ネガポジ解析によるWebデータと株価変動の相関関係評価

佐藤謙太 *　　小高知宏 *　黒岩丈介 **　　白井治彦 ***

Study on the Correlation between Stock Price and Web Data Using Negative-Positive Analysis

Kenta SATO*, Tomohiro ODAKA*, Jousuke KUROIWA, Haruhiko SHIRAI*****

(Received February 6, 2015)

In this study,we investigate the correlation between stock price and Web data.In previous research,it had been reported that there are some relation between the tweets of Twitter.Analysis of the Web data has shown that it is useful for the prediction of stock price from the result. We examine whether approach of previous research is useful for Web data in Japan.We also researched the correlation between stock price and Web data.We applay the negative-positive analysis to news articles of domestic news site.

We quantified the article of domestic news sites to negative or positive of the two types of value emotional value,and analyzed the relationship between stock price.From the experimental results,we found that there is a certain tendency but it is difficult to show the correlation using only negative-positive analysis.

Key words **:** Big Data, Text Mining, Negative-Positive Analysis, Emotional Value, Stock Price,

1. はじめに

本研究ではテキストマイニング手法の一つであるネガポジ解析を用いて, インターネット上のデータと株価変動との間に相関関係があるかどうかを探ることを目的とする.

本研究ではインターネット上のニュース記事を解析対象とする. 近年ではビッグデータが注目を集めている. インターネット上のデータは日々肥大化しており, 今日では巨大なデータリポジトリと言うことが出来る. ビッグーデータを解析して特徴や傾向の発見や新たな市場の創造などが期待されている. 株価推定の分野でも 2000 年以降からインターネットを介して取得できるデータ (SNS やニュース記事) を予測に用いるという研究が多くなされてきた.[1]–[4] その中でも 2010 年に報告された Twitter の感情解析による株価推定の研究が注目を集めている.[5] この先行研究では Twitter のツイートに対して感情解析を行いインターネット上のデータを感情値として数値化を行った. その結果 86%を超える予測精度を出したことに加えて, インターネット上のデータの感情値と株価変動との間に何らかの相関関係があることが示唆された.

株価推定の際にはテクニカル分析やファンダメンタル分析が主に用いられる. 学術分野では階層型ニューラルネットワークやベイジアンネットワークを用いた手法などが提案されている. これらの従来の手法は主に過去の株価や出来高, 企業資産等の定量データのみを扱っていた. これに対して近年提案されているビッグデータを用いた解析手法ではインターネット上のテキスト

*原子力・エネルギー安全工学専攻
**知能システム工学専攻
*** 工学部技術部
*Nuclear Power and Energy Safety Engineering Course,Graduate School of Engineering
**Human and Artificial Intelligent Systems Course,Graduate School of Engineering
***Dept. of Technical. Dept. Engineering

データを一とした非定量のデータを扱うことができる. 先に例に示した研究成果からこれらのデータを用いた株価推定の手法が有用であることが示された. 従来の手法データに加えて新たな種類のデータを推定の材料として使用することが出来るため, 今後さらなる予測精度の向上が期待されている.

先行研究で用いられている Web データは使用されている言語が英語である. ビッグデータを用いた株価推定やテキストデータの感情分析は日本国内でも盛んに行われているが, 英語と日本語では文法構造や単語などが大きく異なる. 加えて先行研究の Twitter をのツイートを用いた点においても日本とは使用しているユーザー層などが大きく異なることが予想される. ニュース記事を初めとする Web データを用いた株価推定手法はいくつか提案されているが, 日本語で構成されている Web データと株価変動の関係について迫った研究報告は少ない.

そこで本研究では国内のインターネット上のニュース記事との間に何らかの相関関係が見出せるのかを調査した. ニュース記事を対象にポジティブかネガティブかを判断するための感情値を抽出するネガポジ解析を行い, 実際の株価変動と照らし合わせて関係があるかどうかを探っていく.

本稿では 2 章では手法をはじめとする株価推定の現状について触れる.3 章ではそれらを踏まえて本研究で行う解析の流れ, 手法について述べる.4 章では行う実験の詳細について述べる.5 章で実験結果の評価並びに考察を述べる.6 章でまとめを述べる.

2. 現在の株価推定手法とその現状

2.1 推定手法の分類

図 1 に現在用いられている株価推定手法を分類したものを示す.

株価推定の手法は大きく確率論の立場に基づいた手法と決定論の立場に基づいた手法に分けられる. 確率論の手法が用いられる分野は金融工学や金融経済学があげられ, 線形分析法と非線形分析法などを用いた研究が挙げられる.[6] 決定論に基づいた手法ではファンダメンタル分析とテクニカル分析に分けることができる. テクニカル分析は過去の株価や出来高を用いて将来の株価を予測する手法である. 過去の株価の推移から傾向や特徴などのパターンを分析し, 株価の予測を行う. そのパターンとして, 現在では主に 11 種類のコンティニュエイションパターンが類型化されている. その例を図 2 に示す. この手法は短期的な予測に向いている手法である.

