

株価予測におけるビッグデータの利用

柳瀬 仁洋* 小高 知宏* 黒岩 丈介** 白井 治彦***

Use of Big Data in Stock Price

Yoshihiro YANASE*, Tomohiro ODAKA*, Jousuke KUROIWA** and Haruhiko SHIRAI***

(Received February 2, 2018)

In this research, stock price is predicted using web data. Many researchers are conducting stock price prediction research. In overseas research, it was announced that highly accurate stock price prediction can be performed using emotional values. By analyzing Web data it has been shown that it helps to predict stock price from results. It is necessary to investigate the usefulness of previous research in Japan.

We investigated the relationship between web data and stock price in Japan by conducting stock price prediction using web data. And we made stock price prediction using decision tree analysis of machine learning. As a result of the experiment, stock price prediction of 63.8

Key words :Big data, stock price, emotion analysis, machine learning

1. はじめに

本研究では、感情解析を用いてインターネット上のデータと株価の変動の間に相関関係があるかどうかを調査することを目的としている。

今日ではビッグデータを用いた研究が注目されている。またビッグデータは、日々増加しており様々な研究や開発に応用・発展されている。その中で、ビッグデータを解析して昨今では予測するのが非常に困難と言われていた経済状況の予測などの研究が注目されてきている。株価予測の分野も今日ではインターネット上で取得することができる SNS やニュース記事などの様々なデータを用いて予測するという研究が多くされてきている。^[1]

その中でも 2010 年にアメリカの論文で発表された「Twitter mood predicts the stock market」^[2] は、非常に

興味深い研究成果が挙げられている。アメリカ国内のインターネット上の Twitter ツイート文を感情解析することによってテキストデータを数値化してダウ平均株価との相関性を調査し実際に株価予測を行った。その結果約 87 % の精度で株価を予測することが可能であると発表された。またこれによりインターネット上のテキストデータの感情値と株価の変動の間には、何かしらの相関関係があることが示された。

しかしながらこの先行研究はアメリカで研究されたものであり、当然言語が英語となっている。日本語と英語では、文法の構造や単語は大きく異なっている。また当然アメリカと日本では、表現方法が異なっておりさらに、Twitter の利用する用途やユーザー層が異なっていることが容易に予想できる。これらの理由により、日本国内でインターネット上のテキストデータを用いた株価予測でこの先行研究のような高い精度での予測が可能といった研究成果は少ない。

そこで本研究では、日本国内のインターネット上のニュース記事と日経平均株価との間に何らかの相関性がないかを調査した。国内で公開されているニュース記事で感情解析をして数値化し、この感情値を用いて実際に株価予測を行った。

* 大学院工学研究科 原子力・エネルギー安全工学専攻

** 大学院工学研究科 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 Intelligence Systems Course,
Graduate School of Engineering

*** Technical Division

テクニカル分析	価格分析	規則時系列分析	トレンド系分析
			オシレーター分析
		規則時系列分析	
	その他の分析	出来高	
		確率論	
		相対比較	

図1 テクニカル分析の種類

本稿では、2章では、株価予測についての現状と手法について述べる。3章では実際に本研究で行う株価予測の流れを説明し、4章では実際に行った実験内容の詳細を述べる。5章では実験結果と結果からの考察を述べ、6章でまとめと今後の課題を述べる。

2. 株価予測の手法

株価予測とは、市場に出回っている株価の推移を予測することである。主に投資家と呼ばれる人たちが、様々な情報を元に分析して予測の精度を高めている。本章では、株価予測における分析手法や本研究に株価予測手法の提案を行う。

