人工知能学会研究会資料
SIG-FIN-011-06

日次データを用いた市場センチメント・インデックスの構築と株価説明力の分析

A Daily Market Sentiment Index and its Predictability of Stock Prices

石島博 [1]　數見拓朗 [2]　前田章 [3]

Hiroshi ISHIJIMA [1], Takuro KAZUMI [2], and Akira MAEDA [3]

[1] 中央大学
[1] Chuo University
[2] ㈱サイバーエージェント
[2] CyberAgent, Inc.
[3] 東京大学
[3] University of Tokyo

Abstract: This study analyzes the sentiment of the Japanese economy that might appear in daily news articles. To quantify such a sentiment, we created an index that accounts for the frequency of occurrence of the words that affirmatively or negatively describe the current economic situation. Using articles taken from the Nikkei, we counted the numbers of “positive” as well as “negative” words in the articles. Constructing a daily summary index, we then conducted statistical analysis to examine correlations between the sentiment index and Tokyo Stock Exchange prices. One of our interesting findings is that the index significantly predicts stock prices of three day ahead.

1. はじめに

デフレ脱却へ向けての経済政策が大きな論争となる中，景気動向を動かす目に見えない感情や雰囲気である「センチメント」が注目されている．センチメントは，統計的なデータのように定量化されたものではなく, 本来形のない定性的なものに過ぎない．本研究では，これを目に見える独自の指標として定義することを試みる．具体的には，日本経済新聞の記事の言語解析により，経済状態を表す指標を作成し，これを市場センチメント・インデックスと名付ける．その上で，このインデックスの株価に対する説明力と予測力について実証分析を行う．

このようにセンチメントを定量化する研究は，近年の IT 技術の発達とそれにより処理可能となった情報量の増加により，コンピュータ科学と社会科学の境界領域として発達しつつある．なかでも，本研究と問題意識が最も関連するのは次の 2 つの研究である．Bollen et al. (2012)の研究では，Twitter より 7 つの感情に対応したセンチメント・インデックスが構築されている．ポジティブ／ネガティブの度合いを測る Opinion Finder，および，POMS と呼ばれる 6 つの尺度から被験者の心理状態を測定するアンケート調査方法によって 7 つのインデックスが構築されている．その上で，これが株価を説明するかを分析している．一方，Boudoukh et al. (2012) によれば，ポジティブであるのかネガティブであるのかを考慮したニュースは，ある程度，株価を説明しうることを Roll (1988) の分析手法を用いることによって示している．

以上のような先行研究を踏まえて，本研究では，伝統的な媒体であるが，我が国おいて最も読まれている日刊の経済新聞である日本経済新聞の記事より抽出したセンチメントが株価を説明・予測しうるかを実証分析する．

本論文は以下のように構成される．第 2 節において，日本経済新聞の記事より，市場センチメント・インデックスを構築する方法について説明し，その性質を調べる．第 3 節において，このインデックスが株価を説明・予測しうるか，実証分析を通じて明らかにする．第 4 節においてまとめをする．

2. 市場センチメント・インデックス

2.1 構築の手順

本研究における市場センチメント・インデックスは，以下に詳述する 3 つのステップを経て構築される．

(i)テキスト・クリーニング

離散時点 t において発信された記事 i を $A_{i,t}$ と書き，その総数を M_t とする．これについて分かち書きを行い，$m_{i,t}$ 個の単語 $a_{ij,t}$ が得られたとする．ただし，本論文における単語は，分析を単純化するために，その品詞を名詞，形容詞，動詞に限定する．また，分かち書きは Text Mining Studio 4.2 を用いて行った(数理システム, 2013).

(ii)スコアリング

一方，読者にポジティブ，あるいはネガティブな印象を想起させる単語の集合を事前に用意しておき，これを辞書と呼ぶ．それぞれの辞書を $D^+:=\{d_k^+: k=1, \ldots, K^+\}$, $D^-:=\{d_k^-: k=1, \ldots, K^-\}$ と書く．本論文では，Takamura et al. (2005)や高村 et al. (2006)の成果に基づいた辞書「単語感情極性対応表」を用いる(高村, 2007)．本辞書においては，D^+ と D^- に属する単語数 K^+, K^- はそれぞれ 5,122 個と 49,983 個である．

単語 $a_{ij,t}$ がポジティブ，あるいはネガティブな印象を想起させるときにのみカウントをする，次のような定義関数を導入する．

$I_{ij,t}^+=1$ (if $a_{ij,t}\in D^+$); $I_{ij,t}^+=0$ (otherwise)