一方ファンダメンタル分析は企業の資産などの経済

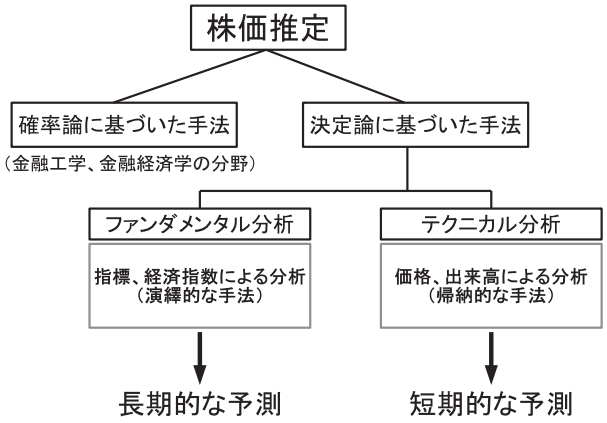

図 1 株価推定手法の分類

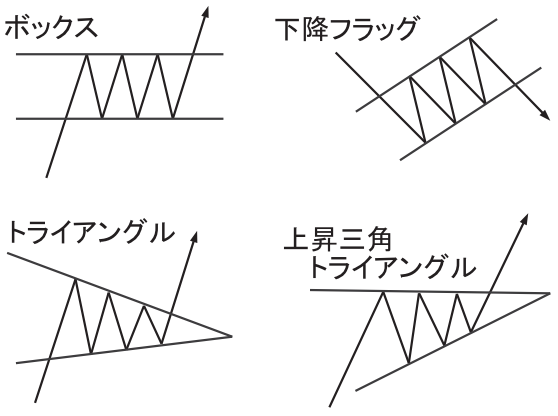

図 2 コンティニュエイションパターンの一例

指数を用いて予測を行う. 分析に比べて扱えるデータの量が多いため様々な角度からの分析を行うことが可能である.[12] 扱うデータの種類, 量が多いため収集に時間がかかることから短期的な予測には向かず, 長期的な予測や市場の将来性を分析する際に用いられる. この二つの手法はそれぞれに向き不向きが存在するため, 実際の分析には両者をうまく組み合わせて行うことが主である.

また, 上記の二つの手法の他にも階層型ニューラルネットワークを用いた手法や, ベイジアンネットワークを用いた手法もある.[7],[8]

近年ではビッグデータが注目を集めるようになりインターネット上のデータを株価推定に用いる手法も多く報告されている.

2.2 ビッグデータ解析を用いた研究

ビッグデータ解析が様々な分野で用いられるに伴い, 株価推定の研究でもビッグデータ解析を用いた研究が多くなされるようになった. 従来の予測では扱うのは定量データのみであった為, 定量化されていないデータなどは一切考慮されていなかった. しかしながら, 株価の変動は過去の株価変動や企業の資産など以外にも変動に関係する要因が多数あるため, 定量化されていないデータを推定の手段に用いることは精度の向上に貢献する余地が十分に考えられる.

ビッグデータの中でも予測に用いられる主なデータは SNS やニュースサイトをはじめとしたテキストデータである. このデータは従来の手法で用いられた定量データより多くの情報が含まれていることが多い. このデータに含まれる情報をうまく数値化することで株価推定に応用が可能である. その手法として行われているのが感情分析である. 感情分析とはテキストデータに含まれる感情の度合いを数値として算出する解析手法のことを指す. いくつかの先行研究でもこの手法が用いられている.

2.3 Web データと株価変動の相関関係分析の試み

Twitter のツイートを用いた予測研究で高い予測精度を出す報告がされたのは既に述べた通りである. 従来までの株価推定の手法は決定論的な手法を用いているため, すべての株価変動を完璧に予測することは非常に困難であるが, 様々な種類の情報が含まれているインターネット上のテキストデータを用いれば精度の向上に寄与出来ることが示された. このことから Web 上のデータを用いる手法が株価推定の手段として有用であることが示されたと言える.

この研究ではこのような高い精度を出すことは出来たが, なぜここまで高い精度になるのかその理由が解明されていない.Web データの感情値と変動予測の結果のつながりが解明されていない為, 今後の精度の向上に向けた明確な手段を出すのは難しいのではないかと考えられる. その為インターネット上のデータと株価変動の間に相関関係の有無を調べ, 相関がある場合はどのような関係や特徴があるのかを調べることが必要だと考えた.

また, 先行研究は解析対象のデータに用いられている言語が異なる. 先行研究で対象となったデータの言語は英語であり日本語とは単語や文法構造が異なる. そのため先行研究と同様のデータや手法を用いて同様の結果を得られるかどうか, 日本国内の Web データが株価推定の手法として有効であるか, 日本語で構成されたテキストデータは株価推定の材料として有効でああるかどうか, などの点を調べる必要があると考えた.

以上の点を踏まえて, 次の章から本研究で行う方法を述べていく.

3. 解析手法

3.1 解析の目的

本研究は本国のインターネット上のテキストデータと株価変動との間の相関性を調べることが目的である. 先行研究では実験が行われた国が異なるために言語仕様や言語解析に用いる手法, インターネットの利用者層並びに利用目的, 思想などが異なる. そのため先行研究と同じ着眼点で解析を行った際に同様の結果が得られるかどうかをまず確認する必要がある. そこでインターネット上にある日本語のテキストデータを解析して相関が得られるかを確認する.

3.1.1 本研究で扱うデータ

ここで本研究で扱うデータを示す. 本研究で行う解析扱うデータは以下である.

- 日経平均株価
- 日本語ニュースサイトのニュース記事

今回は Web 上のテキストデータと日経平均株価との相関を探る. 日経平均株価は YAHOO ファイナンス (http://finance.yahoo.co.jp/) より取得することとする. ニュースサイトについてはその日更新されたトップニュースを全て取得していく. 取得する対象とするニュースサイトは国内の全国紙 5 社の日経新聞, 朝日新聞, 毎日新聞, 読売新聞, 産経新聞とし, 取得するデータ内容は記事

のタイトル, 記事の本文, 記事が公開された時刻をまとめてひとつのデータとして扱う (図 3).

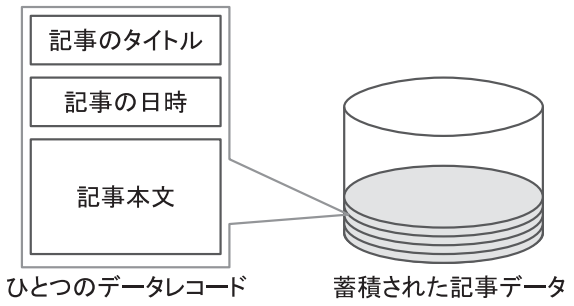

図 3 取得した記事データのフォーマット

本論文では, ニュース記事のデータは 3ヶ月分をあらかじめ専用のスクリプトでニュースサイトから取得し, データベースに蓄積させている. この後説明する記事データの加工や数値化の処理はデータベースから記事データを取り出して行うものとする.

3.2 解析の流れ

本研究で行う解析の流れを以下のようになる. 具体的な流れを図 4 に示す. 各工程の具体的な説明はそれぞれ項を設けて後に説明していく.

