞)

本研究は、石井記念財団、科学研究費基盤 B、ERATO 湊離散構造処理系プロジェクトから研究助成を受けて遂行された。ここに謝意を記したい。

参考文献

- [1] 岡田克彦, 羽室行信「相場の感情とその変動 – 自然言語処理で測定するマーケットセンチメントとボラティリティ」*証券アナリストジャーナル* 49 (8), pp.37-48. (2011)
- [2] 那須川哲也, 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)*, pp.109-116. (2004)
- [3] 羽室行信, 岡田克彦, 森田裕之「周辺文脈アプローチを利用した新聞記事内容と株価に関する分析」*日本オペレーションズ・リサーチ学会秋季アブストラクト集* pp.128-129. (2010)
- [4] Antweiler, W and M.Z. Frank, “Is all that talk just noise?: the information content of internet stock message boards,” *Journal of Finance*, 59 (3), pp.1259-1294. (2004)
- [5] Bollen, J., H. Mao and X. Zeng, “Twitter mood predicts the stock market,” *Journal of Computational Science*, 2 (1), pp.1-8. (2011)
- [6] Fand, L and J. Peress, “Media coverage and the cross-section of stock returns,” *Journal of Finance*, 64 (5), pp.2023-2052. (2009)
- [7] Merton, R., “A simple model of capital market equilibrium with incomplete information,” *Journal of Finance*, 42, pp.483-510. (1987)
- [8] Rennie, J., L. Shih, J. Teevan and D. R. Karger, “Tackling the poor assumptions of naïve bayes text classifiers,” *In Proceedings of the Twentieth International Conference on Machine Learning*, pp.616-623. (2003)
- [9] Takahashi, S., H. Takahashi, and K.Tsuda “Analysis of the effect of headline news in financial market through text categorization,” *International Journal of Computer Application in Technology*, 35, pp.204-209. (2009),
- [10] Tetlock, P., “Giving content to investor sentiment: The role of media in the stock market,” *Journal of Finance*, 62 (3), pp.1139-1168. (2007)
- [11] Tetlock, P., M. Saar-Tsechansky and S. Macskassy “More than words: Quantifying language to measure firms’ fundamentals. *Journal of Finance*, 63 (3), pp.1437-1467. (2008),