2.1 株価予測の分析手法

株価予測の分析手法は、主に2種類に分類できる。それぞれテクニカル分析とファンダメンタル分析である。この2つの分析方法を説明する。テクニカル分析は、統計的に分析を行っていくものである。過去の株価の推移パターンや数値などの様々なデータを分析して予測を行う手法である。この分析法では、主に時系軸と価格軸の統計を見て分析を行う。その中で規則的な時系列をみて行う分析方法を規則時系列分析と呼ばれる。規則時系列分析の中でも、市場の「上昇」、「下降」、「横ばい」の3つの状態の内今後どの状態になるのかまた、今現在どの状態となっているのかを分析する手法をトレンド系分析と呼ばれる。更に規則時系列分析には、オシレーター分析と呼ばれる分析方法もある。また不規則的に時系列を見る不規則時系列分析などがある。その他にもいくつかの分析手法は存在している。テクニカル分析の種類を分類を図1に示す。テクニカル分析は、投資家の中でも主に短期での株取引を行う時に重要視される分析である。ファンダメンタル分析とは、投資する企業の事業方針や、

行事また財政状況などの情報を元に分析を行う手法である。^[5] ファンダメンタル分析では、テクニカル分析とは異なり、統計的に株価を予測するのではなく、安定性、成長性、収益性などの企業価値を見極める分析方法である。これらの企業価値を見極めるために様々な側面から指標を用いて分析する。例えば、安定性を分析する時に用いる指標は、自己資本比率や流動比率またキャッシュフローなどがある。重要な指標と言われる DER(Price Earnings Ratio) や PBR(Price Book-Value Ratio)、また ROE (Return on Equity) などがある。これらの指標の詳細を図2に示す。テクニカル分析とは逆にファンダメンタル分析では、長期的に投資を行う時に重要となる分析方法である。

2.2 ビックデータを利用した株価予測

近年では、ビックデータ解析が様々な研究分野で利用されている。それに伴い株価予測の研究にもビックデータを利用した研究が多く行われてきている。ビックデータを利用した株価予測は、株価の変動する要因と考えられる過去の株価の変動などの定量化されていないデータを株価予測に用いることができる。よって従来の定量的データのみを利用した株価予測の精度よりビックデータを用いた株価予測の精度が良くなるのが容易に考えられる。

様々な研究に多く用いられるようになったビックデータの中で主に株価予測で利用されているビックデータの種類の、TwitterなどのSNSやニュースサイトといったテキストデータが主に利用されている。これらのテキストデータには、株価の変動の要因となる可能性がある定量化されていない様々なデータが多く含まれている。この定量化されていないデータを数値化させることによって株価予測に用いることが可能となる。この数値化して予測する手法の一つとして感情解析がある。感情解析とは、テキストデータの

PER	企業の利益に対する株価の割合
PBR	企業の資産に対する株価の割合
BPS	総資産からどれほどの利益を上げたのかを示す指標
ROE	資本からどれほどの利益を上げたのかを示す指標

図2 ファンダメンタル分析の指標例

単語毎に含まれている感情値の度合いを数値として抽出する解析手法のことである。感情解析の手法は先行研究でもこの手法が用いられているものがある。

3. 本研究における株価予測の手法

本章では、本研究で実際に使用した株価予測手法及び構築した株価予測システムについて述べる。

3.1 本研究での解析目的

本研究では、自国である日本でのインターネット上のテキストデータと株価の変動に相関性があるのかを調査するのが目的である。先行研究で行われていた株価予測では、確かにかなり高い精度で株価の予測が行われていた。しかし先行研究の実施した地域は日本ではなくアメリカで実施された。当然アメリカと日本とは、言語が異なっており日本語と英語とは、単語や文法も異なってくる。

また文化も異なっているので、表現方法やインターネットの利用者層、使用目的などが異なると考えられる。よって先行研究で行われた感情解析と同じ着眼点で解析を行った時に似たような結果を得ることができるのかをまず確認する必要がある。よって日本国内でのインターネット上のテキストデータを使用して同様に感情解析及び株価予測を行い株価変動と相関が得られたかを確認する。

3.2 本研究の解析の流れ

本研究で行う解析の主な流れを説明する。今回の解析手法では、大きく4つの工程となっている。この工程の流れは、以下に示す。また具体的な解析手法の流れを図3に示す。各工程の詳細な説明は次の節から行う。