1. 記事データの加工
2. データの抽出・数値化
3. データの分析

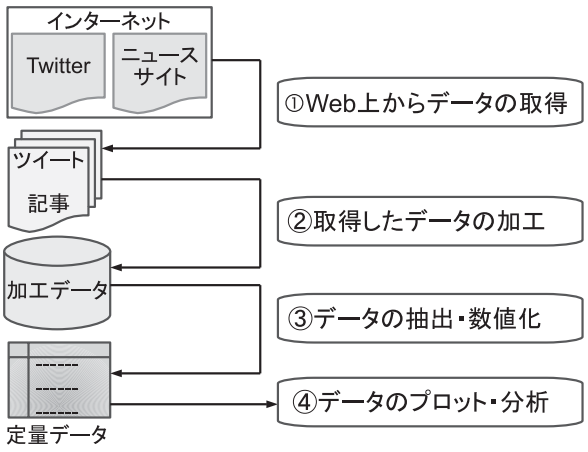

図 4 本研究の解析の流れ

3.3 取得した記事データの加工

ここではニュース記事のテキストデータの加工について説明をする. テキストデータを数値化を行うが, ニュース記事のままのデータは一つの繋がった文章のままであり, このままでは数値化の処理を行うことができない. そこでデータを数値化出来るように加工を行う. テキストデータの加工には形態素解析と呼ばれる自然言語処理の技術を用いる.

3.3.1 テキストマイニング

テキストマイニングとはデータマイニングの一種で, 文字列から構成された大量のテキストデータを解析しその中から有益となる情報を取り出す技術である.[13] テキストマイニングはひとつの固まりとなった文章を細かく区切り, その単語の出現頻度や出現傾向などの調べることで有用な情報を取り出す. 日本語は単語のわかち書きの必要であることや文法の構成がもっているゆらぎから解析が困難とされていたが, 自然言語処理技術の発達により解析が容易に行えるようになった. テキストマイニングを行うには形態素解析と呼ばれる処理が必要となる. 形態素解析については次で説明する.

3.3.2 形態素解析

形態素解析とは自然言語処理の技術のひとつである.[14] ひとつの文章を形態素と呼ばれる言語で意味を持つ最小単位に分割し, 分割した形態素の品詞情報などを取り出す処理のことを指す. 図 5 に形態素解析を行った際の文章の解析結果の例を示す.

形態素解析の例:「お待ちしております。」

文字列	読み	原形	品詞	活用	活用形
お待ち	オマチ	お待ち	名詞-サ変接続		
し	シ	する	動詞-自立	サ変・スル	連用形
て	テ	て	助詞-接続助詞		
おり	オリ	おる	動詞-非自立	五段・ラ行	連用形
ます	マス	ます	助動詞	特殊・マス	基本形
。	。	。	記号-句点		

図 5 形態素解析結果の一例

形態素解析を行うためには解析対象となる言語の文法の情報と品詞情報のついた単語の辞書と形態素解析

エンジンが必要である. 今日ではオープンソースの形態素解析エンジンが公開されている. それが MeCab と呼ばれるライブラリである. また文法及び単語辞書も MeCab 用 IPA 辞書が存在する. 本研究ではこのライブラリを使用することとする.

3.3.3 形態素解析エンジン MeCab

MeCab はライブラリとして実装されたオープンソースの形態素解析エンジンである.[10] このライブラリは奈良先端科学技術大学大学院出身の工藤拓氏によって開発されたライブラリである. このライブラリを使用できる言語は C,C++,C#,java で使用できる他, スクリプト言語の Perl,python,ruby,PHP を介しても使用することが可能である.MeCab は生成したインスタンスに含まれれるメソッドの引数に解析したい文章を与えるだけでテキストの形態素解析を容易に行うことが可能である.

3.4 データの解析

ここでは加工が終わったデータの解析について述べる. 形態素解析を行った後のデータに対して解析処理を行う. ここで行う解析処理の目的はテキストデータを分析・評価しやすい定量データに変換することである. ここで数値化を行った後に実際の株価変動と照らしあわせて分析を行う. 今回本研究で用いる数値化手法は感情解析と呼ばれる手法を用いる.

3.4.1 感情解析

ここでは本研究で用いる感情解析について説明する. 感情解析とは文書などのテキストデータを対象に行う解析である.[15] 自然言語処理の技術を応用しテキスト内の文章を数値化し, ネガティブやポジティブ等のどのような感情を持っているのかを数値として算出する. テキストから取り出せる感情の種類にはいくつかあるが, その中でネガティブ, ポジティブの２種類に絞って分析を行うものをネガポジ解析と呼ぶ. 本研究ではインターネット上のニュースサイトから記事を取得し, 形態素解析した単語データに対して感情解析を行う. テキスト解析における文書からの感情抽出ではネガティブ, ポジティブの他「予期」,「怒り」,「嫌悪」,「驚き」,「恐れ」等の様々な種類の感情の軸が扱われているが, 今回本研究では negative 及び positive の２種類の感情の軸を扱うネガポジ解析を行う. 図 6 に感情解析の大まかな流れを示す.

まず形態素解析によって最小単位に区切られたニュース記事の単語データを取り出す. その後それらの単語に

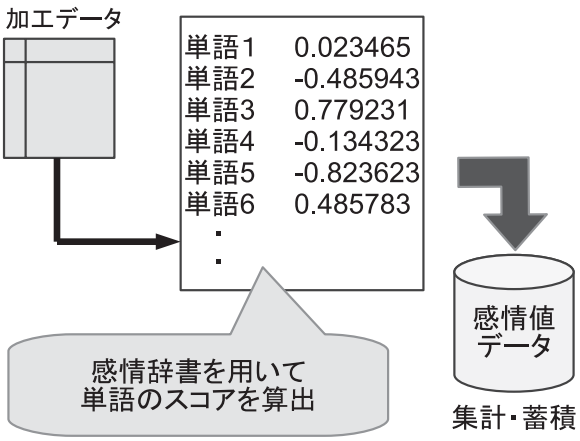

図 6 感情値付与の流れ

対して各単語がネガティブなのかポジティブなのかを判定し, 適当なスコアを付与する. 全ての単語にスコアを付与したらそれらを元に文章全体がどのような感情を持っているかを判定する.

