株式市場を反映するセンチメント・インデックスの構築と株価説明力の実証分析*

數 見 拓 朗†

要　約

本論文では，日本経済新聞（日経）から，日本の経済，特に日本の株式市場を反映するセンチメントを計量化する指数（インデックス）について提案する。具体的には，2つのセンチメント指数（SI: sentiment index）を構築し，その日本株価への説明力を実証する。市場における参加者の心理や雰囲気を表すセンチメントに関する分析は，近年，学術研究のみならず，あらゆる産業実務においても，ますます多くの関心を集めている。センチメント分析には多くのアプローチが考えられるが，本研究は，Ishijima, et al.（2014）を議論の起点とする。彼らは，日経に掲載されているすべての記事を利用して，ポジティブあるいはネガティブな心理を，SIとして計量化する一手法を提案した。さらに，過去5年間の日次データを分析対象として，そのSIが3日後の日本株価を有意に予測し得ることを実証した。本研究の目的は，Ishijima, et al.（2014）の拡張として，日本の雰囲気全体ではなく，日本の経済活動のセンチメントに焦点を当てたSIを提案する。具体的には，日本の経済活動について言及している記事に限定してSIを作成する。本研究で得られた結論は以下の3点に集約される。(1) 本研究で提案するSIは，先行研究の結果とは異なり，翌営業日の対数収益率と出来高に有意な説明力を有する。(2) 提案するSIは，株価を説明する上で，リバウンドが観測されない。(3) SIは，株式市場に対して後追いで反応する，ということである。

JEL Classification：C88, E37, G17
キーワード：センチメント分析，日本経済新聞，テキスト・マイニング，株価予測可能性

1　はじめに

近年，学術界と産業界の両方でセンチメント分析が注目されている。センチメントとは，景気全般や社会心理を漠然と表すもので，これを分析することは，経済やマーケットをより深く理解する助けとなる。

資本市場・証券市場などを念頭においたセンチメントを，市場センチメントと呼ぶことにする。その背景となる議論を以下に概観してみよう。近年の市場センチメントについての関心は，市場の合理性の仮定についての賛否を起点にしている。効率性市場仮説によれば，情報は市場全体に極めて効率的に平等に広がり，それゆえ，リスクを勘案して合理的に予想される投資リターン以上の超過リターンを，何人も得ることができないとされる（Fama, 1965, 1991）。

* 本論文を執筆するにあたり，指導教員である大阪大学大西匡光教授に感謝致します。また，多くの知識や示唆を頂いた中央大学石島博教授にも感謝致します。
† 大阪大学大学院経済学研究科博士後期課程，株式会社サイバーエージェント

これまで多くの実証研究がこの仮説を支持する結果を示してきたが，その一方で，近年，これを否定する研究も増えている。こうした研究の多くは，超過リターンの存在（市場アノマリー）を立証し，それを通して証券価格の予測可能性を立証するものとなっている。

一方，市場アノマリーの実証研究を支える理論的根拠についても，いくつかの理論が提示されている。特に，Kahneman and Tversky（1979）に始まるとされる行動経済学は，経済主体の行動を，その心理的側面に力点をおいて，解釈するものとなっている。その脈絡で，行動ファイナンスについてRitter（2003）は文献を整理し，認知心理学と市場裁定機会の可能性を中心にまとめている。それによると，典型的な分析結果は，市場参加者は必ずしも合理的に取引を行うわけではない，というものである。むしろ，彼らの間で広がっている心理状態に従い，非合理的に取引を行う。そして，そうした取引が市場に超過リターンの機会を生じさせる。そのほかにも，経済主体は当該の経済活動に関係の無い情報に影響されて取引を行う，といった理論も提案されている。社会の雰囲気，世論，社会トレンドなどと言った株価に直接関連しそうにない情報が株価を動かすと考えられる。「センチメント」という言葉は，そうした情報に対する認識全般を表すものであるが，これまでは，実体の無いものとして顧みられることも少なかった。

センチメント分析は，これまではっきりとした実体として捉えられることが少なかったもの，主に文章に現れる心理などを数値として定義し，計量しようとする近年の試みである。これはテキスト・マイニング技術の発達によるところが大きい。合理的市場仮説に対する疑問，行動経済学と関連する新しい理論を背景として，こうしたセンチメント分析が社会経済分析の分野で注目され始めたと言える。

近年急速に増えているセンチメント分析の文献のなかで，株式市場に関連しているものとして特記に値するものがいくつか挙げられる。Tetlock（2007）はマスメディアと株式市場との相関を検証した。その延長線上でさらにTetlock et al.（2008）は言語を計量化したものが個々の企業の会計利益と株式リターンを予測しえるか否か分析を行った。Bollen et al.（2011）は，Twitterに現れる文章についてインデックスを作成することを考案し，センチメント・インデックスとしていくつかのタイプを提示した。Boudoukh et al.（2012）は，新聞などのニュースが株価の動きをどのようにリードするか検証した。