1. インターネット上のテキストデータ・株価データの取得
2. テキストデータの加工
3. 定量化データの抽出及び数値化
4. データの分析・株価予測

3.3 インターネット上のテキストデータ・株価データの取得

この節では、今回の研究で使用したデータについて述べる。本研究で扱うデータは、以下の2つである。

- 日経平均株価

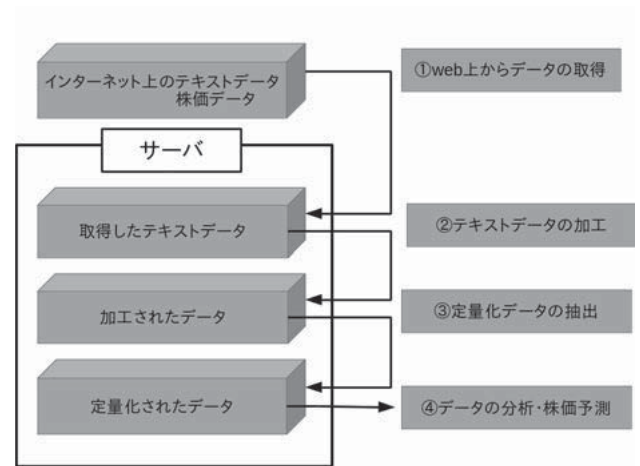


図3 システムの流れ

● ニュース記事

今回の研究では、インターネット上のニュース記事と日経平均株価との相関性を調査する。ニュース記事の取得は、それぞれのニュースサイトのニュース記事がまとめられている Yahoo ニュース (<https://news.yahoo.co.jp/>) から取得することができる。また日経平均株価は、Yahoo ファイナンス (<https://finance.yahoo.co.jp/>) から取得することができる。

取得するニュースサイトは、一日に更新されたトップニュース記事を全てその日の終わりに取得する。取得するニュース記事は、日本国内の新聞社である、朝日新聞、毎日新聞、読売新聞、産経新聞の4社からニュース記事を取得する。取得した記事は、記事のタイトルとニュース記事の本文をまとめて保存しておく。本研究では、日経平均株価とそれぞれのニュース記事を、127日間作成したプログラムで取得しておきデータベースに保存されている。取得したニュース記事を本研究では、テキストデータとする。この保存したデータから本研究では、テキストデータの加工や感情解析および株価予測を行っていく。

3.4 取得したテキストデータの加工

この節では、取得したテキストデータを感情解析できるようにするための加工手法を説明する。テキストデータの数値化を行うには感情解析を行わなければならない。しかしながら感情解析は単語に含まれる感情を解析するため、一つのつながった文章のままでは単語をその文章の中から見つけることが出来ず感情解析を行うことが出来ない。つまり取得して未加工のままのテキストデータからでは、感情解析

ができない。よって取得したテキストデータを単語毎に区切る加工を行わなければならない。今回テキストデータを加工するのに形態素解析を行う。

3.4.1 自然言語処理とテキストマイニング

自然言語処理とは、日本語や英語などの人間が使用している言語（自然言語）をコンピュータが処理することを意味する。自然言語処理の中でも今回はテキストマイニングと呼ばれる技術を使用する。テキストマイニングとは、大量の文字列からなるテキストデータから有益な情報を取得する技術のことである。主に形態素解析を行い文章であるテキストデータを名詞や動詞、形容詞などの単語ごとに区切りそこから単語の頻出頻度や出現傾向等を調べることで情報を取り出す。本来日本語は、英語などの言語と比較して単語に接頭辞や接尾辞のような形態素を付着することで意味をもたせる膠着語であったり、単語と単語の区切りに空白などを入れずにベタ書きで書かれていたりするのでテキストマイニングは困難とされていた。