ネガポジ解析を行うにあたって, それぞれの単語がネガティブかポジティブかを判定するための辞書が必要となる. 本研究では単語感情極性対応表とを用いる.[9] 日本語の主要な語句がネガティブなのかポジティブなのかを一対一で数値化した感情辞書である. これは一般公開されており, 研究目的に限り使用が可能なものとなっている.

対応表にはひとつの単語に対してひとつの実数値のスコアが付与されている. このスコアは対応している単語がネガティブかポジティブを数値として表されたものである. スコアの範囲は-1 から 1 で,1 に近いほどその単語がポジティブであり,-1 に近いほどネガティブであることを表している. 本研究ではこの対応表を用いてニュース記事のネガポジ解析を行う. 対応表から単語とスコアを取り出し, 対応表のデータを配列として扱う. そしてニュース記事の単語に対してスコア付けを行う. 方法としてはニュース記事一つ一つが対応表に登録されている単語かどうかを見ていく. ニュース記事の単語が対応表に登録されている単語である場合は対応するスコアをその単語のスコアとする (図 7).

上記の流れで一つのニュース記事の全ての単語のスコアを算出していく. また, 単語極性対応表は登録されている単語に限りがあるためニュース記事を始めとする全ての日本語の語句には当然対応出来ない. ニュース記事の中に含まれる固有名詞や, それ単体では意味を成さない接続詞などの語句のスコアは今回は 0 とする.

ひとつのニュース記事の単語の全てのスコアを算出したら, ニュース記事が全体でネガティブなのかポジ

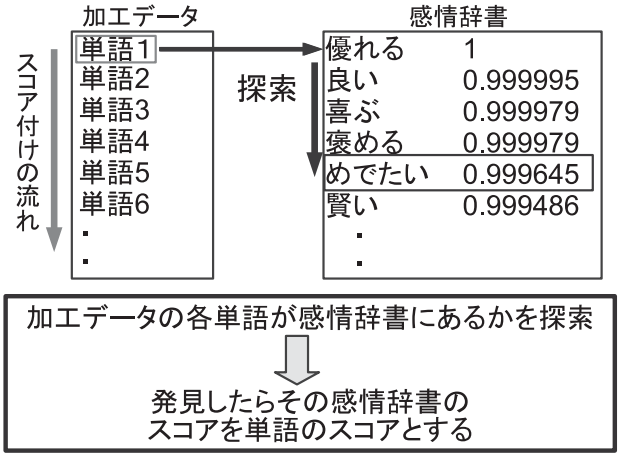

図 7 スコア付けの流れ

ティブなのかを算出する. 一つのニュース記事内の算出したスコアの平均値をとることとする. その平均値が正の値であればそのニュース記事はポジティブ, 負の値であればネガティブと判定する. ここまでの処理を全てのニュース記事に対して行っていく. そのネガポジ解析の結果をニュース記事ごとに保存させる.

ここまでで取得したニュース記事を数値化する処理が終了する. 次で数値化したデータの分析方法について述べる.

3.5 解析データの分析

ここではこれまでの手順で数値化を行ったニュース記事のデータと株価変動の分析の仕方について述べる.

取得したニュース記事のネガポジ解析の結果から, まずはポジティブと判定されたニュース記事の数とネガティブと判定されたニュース記事の数が時系列でどのように推移しているかを見ていく. 最終的にはこれらをグラフとして表示させ推移を見る. しかしニュース記事の判定結果のみでは関係を見ることは難しいため最終的に表示させるグラフは以下のものを一つのグラフ上に表示させる.

- ポジティブと判定されたニュース記事数
- ネガティブと判定されたニュース記事数
- 日経平均株価

これらの推移を見て株価変動とニュース記事との間に何らかの関係がないかを見ていく.

今回の推移は各日付でニュース記事判定結果の数をグラフ上にプロットする為, ニュースで取り上げたれた事柄の時系列の関係などは考慮していない. 株価の変動はその日その日の出来事のみに依存するのではなく, 時系列の出来事のつながりなども大きく影響することから本来は各日ごとのニュースの関係も探ることが望ましいが, 今回は行わない. 今回はニュース記事そのものが株価変動と関係があるかどうかを探ることが目的な為, 時系列での出来事を考慮した分析は今後行うこととする.

表 1 実験環境

新聞名	URL
読売新聞	http://www.yomiuri.co.jp/
毎日新聞	http://mainichi.jp/
朝日新聞	http://www.asahi.com/
日本経済新聞	http://www.nikkei.com/
産経新聞	http://www.sankei.com/

またニュース記事の判定結果のプロットと同時に, 日経平均株価の推移とニュース記事との間に相関があるかどうかを定量的に判断するために相関係数の算出を行う. 以下の式で計算を行う.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\{\sum_{i=1}^{n}(x_i - \overline{x})^2\}\{\sum_{i=1}^{n}(y_i - \overline{y})^2\}}}$$

ρ は相関係数を表し,x がポジティブ及びネガティブどちらかの判定された記事数を表し,y が日経平均株価の数値を表す.$\overline{x}$ と $\overline{y}$ はそれぞれの平均を表す. 上記の式を用いてネガティブと判定された記事における株価変動との相関係数, ポジティブと判定された記事における株価変動との相関係数をそれぞれ算出する. 算出された値をみて両者の間にどれほどの相関が現れているかを確認する.

4. 実験

ここでは本研究で行う実験について述べる. ここまで解析及び分析まで述べた. よってここでは使用するデータや期間などの実験設定等を述べていく.

4.1 使用データ

ここで今回の実験に用いるデータについて述べる. 本研究では Web 上のテキストデータが解析対象であることを既に述べた. その中でも対象となるデータは日本のニュースサイトのニュース記事である. 今回の実験で対象となるニュースサイトは表 1 に示すとおり国内の全国紙 5 社が運営しているニュースサイトである.