日本語のデータを使った研究例も数多く存在する。例えば，沖本 et al.（2014）や五島（2016）は，Tetlock（2007）で提唱された仮説が日本の株式市場でも成り立つかどうかを検証している。彼らの分析では，日経QUICKやロイターニュースのデータから作成したニュース指標が，翌日のリターンや出来高に対して有意な説明力を持つことが示されている。また，インターネット上のデータを利用して，株価を予測する研究例として，坪内 et al.（2015）が挙げられる。彼らの研究では，感情辞書を準備して，Yahoo!ファイナンスの株価掲示板からインデックスを作成し，株価の予測性能を検証している。新聞記事を利用した研究として，Ishijima et al.（2014）は，毎日の新聞ニュースに現れる日本経済のセンチメントを計量的に分析した。彼らは，日々の経済状況を肯定的あるいは否定的に説明する単語の出現頻度をカウントし，指数（インデックス）とした。新聞ニュースとしては日本経済新聞（以下，日経）が使われている。その上で，彼らはそうした指数をセンチメント・インデックス（以下，SI）とよび，株価指標である日経 225 との相互関係を統計分析した。興味深いことに，彼らの結論は，日経に掲載された全ての記事から作成したSIが 3 日先の株価を予測しえるとのことであった。

以上のように，テキスト情報からセンチメントを抽出し，株式市場を予測する研究は多く存

在するが，株式市場を説明するために，どのようなソースからセンチメントを作成すべきかについて，統一的な見解はない。例えば，Twitterのデータを利用したBollen et al.（2012）や，全ての日経記事を利用したIshijima et al.（2014）のように，経済活動に関するニュースを対象にしていない記事を含めてSIを構築する研究事例がある。彼らの研究は，特に世の中の「全体」の雰囲気に着目をして，株式市場の関係を明らかにする研究と言えよう。一方で，沖本 et al.（2014）や五島 et al.（2016）のように，マーケットニュースなど株式市場に影響を与える可能性のあるニュースからSIを構築する研究も多い。特に，沖本 et al.（2014）では，どのような種類のニュースが株価に対して大きな影響を有するかを実証している。彼らの分析では，コーポレートアクションとマーケットコメントに関する記事が株式リターンに有意な説明力を有していることを示している。

以上を踏まえて，本研究の目的は，2007年1月1日から2014年12月31日までの日経の記事を分析対象とし，そのSIを計量化するにあたっては，市場に関連すると思われる記事に限定して作成したSIと，記事を限定せずに全ての記事を用いて作成したSIが，株価の説明力という点において差異があるかどうかを実証することである。具体的には，以下の手順を踏んで実証を行う。第一に，日経記事の内容ごとに分類を行う。日経記事の元データには，どのような内容が書かれているかについて分類がなされていないため，トピックモデルを利用して各記事の分類を行い，経済活動に関連したニュースを特定する。第二に，経済活動に関連したニュースに分類された記事から作成するSI，株式収益率，出来高からなる3変量VARモデルと，全ての記事を利用するSI，株式収益率，出来高からなる3変量VARモデルの，2つのVARモデルを構築する。第三に，VARモデルの推定結果の符号条件を確認する。Tetlock（2007）で指摘された，ニュースの株価への影響に関する3つの理論（情報理論，センチメント理論，無情報理論）のうち，どの理論が成立する可能性が高いかを調査する。第四に，統計的なモデルの当てはまり度合いを確認する。以上の手順を踏むことで，本研究で提案するSIと，Ishijima et al.（2014）のSIのどちらが，株式市場をより反映するSIであるかを検証する。Tetlock（2007）の3つの理論に関しては，以下で説明する。

第一に，情報理論とは，SIの株価に対する影響は，恒久的に消滅しないという理論である。情報理論が成立する場合，SIの株価に与える影響の方向は，一貫して同じ方向を向いているはずである。第二に，センチメント理論とは，SIは短期的に株価に影響を与えるが，その影響は長期的には消滅するという理論である。第三に，無情報理論とは，SIは株価に対して影響を全く与えないという理論である。

本研究の貢献として，以下の3点が挙げられる。第一に，経済活動に関する記事を抽出する際の，記事の分類方法の提案である。日本の株式市場を説明しうるSIの構築に関して，LDAが有効であることを示す。第二に，提案するSIは，翌営業日の対数収益率と出来高を有意に説明し，株式市場に対して，後追いで反応するということである。第三に，提案するSIの株価への推定係数が，一貫して同じ方向であることを示す。これはTetlock（2007）とは異なる結果である。

本研究は次のように構成される。第2節では，分析対象とする日経記事について説明する。第3節では，それぞれの日経記事を書かれている内容をもとにクラスタリングを行い，経済活動に関連したニュースの記事を特定する。第4節では，SIの構築方法について述べる。第5節では，株式市場との相関分析のためのモデルを記す。第6節でまとめとする。

2　日経記事の分析

本節では，SIを構築する上で，対象となるテキスト情報がどのようなニュースソースなのかを理解するために，日経記事について概観する。まず，データの前処理について説明する。次に，日ごとに現れる特徴語の推移を確認しながら，日経記事と社会的イベントの関係について明らかにする。尚，本研究で利用する記事は，2007年1月1日から2014年12月31日の間に掲載された，718,743記事の日経の朝刊である。