しかし自然言語処理の技術が向上されてきたことによって日本語でもテキストマイニングすることが容易になった。今回のテキストマイニングでは、形態素解析を行う必要がある。次の節で形態素解析についての説明をする。

3.4.2 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。そもそも形態素とは、文章のなかで意味を持った最も短い文字列のことである。形態素解析では、取得したテキストデータを形態素毎に区切り、さらに抽出した形態素の品詞情報なども解析する処理を行う。次の図4に実際にテキストデータの形態素解析を行った時の解析結果の例を示す。

形態素解析をパソコン上で行うには、形態素に区別するためにそれぞれの文法情報や品詞情報が掲載されている単語の辞書と専用の形態素解析エンジンが必要である。形態素解析エンジンの中には、インターネット上でオープンソースとして公開されているものがある。そういったオープンソースの一つが Mecab と呼ばれる形態素解析エンジンである。Mecab は、インターネット上で公開されているライブラリである形態素解析エンジンである。^[3] この形態素解析エンジンは、奈良先端科学技術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

ライブラリは、様々なプログラム言語から応用して利用することができ、また自身のパソコン上からでも動作させることができる。Mecab を利用すれば Mecab の引数に形態素解析したいテキストデータを与えるだけでテキストデータの形態素解析を行うことができる。また形態素解析に必要な単語の辞書も Mecab のオープンソースの中に含まれており、Mecab 専用の IPA 辞書が存在している。本研究では、この Mecab ライブラリを使用して形態素解析を行う。

3.5 定量化データの抽出及び数値化

ここでは、上述した加工方法により加工されたテキストデータから定量化されたデータの抽出・数値化方法について述べる。Mecab によって形態素解析を行ったデータを解析処理する。今回の解析を行うことによって定量データではないテキストデータを数値化して定量データにすることが出来る。テキストデータを数値化された定量データにした後に株価変動と分析して株価予測を行う。今回の研究で用いる解析手法は、感情解析である。

3.5.1 感情解析

ここでは、本研究に用いる感情解析についてと実際の解析方法についての説明を行う。本研究での感情解析とは、テキストデータから解析を行うことである。形態素解析を行い加工された定量データを感情解析して数値化し、いくつかの種類の感情のどの感情が含まれているかを数値化する。またテキストデータから取得できる感情の種類はいくつか存在している。例えば、非常に高い精度で株価予測を行っている海外の先行研究では、[Calm], [Alert], [Sure], [Vital], [Kind],

例文：株価予測を行った。

文字列	読み	原形	品詞	活用	活用形
株価	カブカ	株価	名詞		
予測	ヨソク	予測	名詞		
を	ヲ	を	助詞		
行っ	オコ ナツ	行う	動詞	五段・ ワ行促 音便	
た	タ	た	助動 詞	特殊・ タ	基本 形
。	。	。	句点		

図4 形態素解析の例

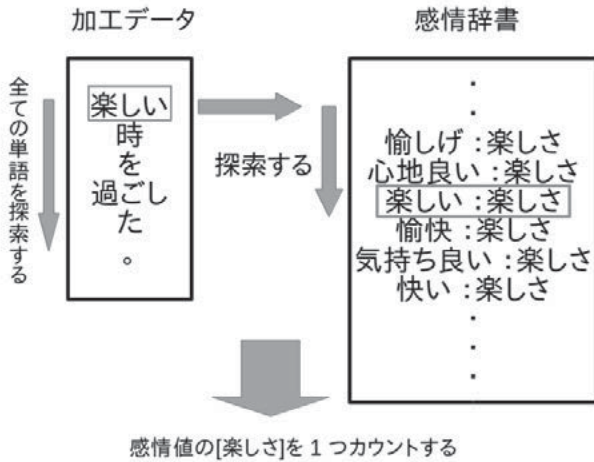


図5 感情解析の流れ

[Happy]の5種類の感情因子を用いた感情解析が行われている。他にもテキストデータの単語が、ネガティブかポジティブの2種類のみで感情解析を行うネガポジ解析と呼ばれるものもある。今回の研究では、インターネット上のニュースサイトからニュース記事を取得して、取得したテキストデータを形態素解析して取得した単語データに対して感情解析を行う。上記に例として述べたように感情解析の方法には、いくつもあり様々な感情因子で区別して解析を行うことができるが、今回は[喜び][怒り][哀しみ][楽しさ][愛しみ][憎しみ][穏やか]の7つの感情因子に分けて感情解析を行う。