あらかじめ蓄積させ用意しておいた上記のニュースサイト 5 箇所のトップニュース (公開順に並んだ最新ニュース) の解析を行っていく. またニュース記事の解

表 2 用いるデータの期間

ニュース記事	2014/7/7 〜 2014/9/30

析結果と株価の推移を照らし合わせるため, 株価の値もインターネット上の株価情報サイトから取得してくる. 取得してくる. サイトは先に述べたように Yahoo ファイナンスより日経平均株価を取得してくるものとする.

また、今回表 2 に示す期間のデータを用いる.

4.2 実験方法

実験方法について述べる. 今回はこれまでに述べた流れに沿って分析を進めていく. 蓄積させたデータに対して解析処理を行い結果をプロットさせ, 同時に相関係数も算出し関連性があるかどうかを見ていく.

5. 結果

ここでは実験の結果を示す. まずニュースサイトの記事の解析結果を示していく. ニュース記事に関しては全国紙 5 社のニュース記事を扱った. まずは各ニュースサイトごとの解析結果を示す. これから示すグラフには 3 つのデータが表示されている. それぞれ日経平均株価, ポジティブと判定されたニュース記事数, ネガティブと判定されたニュース記事数である.

まずは各ニュースサイトごとの結果を図 8,9,10,11,12 に示す. 全てのニュースサイトのニュース記事の総数で見たものが図 13 である. ポジティブと判定された記事とネガティブと数値を示したものが表 3, これらのニュース記事の判定結果と株価変動との相関係数は表 4 である.

6. 考察

まずここでは 5 章での結果を受けて, 実験全体に関する考察を行っていく.

まずニュースサイトのネガポジ解析の結果では、全てのニュースサイトの総数の場合, 7 月ではポジティブと判定された記事の数が 887 個, ネガティブと判定された記事の数が 3762 個となっておりネガティブと判定された記事の割合が多い結果となった. 8 月の間でも同様にポジティブと判定された記事が 913 個, ネガティブと判定された記事が 4455 個となっている. 9 月の間でも同様にネガティブと判定された記事の割合が多くなっている. また株価と解析結果の相関係数はポジティブと判定された記事と株価の場合が 0.002, ネガティブと判定された記事と株価の場合が-0.322 となった.

記事全体で見た場合と各ニュースサイトごとに見た場合の両方で, 全ての期間を通してネガティブの割合が多い傾向が見られた. 全体的にネガティブに偏った傾向であるが, 相関係数を見るとどれも低い値で推移している. このことから今回のこの記事全体から見た結果からはニュース記事と株価変動との間に相関がほとんど見られないことが示されている. ニュースサイトごとの個別の結果を見てもネガティブと判定されている記事数の割合が多いことに加えて相関係数が低く相関がほとんど見られない結果となった. どのニュースサイトに対しても同様の割合が出ることから, まず今回行った手法がネガティブと判定されやすいものだったのではないかという点がまず挙られる.

表 3 記事の判定結果

新聞名	期間	ポジティブ	ネガティブ
読売新聞	7 月	255	979
	8 月	241	1032
	9 月	278	964
毎日新聞	7 月	213	729
	8 月	211	880
	9 月	158	744
朝日新聞	7 月	160	683
	8 月	140	916
	9 月	142	818
日本経済新聞	7 月	119	685
	8 月	168	845
	9 月	130	763
産経新聞	7 月	140	686
	8 月	153	782
	9 月	138	834

今回行った解析手法は感情解析の中でもネガティブかポジティブかの 2 つの軸でテキストを判定するネガポジ解析を用いた. ネガポジ解析を行った際に用いたのが単語極性対応表である. の対応表は単語ごとにスコアが付与されているものである. 今回のネガポジ解析では対応表のスコアをそのまま使用して行った. この対応表はポジティブなスコアの単語とネガティブなスコアの単語の数と比較するとネガティブなスコアの単語の方が割合的に多く登録されている. 今回の実験結果のネガティブだと判定されたニュース記事数が多かったが, これはこの対応表の割合と概ね同じであった.