$I_{ij,t}^-=1$ (if $a_{ij,t}\in D^-$); $I_{ij,t}^-=0$ (otherwise)

これらを用いて，記事 $A_{i,t}$ に含まれるポジティブあるいはネガティブな印象を想起させる単語数をそれぞれ次のように数える．

$$n_{i,t}^+=\sum_{j=1}^{m_{i,t}} I_{ij,t}^+ \qquad n_{i,t}^-=\sum_{j=1}^{m_{i,t}} I_{ij,t}^-$$

したがって，時点 t において発信された記事の全体では，ポジティブあるいはネガティブな印象を想起させる単語数はそれぞれ

$$N_t^+=\sum_{i=1}^{M_t} n_{i,t}^+ \quad N_t^-=\sum_{i=1}^{M_t} n_{i,t}^-$$

となる．その合計は，$N_t = N_t^+ + N_t^-$ である．

(iii)インデックスの定義

本論文における市場センチメント・インデックスを，時点 t における発信された記事の全体において，ポジティブな印象を想起させる単語が出現した割合として定義する．

$$x_t = N_t^+ / N_t \qquad (1)$$

以下の分析では，便宜上その平均が 0 になるように基準化して用いている．

2.2 データ

本論文で用いた記事は，2007 年 1 月 1 日から 2012 年 9 月 30 日までの 2,094 日間において，日本経済新聞の朝刊に掲載されたすべての記事である．その総数は 1,026,575 であり，月平均 14,878 である．日本経済新聞記事アーカイブについて，日本経済新聞デジタルメディアよりその利用許諾を受けた．その記事は毎日発信されるため，日次という離散時点において，市場センチメント・インデックスを構築し，これを

$x=\{x_t: t=1, \ldots, T\}$ と書く．

一方，同一期間における日経平均株価(以下，株価)の収益率を導入する．これは，日次の時系列上で隣り合った 2 つの株価(調整済終値)の増減率である．これを「株価収益率」と呼び，

$y=\{y_t: t=1, \ldots, T\}$ と書く．

2.3 市場センチメント・インデックスの特徴

図 1 は市場センチメント･インデックス(上段の線)と株価収益率(下段の線)を時系列表示したものである．ただし，前者は 3 日間だけ時点を進めて表示している．この図を見ると，数多くの期間において，両者の挙動が似通っていることが分かる．特に，株価収益率のボラティリティが大きな時期には，市場センチメント・インデックスも大きく変動して，そのタイミングが一致していることが分かる．図中には A, B, C のマーカーを記した．それぞれ，2007 年 8 月欧米金融市場においてサブプライム問題が顕在化した時期，2008 年 9 月リーマンショックにより金融危機が認識された時期，2011 年 3 月東日本大震災が発生した時期である．

ここでの重要な視点は，3 日間だけ時点を進めて表示した市場センチメント・インデックスと株価収益率の挙動が一致する期間があり，特に，大きなニュースに起因して市場が大きく変動する期間にはその挙動が顕著に一致する，ということである．つまり，市場センチメント・インデックスは株価収益率を説明・予測する可能性があることを示唆する．次節では，こうした着想を実際に計量分析してみたい．

3. 実証分析

3.1 モデル

時点 t と $t-1$ で挟まれた時間間隔を期間 t と呼ぶ．期間 t における株価収益率を y_t，期間 t の期首，つまり時点 $t-1$ における市場センチメント・インデックスを x_{t-1} と書く．以下の p 次 VAR モデル(2)を利用

して，市場センチメント・インデックス x により，株価収益率 y が予測であるかどうか分析を行う．

$$y_t = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{i=1}^{p} \gamma_i x_{t-i} + \varepsilon_t \qquad (2)$$

3.2 データセット

本研究では日次のデータを扱っているため，株価収益率と市場センチメント・インデックスの欠損値の頻度が異なる．欠損値について，前者は株式市場休業日に，後者は日本経済新聞休刊日に発生し，両者の観測数も異なっている．これらの欠損値を取り扱う方法には 2 種類ある．

A. 市場休業日の市場センチメントを考慮する処理方法：

データセットに欠損値があったとしても，予めその削除を行わない．モデル(2)の推定に際して，$\{y_t; y_{t-i}, x_{t-i} \quad (i=1, \ldots, p)\}$が揃わないデータについてのみ，欠損値として削除する．