図5に本研究で実装した感情解析の流れを示す。まずMecabによって形態素解析され形態素に区切られたニュース記事の単語を取り出す。取り出したそれぞれの単語に対して一つ一つの単語が上記に述べた7つの感情の中でどの感情が含まれているかを定める。全ての単語データを照合し終えたらそれを元にそれぞれの感情がそのテキストデータにいくつ含まれているかを算出する。この算出結果が今回の感情値となる。

今回の感情解析を行っていく中で、単語にどの感情が含まれているのかを判断するためには、あらゆる単語それぞれがどの感情に含まれているのかの情報が記載されている辞書を用いなければならない。本研究では、データソースであるWordNet, WordNet-Affect, 日本語WordNetを利用して辞書を作製する。^[4] WordNetは英語版の概念辞書となっており、WordNet-AffectはWordNetの単語に感情をさらにラベリングするためのデータソースである。また日本語WordNetは、英語で構成されているWordNetを日本語に変換するた

例文

野球の試合を行ったが、結果は負けてしまった。
残念だったが、初めてレギュラーになれて
うれしかったが緊張した。

感情解析を行った結果

喜	怒	哀	楽	愛	憎	穏
1	0	2	0	0	0	0

図6 感情解析の例

めのデータソースとなっている。これらのデータソースは、インターネット上で公開されており研究のための利用が可能となっている。作成した感情辞書では、7000語以上の単語が掲載されておりそれぞれの単語に7つの感情のいずれかが付与されている。今回の研究では、感情解析はテキストデータを定量化するための一つ的手段とみなして、感情それぞれに意味がないものとしている。この感情辞書を用いてニュース記事の感情解析を行う。まずニュース記事の単語が感情辞書に含まれているか一つ一つの単語を感情辞書で検索をする。もしその単語が感情辞書に含まれているのであれば、その単語の感情をカウントする。ニュース記事全てが検索し終わった時にそれぞれの感情値がいくつあったかを算出する。具体的な動作例を図6に示す。また当然ニュース記事の中には、感情辞書に含まれていない単語もいくつか存在している。今回はそのような固有名詞やスラング、単語などはカウントしないものとする。最終的に一日分のニュース記事のテキストデータの感情解析を行い、感情値の合計値を定量データとする。

これで取得したニュース記事から感情解析をすることによって数値化された定量データの修得が出来る。次の節で取得した定量データを用いた分析方法について説明する。

3.6 データの分析・株価予測

この節では、これまでの方法で定量データにしたテキストデータであるニュース記事と日経平均株価を用いた株価予測の方法について述べる。

テキストデータであるニュース記事の感情解析の結果からまずそれぞれの感情因子の推移と日経平均株価の推移が時系列でどのようになっているのかを見る。これらをグラフにして時系列でどのように変

表 1 取得するニュースサイト名

ニュースサイト名	URL
朝日新聞	https://headlines.yahoo.co.jp/list/?m=asahi
毎日新聞	https://headlines.yahoo.co.jp/list/?m=mai
読売新聞	https://headlines.yahoo.co.jp/list/?m=yom
産経新聞	https://headlines.yahoo.co.jp/list/?m=san