今回は単語レベルでの数値化の処理を行ったが, 本来文章内に含まれる感情は単語だけではなく単語の係り受けや, 文脈など単語より大きな単位で決定されるもの

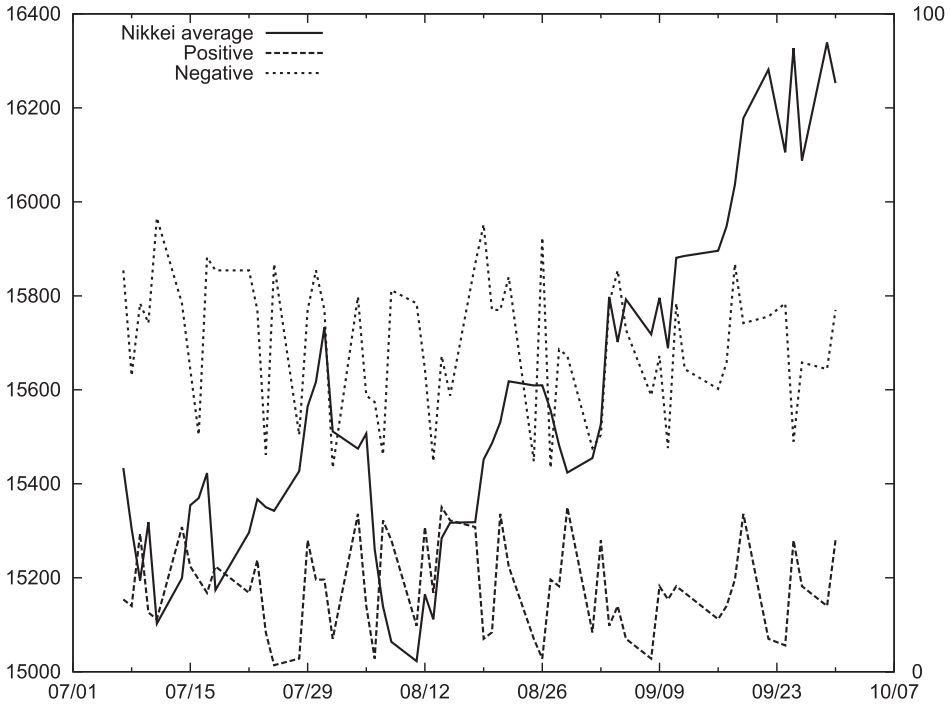

図 8 ニュース記事 (読売新聞) のネガポジ解析結果と日経平均株価の推移

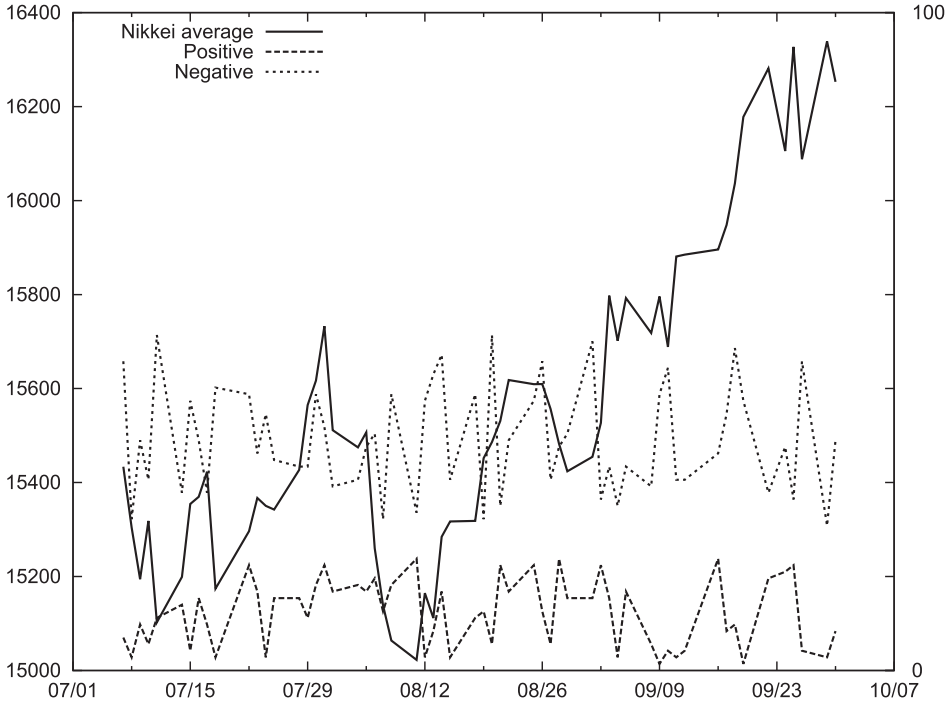

図 9 ニュース記事 (毎日新聞) のネガポジ解析結果と日経平均株価の推移

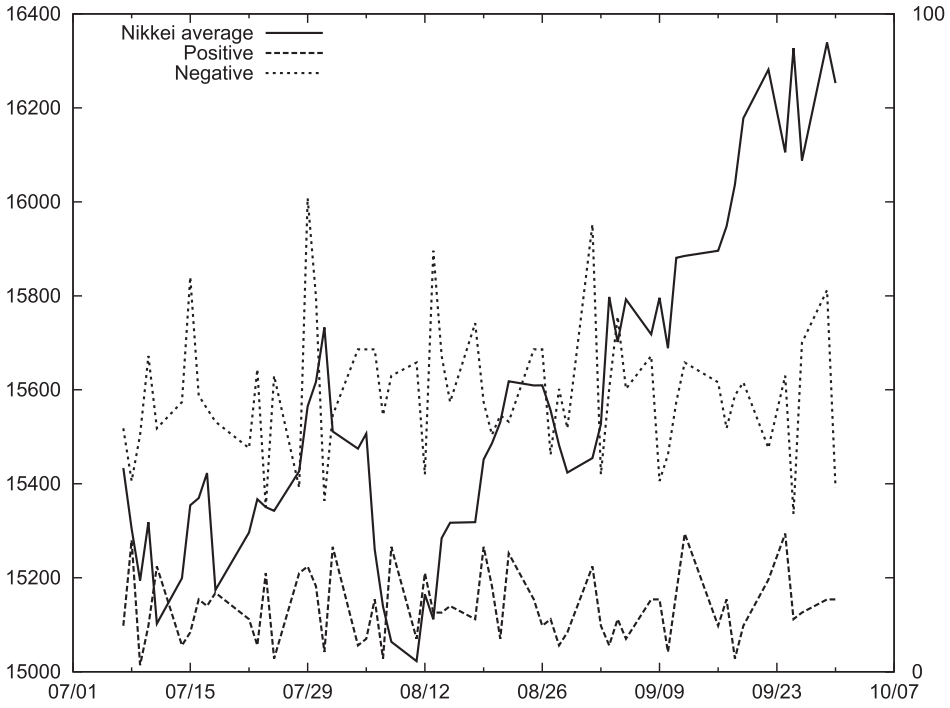

図 10 ニュース記事 (朝日新聞) のネガポジ解析結果と日経平均株価の推移

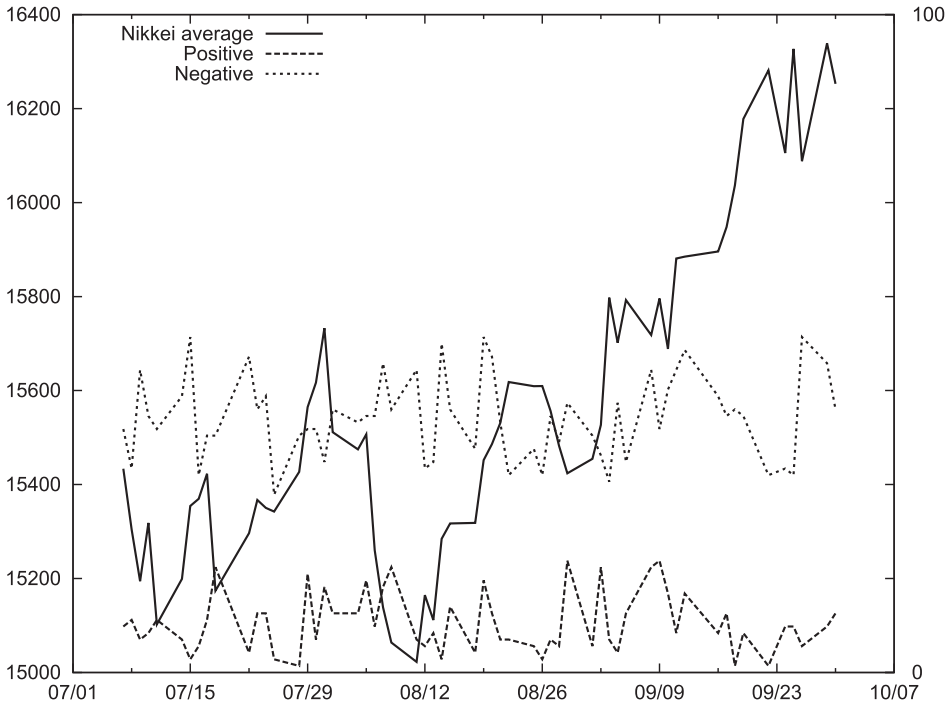

図 11 ニュース記事 (日経新聞) のネガポジ解析結果と日経平均株価の推移

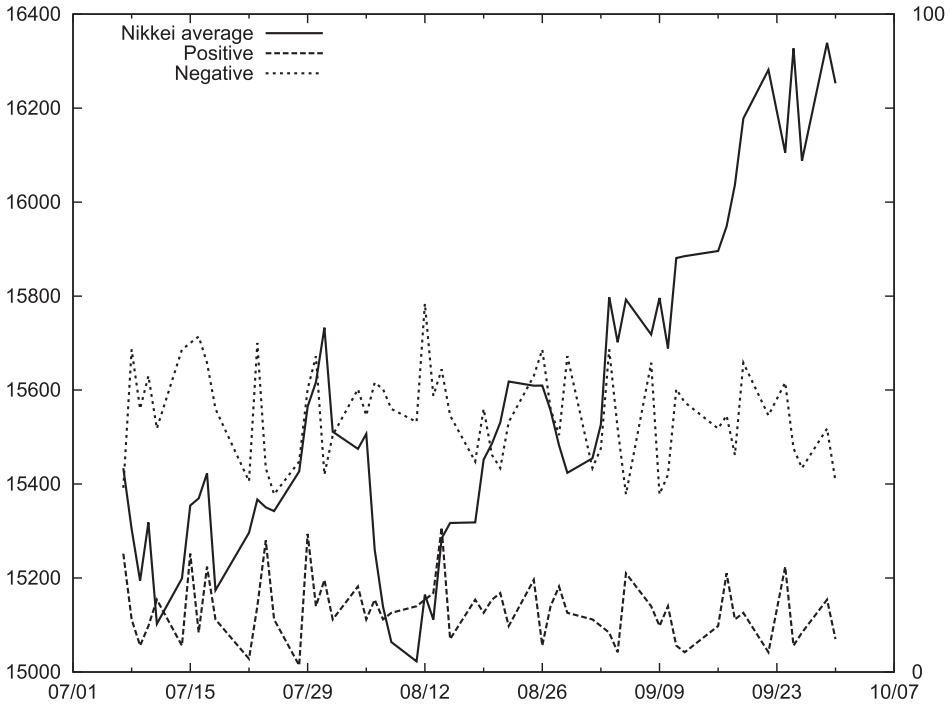