2.1. データの前処理

本研究の分析では，以下の記事を分析の対象外とする。第一に，決算数字のみが掲載されている記事である。第二に，会社人事情報のみが掲載されている記事である。上記の記事を除くと，最終的に分析に利用する記事数は，603,063記事になる。

また，ストップワードを設定する。新聞記事の性質上，「～月」や「～日」など日付に関する言葉が多く出現する。こうした言葉は，センチメント分析で有用でないため，分析に利用しないことにする。

2.2. 社会的イベントと日経記事に現れる特徴語

サブプライムローン問題を発端とした世界金融危機や東北大震災などの社会的イベントに対して，日経でどのような話題が掲載されていたのかを確認するために，日ごとに特徴的に現れた単語を抽出する。抽出対象とした単語は全期間を通じて，100回以上出現した名詞に限定する。ただし，数字が特徴的な語として現れることを防ぐために，文書全体で30%以上使用されている名詞は対象外とする。名詞の抽出には，形態素解析ライブラリであるMeCab[1]を使用し，最新の言葉を抽出するために辞書としてmecab-ipadic-neologdを利用する[2]。

次に，特徴語を定義する。ここで，文書dは，ある1日に掲載された日経記事とし，$|D|$は日経が発行される全日数，$|d: d \ni w_i|$は，単語w_iが出現する日数であるとする。ここで，特徴語は，ある日に単語w_iが出現した回数を$N_{w_i,d}$としたときに，

$$tf_{w_i,d} = \frac{N_{w_i,d}}{\sum_k N_{w_k,d}} \quad (1)$$

$$idf_{w_i,d} = log \frac{|D|}{|d: d \ni w_i|} \quad (2)$$

$$tfidf_{w_i,d} = tf_{w_i,d} \cdot idf_{w_i,d} \quad (3)$$

によって求められる$tfidf_{w_i,d}$が高い単語とする。

こうして得られた特徴語を，リーマンブラザーズの破綻などの社会的イベント前後に分けて，表1に示す。表1から，社会的イベントごとにどのような言葉が特徴的であったかを概観する。尚，表1の網掛けになっている日付けは，社会的イベントが発生した日を表している。

(1) イチロー4000本安打達成（2013年8月21日）

アメリカ時間の2013年8月21日（日本時間22日）のブルージェイズ戦で，イチローは日米通算で4000本安打を記録した。時差の関係で，23日からイチローに関する話題が特徴的になっていることがわかる。

(2) リーマンブラザーズの破綻(2008年9月15日)

2008年9月15日は，リーマンブラザーズが破綻し，バンク・オブ・アメリカがメリルリンチの買収に合意した日である。また翌日に

[1] MeCabとは京都大学大学院情報学研究科とNTTコミュニケーション科学基礎研究所によって2013年に開発された形態素解析を行うためのアプリケーションである。詳細は次のサイトを参照のこと。URL http://taku910.github.io/mecab/（アクセス日：2016年8月21日）