この方法のメリットは，例えば，市場休業日の翌日である月曜日の株価収益率は，月曜日・日曜日・土曜日・金曜日…の市場センチメントによって説明され，土曜日と日曜日という市場休業日における市場センチメント情報が積極的に利用される．

B. 市場休業日の市場センチメントを考慮しない処理方法：

株価収益率が観測されない市場休業日，および市場センチメントが観測されない日本経済新聞の休刊日を，データセットより予め削除する．

メリットしては，処理方法 A と比べて，より高い次数 p を設定可能な点が挙げられる．一方，デメリットは，$p \geq 2$ 以上の場合，例えば，市場休業日の翌日である月曜日の株価収益率は，月曜日・金曜日・木曜日…の市場センチメントによって説明がされることになり，土曜日と日曜日という市場休業日における市場センチメント情報は利用されない．

3.3 分析結果

本研究が対象とするデータ期間は，2008 年 9 月の金融危機を含んでいる．図 1 からも直観的にわかるように，その前後において株価収益率の構造が大きく変化した可能性がある．また，事前にすべてのデータ期間に対してモデル(2)の推定を行った結果，単位根は存在しないものの，市場センチメント・インデックスは，株価収益率を有意に説明しているとは言い難いことが分かった．

そこで，モデル(2)を記述するパラメータ構造がその前後において変化しているのか，とりわけ，市場センチメント・インデックスの株価収益率の予測有意性の変化を検証する．

まず，データ期間を金融危機の前後の 2 つの区間に分け，観測データがどちらのデータ区間に属するかを表すダミー変数を導入したモデル(2)に基づいて Chow 検定を行った（表 1）．その結果，次のことが言える．

市場休業日の市場センチメントを考慮する場合には，ラグ p=3 と p=4 の場合に，帰無仮説が棄却され，有意に構造変化したといえる．特に，ラグが p=4 の場合には 1%有意となっている．一方，市場休業日の市場センチメントを考慮しない場合には，どのラグにおいても，帰無仮説が棄却されず，構造変化があったとは言えない．

そこで，有意な構造変化を示した，市場休業日のセンチメントを考慮する場合について，金融危機の前後のそれぞれの区間において，ラグを p=4 と設定したモデル(2)を推定した．その結果を表 2 に示す．

推定に用いるデータが異なるので，厳密な比較は難しいが，自由度調整済み決定係数を見ると，後半のデータへの適合度が高いことが分かる．また，後半のデータに対しては，パラメータが有意に推定されることが多い．特筆すべき重要な発見は，「株価収益率は，3 日前の市場センチメント・インデックス x_{t-4} と正の相関によって説明される」という点である．

追加して Granger の因果性テストも行ったが，株価収益率の予測において，3 日前の市場センチメント・インデックスが 1%有意な変数となることが確認された（表 3）．

4. おわりに

本研究においては，日本経済新聞から単純な 1 つの市場センチメント・インデックスを構築した．この市場センチメント・インデックスは，2008 年 9 月の金融危機後の期間において，3 日後の株価収益率を有意に説明・予測しうることが分かった．このようなラグが生じる理由として予想されるのは，一つには，感情が伝播するのにはある程度の時間がかかるということ（情動伝染），もう一つには，週末を挟んで金曜日の取扱いが影響を与える場合があること（曜日効果）が挙げられる．こうした予想を本研究で対象とした日次データに加え，より長期的な頻度

の観測データについても検証することは，今後の課題である．

謝辞

本研究発表の機会を与えていただいた金融情報学研究会(SIG-FIN)の方々，また，日本経済政策学会第 70 回全国大会，日本金融・証券計量・工学学会(JAFEE)高頻度データ・ビッグデータ活用法研究部会，第 39 回 2013 年度夏季 JAFEE 大会にて有益なご助言を頂いた参加者の方々に深く感謝したい。

参考文献

[1] 高村大也，乾孝司，奥村学：スピンモデルによる単語の感情極性抽出，情報処理学会論文誌ジャーナル，Vol. 47, No. 2, pp. 627-637, (2006)

[2] 高村大也：単語感情極性対応表，http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html (accessed 2013-04-16), (2007)

[3] Bollen, J., Mao, H. and Zeng, X.: Twitter Mood Predicts the Stock market, J. Computational Science, Vol. 2, No. 1, pp. 1–8, (2011)

[4] Boudoukh, J., Feldman, R., Kogan, S. and Richardson, M.: Which News Moves Stock Prices? A Textual Analysis, NBER Working Paper 18725, (2012).