図 12 ニュース記事 (産経新聞) のネガポジ解析結果と日経平均株価の推移

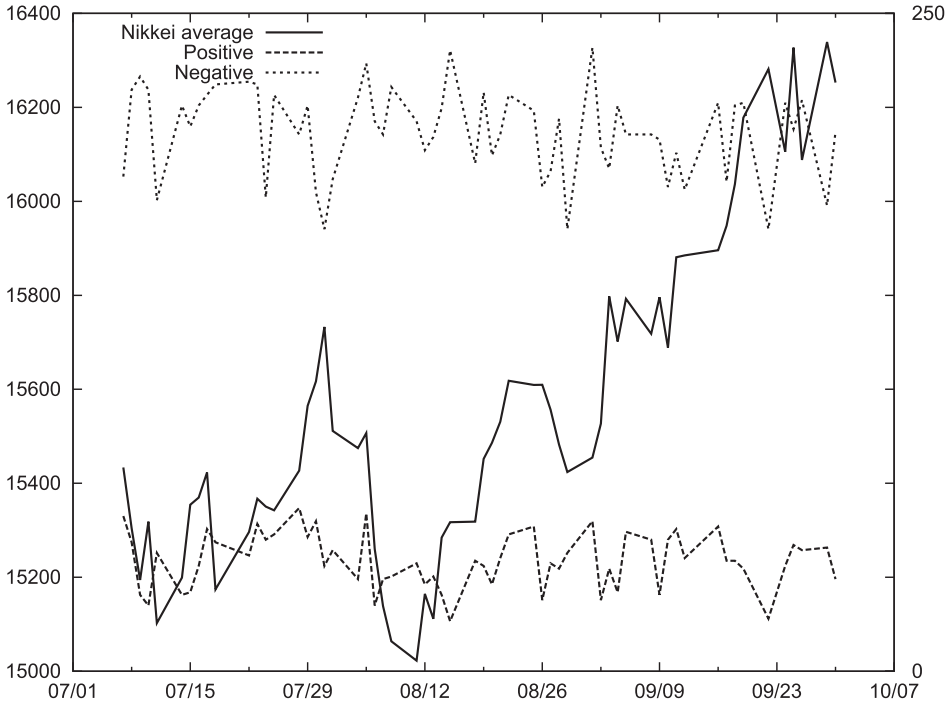

図 13 全ニュースサイトのネガポジ解析結果と日経平均株価の推移

表 4 記事の判定結果

新聞名	期間	相関係数 (positive)	相関係数 (negative)
読売新聞	7 月	0.415	-0.232
	8 月	0.289	-0.122
	9 月	0.074	-0.194
毎日新聞	7 月	0.290	-0.157
	8 月	0.100	-0.112
	9 月	-0.605	-0.235
朝日新聞	7 月	-0.093	-0.222
	8 月	0.202	-0.046
	9 月	0.075	-0.095
日本経済新聞	7 月	-0.129	-0.324
	8 月	0.092	-0.085
	9 月	0.085	-0.134
産経新聞	7 月	0.048	-0.358
	8 月	-0.256	-0.196
	9 月	-0.115	-0.120