[2] Ishijima et al.（2014）では，標準のシステム辞書を利用している。

表１　イベント前後の特徴語の推移

	2008/8/20	2008/8/21	2008/8/22	2008/8/23	2008/8/24
1	カジノ	競争政策	標本	イチロ	イチロ
2	Cluster	疎開	汚染水	汚染水	水中り
3	新電力	暫定政権	カリウム	安井	卵子
4	チャットアプリ	国有企業	lte	日米通算	イチ子
5	うちわ	山本さん	円借款	mrj	待機児童
6	LIXIL	派遣会社	評議員	鉱物	考古学
7	ビッグデータ	汚染水	暫定政権	三菱航空機	同胞
8	代謝	丸太	国保	保釈	長谷工
9	騒乱	同胞	line	中尾	野党再編
10	もんじゅ	his	化学兵器	飼い主	開催都市
	2008/9/13	**2008/9/14**	**2008/9/15**	**2008/9/17**	**2008/9/18**
1	麻生氏	モネ	事故米	リーマン	aig
2	事故米	庵	有料道路	事故米	リーマン
3	アウトレット	麻生氏	捕虜	製菓	事故米
4	アーバン	ユダヤ人	分配金	aig	リーマン・ブラザーズ
5	献金	産科	リーマン	リーマン・ブラザーズ	落雷
6	与謝野	検針	共青団	メリル	つゆ
7	小池	ネスレ	内部統制	バンカメ	裁定
8	公開討論会	老い	艦	堂	サムライ債
9	石破	貧血	総合病院	ベアー・スターンズ	MBO
10	資金管理団体	リーマン	秋田県	特別養護老人ホーム	シダックス
	2009/9/14	**2009/9/15**	**2009/9/16**	**2009/9/17**	**2009/9/18**
1	デルタ	車検	入閣	鳩山首相	商業地
2	減反	デルタ	亀井	鳩山内閣	アウトレット
3	宇宙ステーション	カシオ	郵政	衆	鳩山内閣
4	蚊	閣僚人事	環境税	鳩山政権	住宅地
5	正答	イチロ	飲酒運転	藤井裕久	地価
6	インサイダー取引	雅山	三好	政治主導	チーズ
7	有人	スプレッド	増派	水量	hiv
8	ドッキング	水族館	モリブデン	脱・官僚	MD
9	植物工場	新型インフルエンザ	郵政事業	横顔	サンマ
10	西川	ネットスーパ	カブ	節水	用船
	2011/3/9	**2011/3/10**	**2011/3/11**	**2011/3/12**	**2011/3/13**
1	カバン	日立物流	SWF	津波警報	炉心
2	ソバ	活字	古紙	震源	炉心溶融
3	林業	リビア	リビア	大津波	格納
4	リビア	監察	租	余震	冷却水
5	世界観	sap	信組	帰宅困難者	カレンダ
6	未納	音質	中国産	災害対策本部	灯火
7	フック	録音	証取	水没	避難所
8	緒方	上場会社	土木	巨大地震	海水
9	法大	紙おむつ	明細書	仙台空港	チェルノブイリ
10	ダイバーシティ	農業委員会	ニンジン	プレート	停電
	2013/4/2	**2013/4/3**	**2013/4/4**	**2013/4/5**	**2013/4/6**
1	長嶋	isa	完全試合	マネタリーベース	サーベラス
2	モリブデン	サーベラス	alsok	黒田	以南
3	松井氏	黒鉛	月齢	物価目標	普天間
4	産業政策	roe	地熱発電	禁	文化財
5	民主派	教育長	イー・アクセス	黒田東彦	イチ子
6	以南	九電	ダルビッシュ	量的・質的金融緩和	ダイエ
7	紀州	教育行政	1seg	当座預金	辺野古
8	北越	発送電分離	八百長	roe	roe
9	焼き鳥	貸し切り	ヒューストン	大王	n
10	再選挙	返戻	選抜高校野球	サーベラス	嘉手納基地

はAIG株が急落した。そのため，17日以降は，リーマンブラザーズ・AIG・バンカメ・メリルなど，一連の金融危機によって，損害を受けた金融機関が特徴的な言葉として現れている。

(3) 民主党政権の発足（2009年9月16日）

民主党政権が発足した2009年9月16日以降のニュースは，「脱・官僚」や「政治主導」政権のスローガンや，政権のキーマンに関する言葉が特徴的になっている。

(4) 東北大震災（2011年3月11日）

地震が起こった次の日の記事は，前日の地震が大きな地震であったことや，津波により大きな被害を与えたこと，東京などで帰宅困難者が発生したことなどを報道する記事が多いことがわかる。しかし，地震発生から2日後には，炉心溶融が起こったことを伝える記事が増え，原発の停止により計画停電が実施されることが伝えられる記事が多くなっていることがわかる。

(5) 異次元の金融緩和（2013年4月4日）

就任直後の日銀総裁の黒田総裁は，「2年程度で，2%の物価上昇」という目標をアナウンスした。そのため，翌日の記事では，「マネタリーベース」や「物価目標」など，金融政策でよく利用される言葉が特徴的な語として現れている。

上記のように，他の新聞同様，日経でも様々な記事が記載されるので，新聞記事と株式市場の関係を分析する際は，記事の分類を行い，企業の経済活動に関するニュースなど，株式市場に影響を与える可能性のある記事を特定することが重要である。

3　日経記事の分類

Ishijima et al.（2014）では，分析対象時期の全ての日経記事を利用して，SIを作成し，株価との関係について考察しているが，スポーツ情報など，金融市場には関係のないニュースが含まれている。そこで，本研究では，LDAと呼ばれる文書クラスタリング技術を利用して，ある日経記事が，経済，政治，スポーツなど，どのようなことを話題にしているか特定する。文書クラスタリングの結果，経済活動について言及している日経記事を特定し，SIを作成する。

3.1. 潜在的ディリクレ配分法（LDA）

LDA（Latent Dirichlet Allocation）はトピックモデルの一種で，ある文書に含まれる単語がどのようなトピックによって生成されたかを知ることできる。LDAでは，各文書はK個の観測できないトピックから発生した単語で構成されている，と仮定[3]する。単語vを含んだ文書集合Dと，トピック数Kを入力として，以下の2つの確率分布を推定することになる。第一に，トピックkにおける単語vの出現する確率の推定である。第二に，文書dにおけるトピックkの出現する確率を推定することである。トピックごとに出現確率の高い単語を確認すれば，そのトピックがどのようなトピックであるかを分析することができる。

3.2. クラスタリング結果

前節の特徴語抽出で使用した603,063記事を用いて，クラスタリングを行う。ここで，トピック数は30個とする。表2には，トピック番号と，その各トピックで出現する確率の高い上

[3] 例えば，「東京オリンピックによる経済効果は○○億円と推定される」という記事は「スポーツ」と「経済」という2つのトピックから構成されていると考えられる。

位5個の単語を示す。表2の単語の出現情報から，トピック1は生産量や輸出量などが変化したことについて言及しており，トピック21は，金融政策に関する話題である可能性がある。一方，トピック16・22・30はスポーツに関する話題である。単語の出現確率から，経済や市場に関連すると考えられるトピック番号と単語に影を付けている。