[5] Takamura, H., Inui, T. and Okumura, M.: Extracting Semantic Orientations of Words using Spin Model, In: Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005), pp. 133-140, (2005).

図表

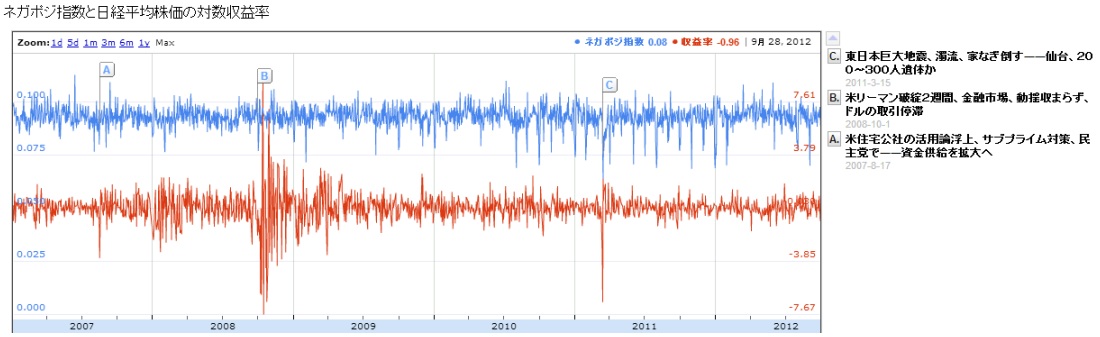

図 1：構築した市場センチメント・インデックス(上段)と株価収益率(下段)．比較のため，前者を 3 日間だけ時点を進めて表示．

表 1:データ期間の前半と後半における構造が変化したかどうかに関する Chow 検定. カッコ内は Newey-West の標準誤差を用いて計算した p 値を表す． **印は 5%有意，***印は 1%有意を表す．

考慮あり		考慮なし	
Lag (p)	検定統計量	Lag (p)	検定統計量
1	1.1759	1	0.5323
	(0.3089)		(0.5874)
2	0.1642	2	0.5154
	(0.9565)		(0.7245)
3	2.1506**	3	1.6479
	(0.0465)		(0.1304)
4	2.8851***	4	1.4376
	(0.0045)		(0.1760)
		5	1.4079
			(0.1708)
		6	1.2175
			(0.2649)
		7	1.1500
			(0.3087)

表 2：市場休業日の市場センチメントを考慮する場合の推定結果．AIC は赤池の情報量規準，Res. Var.は残差分散，R^2 は自由度調整済み決定係数を表す．カッコ内は Newey-West の標準誤差を用いて計算した p 値を表す．*印は 10%有意，***印は 1%有意を表す．

	AIC	Res. Var.	R^2	市場センチメント・インデックス				定数項	対数収益率			
				x_{t-1}	x_{t-2}	x_{t-3}	x_{t-4}	α	y_{t-1}	y_{t-2}	y_{t-3}	y_{t-4}
前半	-414.68	1.16%	-6.12%	-0.0352	0.3374	0.2175	-0.2878	-0.023461	0.2048	-0.0286	-0.1187	-0.1238
				(0.9337)	(0.2961)	(0.4978)	(0.4534)	(0.6120)	(0.1600)	(0.8230)	(0.4324)	(0.3868)
後半	-945.15	1.24%	20.32%	-0.2582	0.0110	-0.2269	0.4542*	-0.000904	-0.3307***	0.3943	-0.1336*	0.0275
				(0.1877)	(0.9489)	(0.4116)	(0.0950)	(0.9689)	(0.0037)	(0.1076)	(0.0981)	(0.8427)

表 3：Granger 因果性テスト

考慮あり

Lag (p)	前半	後半
1	0.2760	0.4746
	(0.5994)	(0.4909)
2	5.0731*	0.1802
	(0.0791)	(0.9138)
3	6.9341*	30.4253***
	(0.0740)	(0.0000)
4	71.0274***	41.2864***
	(0.0000)	(0.0000)

考慮なし

Lag (p)	前半	後半
1	0.0487	0.1268
	(0.8254)	(0.7218)
2	1.9904	0.6603
	(0.3697)	(0.7188)
3	2.1961	23.2067***
	(0.5327)	(0.0000)
4	2.0061	24.3651***
	(0.7346)	(0.0001)
5	3.8761	29.1773***
	(0.5674)	(0.0000)
6	4.3367	30.1460***
	(0.6312)	(0.0000)
7	6.6544	36.5714***
	(0.4657)	(0.0000)