もある. 対応表をそのまま使用した今回の実験では係り受けや文脈などは一切考慮しておらず, 文脈によってはネガティブにもポジティブに取れるものの存在を無視していた. 加えて対応表には登録されている単語に限りがあり, 今回の実験では対応表に記載されていない未知の単語のスコアを 0 として解析を行っていた. これらの点が今回の実験結果となった要因であると考えられる.

相関があるかどうかを確認するためには上記の点を考慮した上での再実験が必要となってくる. 具体的にはニュース記事の形態素解析結果を単語ベクトルとして表し, 共起関係などを見て未知の単語のスコアを機械的に新たに付与したり, 対応表のスコアの更新・変更したりする処理を行うことが必要なのではないかと考えられる. この処理で単語レベルでの解析が向上されるのではないかと考えられる.

よって今後はまずはスコアの更新, 未知の単語のスコア付けの処理を行った後, 単語のみではなく構文なども考慮した解析処理を行う必要があると考えられる. 未知の単語の感情値の重み付け処理には Bracewel らの WorldNet を用いた半自動での感情辞書構築の先行研究がある.[11] この研究で行われた手順を用いることで未知の単語の感情値の付与が行えるのではないかと考えられる.

また, 今回はニュース記事を対象としたが, 先行研究では SNS のデータを使用していた. ニュース記事と SNS のデータでは同じテキストでも内容に大きく違いがある. ニュース記事は起こった出来事を端的に説明しているのに対し,SNS のつぶやき等のデータは投稿者の感情が大きく関与している場合が多い. そのためニュース記事と SNS に同様の解析を行った場合,SNS の方が特徴が出やすいのではないかと考えられる. よって今後国内の SNS を対象にした解析も検討していく.

7. まとめ

本研究はインターネット上のテキストデータと株価変動の間に相関関係があるかどうかを探った. 先行研究では twitter のツイートの感情値用いた株価推定では高い精度を出していた. この結果から Web 上データと株価変動の間に何らかの相関関係があることが示唆されたと言えるが, 厳密な関係が解明されたわけではなく, なぜ高い精度になるかは不明のままであった. 加えて言語仕様が異なる日本国内の Web データに対して先行研究の手法を応用しようとした場合に同様の結果が得られるかどうか, 国内での研究成果は報告されていなかった.

そこで本研究では日本国内の Web サイトに対して感情解析を行い, その結果が株価変動との相関関係を見出すことが出来るのかを目的とした. 本研究では解析対象として国内のニュースサイトを使用した. ニュース記事に対してネガポジ解析を行い, その結果と株価の変動を見た. 結果は全体的にネガティブに偏る傾向が見られたが, 相関係数はどれも低い値を示していた. そのため今回の結果からは両者の間に相関を見出したとは言えな

い結果となった. 今後の展望として今回行った手法そのものの改善が挙げられる. さらに今後はネガティブかポジティブかの一つの感情の軸ではなく複数の感情の軸からの解析を検討, ニュース記事だけでなく国内 SNS の解析も行っていくことが挙げられる.

参考文献

[1] W. Antweiler and M. Z. Frank : Is all that talk just noise? the infermation content of internet stock message boards, Jeurnai of Finance, 59-3, 1259-1294 (2004).

[2] P.C.Tetlock : Giving content to investor sentiment:The role of media in the steck market, Journal of Finance, 62-3, 1139-1168 (2007).

[3] P.C.Tetlock,M.Saar-Tsechansky and S.Macskassy : More than words:language to Quantifying measure firms fundamentals, Joumal of Finance,63-3, 1437-1467 (2008).

[4] 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 (2011).

[5] Johan Bollen,Huina Mao,Xiao-Jun Zeng : Twitter mood predicts the stock market, Journal of Cmputation Science, 2, 1-8 (2010).

[6] E.K.Berndt,B.H.Hall and R.E.Hall : Estimation and Inference in Nonlinear Structural Models, Annals of Economic and Social Measurement, 3-4, 103-116 (1974).

[7] Jason E. Kutsurelis : Forcasting financial markets using neural networks: An analysis of methods and accuracy, Naval PostGraduate School (1998).

[8] 左 毅, 北 栄輔 : ベイジアンネットワークを用いた株価予測について, 情報処理学会論文誌 数理モデル化と応用, 3-3, 80-90 (2010).

[9] Hiroya Takamura, Takashi Inui, Manabu Okumura : Extracting Semantic Orientations of Words using Spin Model, In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005), 133–140 (2005).

[10] MeCab:Yet Another Part-of-Speech and Morphological Analyzer,http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.

[11] David B. Bracewell : Semi-Automatic Creation of an Emotion Dictionary Using WordNet and its Evaluation, In Proceedings of 2008 IEEE Conference on Cybernetics and Intelligent Systems, 1385-1389 (2008).

[12] 薄井 彰 : 企業評価とファンダメンタル分析, 年報経営分析研究, 17, 2-9 (2001).

[13] 和泉 潔, 松井 宏樹, 松尾 豊 : 人工市場とテキストマイニングの融合による市場分析, 人工知能学会誌,22-4, 397-404 (2007).

[14] 松本 裕治, 形態素解析システム「茶筌」: 情報処理, 41-11, 1208-1214 (2000).

[15] 倭本一輝, 川本淳平, 浅野泰仁, 吉川正俊 : 感情解析のための分布モデルと相互強化型解析手法, 電子情報通信学会第,2 回データ工学と情報マネジメントに関するフォーラム (2010).