4　SIの作成と社会的イベントとの関連

本節では，SIの作成について述べる。本研究では，2種類のSIを作成する。第一に，経済について話題にしている記事から作成するSIである。具体的には，経済や市場に関連するトピック$k=\{1,2,3,11,18,19,21,23,25,26\}$の出現する確率が，正の値をとる記事，461,055件を抽出し，SIを作成する。第二に，比較対象として，全ての記事から作成するSIを作成する。

表2　トピック番号と，各トピックに出現する確率が高い上位5個の語

トピック番号	単語				
1	前年	比	増	減	回復
2	億	期	純利益	増	売上高
3	工場	生産	中国	インド	事業
4	店	店舗	商品	出店	販売
5	氏	大統領	選挙	政権	投票
6	北朝鮮	中国	米	ロシア	会談
7	研究	細胞	薬	治療	患者
8	感染	台湾	卸売市場	元代表	ウイルス
9	容疑者	逮捕	容疑	人	捜査
10	首相	民主党	自民党	氏	党
11	株	東証	貸し借り	銘柄	指定
12	原発	東電	福島	稼働	電力
13	配	予	発売	newface	販売
14	スマホ	通信	装置	半導体	パネル
15	さん	私	人	著	作品
16	位	ゴルフ	アンダ	番	打
17	空売り	人事	課長	比率	維新
18	価格	トン	値上げ	原料	輸入
19	数表	先物	原油	相場	残高
20	社長	氏	取締役	就任	入社
21	日銀	金利	国債	応札	物
22	試合	監督	回	戦	チーム
23	発行	格付け	銀	運用	融資
24	サービス	ネット	サイト	情報	企業
25	企業	経済	政府	制度	改革
26	ドル	安	高	銭	上昇
27	空港	日航	便	ホテル	航空
28	死去	歳	さん	人	氏
29	週	馬	特許	gi	公社債
30	位	女子	男子	五輪	大会

4.1. SIの作成

各記事に現れる単語からセンチメントを計量化する方法について述べる。Ishijima et al.（2014）と同様に，高村（2007）による「単語感情極性対応表」というセンチメント辞書を利用する。この辞書を$\mathcal{D} \coloneqq \{(D_k, S(D_k))|k = 1 \cdots K\}$と書く。つまり本辞書は，単語$D_k$とそのセンチメント・スコア$S(D_k)$の組より構成される。このセンチメント・スコアは，登録された単語が，どれだけポジティブ，あるいはネガティブな心理を日本人に想起させるか，という度合いを計量化したものであり，-1から$+1$までの定義域を持つ。すなわち，$-1(+1)$に近づくほど，日本人に，よりネガティブ（ポジティブ）な心理を想起させる度合いが強くなる。ポジティブなスコア$(S(D_k) > 0)$を持つ登録単語数は5,122である一方で，ネガティブなスコア$(S(D_k) < 0)$を持つ登録単語数はその約10倍の49,983である。

単語感情極性対応表に記載されている単語で，スコアが0よりも小さい語をネガティブな単語とし，0よりも大きい語をポジティブな単語とする。発行日tにおいて配信された日経新聞記事における，ネガティブな単語数をN_tとし，ポジティブな単語数をP_tとする。これを以下のように集計したものを日次のSIとする。

$$SI_t = \frac{-N_t}{P_t + N_t + 1} \tag{4}$$

但し，分母に1を足すのは，ポジティブな単語とネガティブな単語が出現しないときに，定義できなくなることを防ぐためである。

4.2. SIの作成

作成したSIと，社会的イベントの関係を明らかにするために，SIの5日移動平均線を図1に示す。期間中で，SIが最も小さくなるのは，2011年3月16日である。東日本大震災から5日後であり，津波による被害，福島原発の炉心溶解，東京電力の計画停電など，被害の全体像が明らかになった時期である。一方，SIが最も大きくなるのは2014年1月6日である。この日は大発会であり，日経平均株価の終値が前年末比382円43銭安の1万5908円88銭の大幅安となった。しかし，この日でSIが最高値となっているのは，昨年の大納会までの9日間で日経平均株価は9日連続で上昇しており，大発会への期待が高かったためであると考えられる。

5　日本株式市場における実証分析

本節では，前節で作成したSIをもとに分析を行う。具体的には，センチメント・インデックス（SI_t），日経225の対数収益率（$N225_t$），日経225の構成銘柄の出来高の合計の対数値（Vol_t）からなる3変量VARモデルを考えて，Tetlock（2007）と沖本 et al.（2014）で検証され

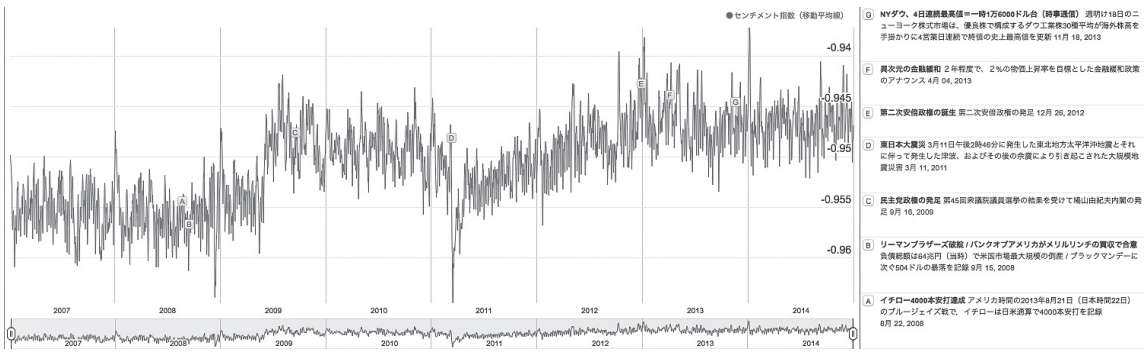

図1　SIの5日移動平均線と社会的イベント

ている，SIの株価に与える影響について考察する。Ishijima et al.（2014）では，出来高について考慮していないが，本研究では，Tetlock（2007）と沖本 et al.（2014）に準じて，出来高を考察対象に含める。また，SIは，全期間で平均0，分散1になるように標準化を行う。

5.1. SIの対数収益率への影響

SIの対数収益率への影響を調査するために，3変量VARモデルのうち，日経225の対数収益率（N225）に関する回帰モデル：

$$N225_t = \alpha_1 + \sum_{j=1}^{p} \beta_{1j} N225_{t-j} + \sum_{j=1}^{p} \gamma_{1j} SI_{t-j} + \sum_{j=1}^{p} \delta_{1j} Vol_{t-j} + \varepsilon_{1t} \quad (5)$$

について，$SI_{\text{t-1}}$の係数γ_{1j}を確認する。pはVARモデルのラグ次数を表し，最大のラグ次数を5として，AICが最も低くなるような次数を選択する。AICが最も低くなった次数は5であった。推定結果をまとめたものが，表3である。ここでカッコ内はNewey-West標準誤差を用いて計算したt値を報告している。また，*，**，***は，それぞれ10％，5％，1％の有意水準で有意であることを表す。これらのことは，以下の表4と表5でも同じとする。

LDAを用いて，経済活動に関連する記事のみを利用する場合，$SI_{\text{t-1}}$の係数と$SI_{\text{t-3}}$の係数が，有意水準10％で有意に正になっており，有意に負になっている他の係数はない。この結果は，沖本 et al.（2014）の結果と整合的である一方，Tetlock（2007）とは異なる結果である。沖本 et al.（2014）が指摘するように，経済活動に関するニュースが株価に対して本源的な情報を保有しているという情報理論が成立している可能性がある。

一方，全ての記事を利用する場合，4営業日前の$SI_{\text{t-4}}$の係数が負で有意になっており，Tetlock（2007）で指摘されているリバウンドが観測される。この結果は，Tetlock（2007）の結果と整合的であり，SIの株価への影響は一時的なものである，というセンチメント理論が成立している可能性があることを示している。

記事を限定して作成したSIの推定結果に注目すると，有意な係数は，全て正である。一方，$SI_{\text{t-4}}$の係数と$SI_{\text{t-5}}$の係数は負になっているが，有意ではない。つまり，どの時点においても，ポジティブ度合いが増せば，対数収益率が上昇することを意味している。これは，記事を経済ニュースのみに限定しているので，株式市場に影響のあるSIを作成できているからだと考えられる。

一方，全ての記事を利用して作成したSIでは，対数収益率への影響が正と負で逆転してしまう。これは，SIが，社会面で掲載される殺人事件のように，株式市場とは関係のない記事を含んでしまっているからだと考えられる。

さらに，経済活動に関連する記事から作成したSIのモデルの方が，全ての記事を利用して作成したSIのモデルよりも，調整済み決定係数がわずかであるが，高くなっている。このこ

表3　SIの対数収益率への影響に関する推定結果

$SI_{\text{t-j}}$	LDAによって記事を限定した場合		全ての記事を利用した場合	
$SI_{\text{t-1}}$	0.00091*	(1.76056)	0.00088*	(1.70551)
$SI_{\text{t-2}}$	0.00029	(0.59390)	0.00005	(0.10246)
$SI_{\text{t-3}}$	0.00083*	(1.65662)	0.00091*	(1.81535)
$SI_{\text{t-4}}$	-0.00073	(-1.50278)	-0.00081*	(-1.64917)
$SI_{\text{t-5}}$	-0.00019	(-0.40743)	-0.00017	(-0.35028)
Adj. R	0.00531		0.00490	

とから，株価を説明するSIの作成には，経済活動には関係のない記事を除いて作成した方がよいと考えられる。

5.2. 対数収益率のSIへの影響

次に，対数収益率のSIへの影響を考察する。このことを考察するために，3変量VARモデルのうち，センチメント・インデックス（SI_t）の回帰式（6）に注目する。上記の分析同様，pはVARモデルのラグ次数を表し，最大のラグ次数を5として，AICが最も低くなるような次数を選択する。AICが最も低くなった次数は5であった。

$$SI_t = \alpha_2 + \sum_{j=1}^{p} \beta_{2j} N225_{t-j} + \sum_{j=1}^{p} \gamma_{2j} SI_{t-j} + \sum_{j=1}^{p} \delta_{2j} Vol_{t-j} + \varepsilon_{2t} \quad (6)$$

もし，新聞記事の中に，株価に関する記述があれば，$N225_{t-j}$の係数が正か負かのいずれかで有意になるはずである。推定結果を表4に示す。

記事を限定して作成したSIと，全記事から作成したSIのどちらの場合も，$N225_{t-2}$の係数と$N225_{t-3}$の係数が正で有意である。これは，作成したSIに，2・3営業日前の，日本の株式市場の状況を表現していることを示している。つまり，作成した2種類のSIは，株式市場に後追いで反応している，ということである。

また，SIとして，記事を限定して作成したSIを採用した方が，調整済み決定係数が高くなっており，モデルの当てはまりが良くなる。

しかし，なぜSIが前営業日の対数収益率に反応していないのかは，今後の研究課題である。

5.3. SIの出来高への影響

最後に，SIの出来高への影響を考察する。Delong et al.（1990）やCampbell et al.（1993）に基づき，Tetlock（2007）は，対数収益率・SI・出来高の3変量VARモデルだけではなく，SIの絶対値を含めた4変量VARモデルを考えている。本研究も，それに倣い，SIと出来高の関係に注目するために，4変量VARモデルの下記の回帰式（7）を考える。ここで，SI_{t-j}の係数と$|SI|_{t-j}$の係数が重要である。

$$Vol_t = \alpha_3 + \sum_{j=1}^{p} \beta_{3j} N225_{t-j} + \sum_{j=1}^{p} \gamma_{3j} SI_{t-j} + \sum_{j=1}^{p} \psi_{3j} |SI|_{t-j} + \sum_{j=1}^{p} \delta_{3j} Vol_{t-j} + \varepsilon_{3t} \quad (7)$$

上記までの分析同様，pはVARモデルのラグ次数を表す。ラグ次数を1から5まで変化させてみて，AICが最も低くなった次数5を選択した。推定結果を表5に示す。

前営業日のSIの出来高への影響を表わすSI_{t-1}の係数に注目する。記事を限定して作成したSIのモデルでは，SI_{t-1}の係数が正で有意になっているが，全ての記事を利用して作成したSIのモデルでは，有意になっていない。つまり，この結果は，前営業日の経済活動に関するポジティブなニュースは出来高を増加させ，経済活動に関するネガティブなニュースは出来高を減少させることを示している。一方，前営業日のSIの絶対値の出来高への影響を表す$|SI|_{t-1}$

表4　対数収益率のSIへの影響に関する推定結果

$N225_{t-j}$	LDAによって記事を限定した場合		全ての記事を利用した場合	
$N225_{t-1}$	1.57024	(1.36680)	1.43639	(1.70551)
$N225_{t-2}$	2.49047**	(2.12557)	2.60836**	(0.10246)
$N225_{t-3}$	2.20995**	(2.19445)	2.60992***	(1.81535)
$N225_{t-4}$	0.47296	(0.42204)	0.96105	(-1.64917)
$N225_{t-5}$	0.63010	(0.57654)	0.26785	(-0.35028)
Adj. R	0.36680		0.36570	

表５　SIの出来高への影響に関する推定結果

SI_{t-j}	LDAによって記事を限定した場合		全ての記事を利用した場合	
SI_{t-1}	0.01010**	(2.00535)	0.00730	(1.45662)
SI_{t-2}	0.01287***	(2.64606)	0.01057**	(2.12214)
SI_{t-3}	0.00385	(0.80720)	0.00416	(0.81910)
SI_{t-4}	0.02001***	(3.62202)	0.02411***	(4.39540)
SI_{t-5}	-0.02834***	(-5.54388)	-0.02459***	(-4.85821)
$\|SI\|_{t-j}$	LDAによって記事を限定した場合		全ての記事を利用した場合	
$\|SI\|_{t-1}$	0.00493	(0.69298)	0.00454	(0.61831)
$\|SI\|_{t-2}$	0.00123	(0.17111)	0.00819	(1.08125)
$\|SI\|_{t-3}$	0.01067	(1.54344)	0.01885**	(2.55177)
$\|SI\|_{t-4}$	0.00695	(0.83503)	0.00438	(0.52270)
$\|SI\|_{t-5}$	-0.01180	(-1.52057)	-0.01462*	(-1.89555)
Adj. R	0.63690		0.63680	

の係数は，SI_{t-1}の結果と異なり，両方のSIのモデルで，有意な結果とはならなかった。

また，記事を限定したSIを変数として採用する方が調整済み決定係数は高くなり，モデルの当てはまりが良くなる。

6　結論

本研究では，日経記事を書かれている内容ごとに分類し，経済活動に関する記事より計量された，ポジティブさとネガティブさを反映した日次SIを作成した。その日次SIと株式市場との関連についてVARモデルにより実証分析を行ったところ，次のような知見を得た。日経記事の，経済や市場を話題にした記事から作成したSIは，（1）翌日の対数収益率と出来高を有意に説明し，（2）株式市場に対して，後追いで反応するということである。

7　参考文献

Asur, S. and Huberman, B.A. (2010) “Predicting the Future with Social Media,” http://www.arxiv.org arXiv:1003.5699v1.

Bollen, J., Mao, H. and Zeng, X. (2011) “Twitter Mood Predicts the Stock Market,” *Journal of Computational Science*, 2(1), 1-8.

Boudoukh, J., Feldman, R., Kogan, S. and Richardson, M. (2012) “Which News Moves Stock Prices? A Textual Analysis,” *NBER Working Paper*, No. w18725.

Campbell, J. Y., Grossman, S. J., and Wang, J. (1992) “Trading volume and serial correlation in stock returns,” *National Bureau of Economic Research*, No. w4193.

De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J. (1990) “Noise trader risk in financial markets,” *Journal of political Economy*, 703-738.

Fama, E.F. (1965) “The Behavior of Stock-market Prices,” *The Journal of Business*, 38(1), 34-105.

Fama, E.F. (1991) “Efficient Capital Markets: II,” *Journal of Finance*, 46(5), 1575-1617.

Gruhl, D., Guha, R., Kumar, R., Novak, J. and

Tomkins, A. (2005) “The Predictive Power of Online Chatter,” *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, New York: ACM Press, 78-87.

Ishijima, H., Kazumi, T. and Maeda, A. (2015) “Sentiment Analysis for the Japanese Stock Market,” *Global Business and Economics Review*, 17(3), 237-255.

Kahneman, D. and Tversky, A. (1979) “Prospect Theory: An Analysis of Decision under Risk,” *Econometrica*, 47(2), 263-291.

Liu, Y., Huang, X., An, A. and Yu, X. (2007) “ARSA: A Sentiment-aware Model for Predicting Sales Performance using Blogs,” *SIGIR ‘07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York: ACM, 607-614.

Loughran, T. and McDonald, B. (2011) “When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks,” *Journal of Finance*, 66(1), 35-65

Mishne, G. and Glance, N. (2006) “Predicting Movie Sales from Blogger Sentiment,” *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*.

Ritter, J.R. (2003) “Behavioral Finance,” *Pacific-Basin Finance Journal*, 11(4), 429-437.

Schumaker, R.P. and Chen, H. (2009) “Textual Analysis of Stock Market Prediction using Breaking Financial News: The AZFin text system,” *ACM Transactions on Information Systems*, 27(2), 12:1-12:19.

Tetlock, P.C. (2007) “Giving Content to Investor Sentiment: The Role of Media in the Stock Market,” *Journal of Finance*, 62(3), 1139-1168.

Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S. (2008) “More than Words: Quantifying Language to Measure Firms’ Fundamentals,” *Journal of Finance*, 63(3), 1437–1467.

石島博・數見拓朗・前田章（2015）「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：日次データによる検証」『経済政策ジャーナル』11 巻 2 号，7–10 頁.

沖本竜義・平澤英司（2014）「ニュース指標による株式市場の予測可能性」『証券アナリストジャーナル』52 巻 4 号，67–75 頁.

五島圭一・高橋大志（2016）「ニュースと株価に関する実証分析：ディープラーニングによるニュース記事の評判分析」『証券アナリストジャーナル』54 巻 3 号，76–86 頁.

高村大也（2007），単語感情極性対応表，http://www.lr.pi.titech.ac.jp/~takamura/pubs/pn_ja.dic（accessed 2015-08-01）.

坪内孝太・山下達雄（2015）「株価掲示板情報の感情解析と株価との相関の研究」『2015 年度人工知能学会全国大会講演集』1J5-OS-13b-2in.

Quantifying Market Sentiment for the Empirical Analysis in the Japanese Stock Market

Takuro Kazumi

The purpose of this paper is to quantify the market sentiment as two indexes and examine whether they can help predict stock prices in the Japanese market. Sentiment analysis is gaining increasing interest in both academia and business. Along these lines, Ishijima et al. (2014) created a sentiment index that quantifies the positive or negative emotion that might appear in entire articles of Nikkei which is the most popular business newspaper in Japan. They concluded that the sentiment index significantly predicts stock prices three days in advance. We re-examine their results by suggesting a new sentiment index quantified from the articles limited to the economic-activity-related news and explore the implication on how the sentiment index can help explain Japanese stock price. Our findings are three-fold: (i) Sentiment index created from the articles limited to the economic-activity-related news significantly allows us to explain Nikkei 225 and the market trading volume of the next business day. (ii) We cannot observe the return reversal referred in the literature. (iii) SI will follow the stock price.

JEL Classification: C88, E37, G17
Keywords: Sentiment Analysis, Nikkei, Text Mining, Predictability of Stock Price