化しているのかを確認する。今回グラフとして表示するデータは以下の通りである。

- 日経平均株価
- それぞれの感情値

これらのデータの推移を見て株価の変動とニュース記事との間の相関性の有無を確認する。

株価予測を行う前に感情値と相関係数を算出する。相関係数の算出は以下の式で行う。

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

r_{xy} は相関係数を表す。 n 組のデータ $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ がある時、これらの x, y のそれぞれの平均をそれぞれ \bar{x}, \bar{y} とする。相関係数は、2 つの変数間の直接的な関係の強さを示す係数であり算出値は -1 から 1 までの値となる。一般的に算出値の絶対値が 1 に近いほど関係性が強いと言える。相関係数の算出結果も加味しつつ株価予測を行う。以降に機械学習を用いた株価予測の手法を示す。

3.6.1 機械学習を用いた株価予測

機械学習とは、近年注目されている分野であり蓄積されたデータから傾向や特色などを機械が学ぶアルゴリズムのことである。今回の研究では、機械学習の中の一つである、決定木分析を行った。

決定木分析とは、学習時に決定木と呼ばれる探索木を作成し、予測時にはこの決定木を元に株価予測を行う学習方法である。決定木分析では、目的変数と呼ばれる最終的にこの変数がどうなるかを決定する変数と説明変数と呼ばれる目的変数を決定するための条件となる変数が必要になる。本研究では目的変数は日経平均株価の増減とし、説明変数に 7 つの感情値とする。決定機分析による株価予測には、主な流れとして 2 つの作業が行われる。

まず学習作業である。学習作業では、感情値と日経平均株価を元に決定木を作製する。この時に株価が上昇する時の感情値の傾向と下降する時の傾向を学習している。次に予測作業である。予測作業では、感情値と学習時に作成した決定木を用いる。決定木を元にその日の感情値の条件にあった株価の変動を予測する。予測後その予測結果が実際に正しいのかを確認する。

決定機分析による株価予測によってどれほどの予測精度かを見ることにより株価とニュース記事との相関があるかを見る。精度が高ければ高いほど相関性があると言える。

4. 実験

本章では、3 章で述べた株価予測システムを実装して実際に株価予測を行う。ここでは、実験の期間や使用したデータを述べる。

4.1 実験の期間及び使用したデータ

今回の実験の期間及び実験で使用したデータについて述べる。今回の研究では、日本国内のインターネット上のテキストデータが解析対象となっている。今回は、すでに述べたように 4 つのニュースサイトの記事を Yahoo ニュース (<https://news.yahoo.co.jp/>) から取得する。表 1 に今回の実験で使用したニュースサイトと取得してきた URL を示している。実験を行う前に予め上記のサイトからニュース記事のデータを蓄積させておいたこれらのテキストデータを使用して解析を行う。また日経平均株価も Yahoo ファイナンス (<https://finance.yahoo.co.jp/>) から取得しておいた。これらの 2 つのデータを学習期間と予測期間の 2 種類の期間を設けて取得する。それぞれの取得期間は表 2 のとおりとなっている。

表 2 データの取得期間

学習期間	80 日間
予測期間	47 日間

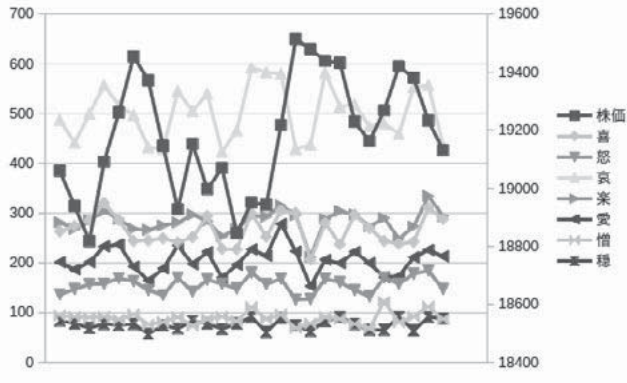


図 7 4つのニュース記事の感情解析の合計と日経平均株価の推移

表 4 相関係数の結果

パラメータ	相関係数
朝日新聞 [喜]	0.28
産経新聞 [憎]	-0.19
4 社合計 [憎]	-0.22

4.2 実験方法

実験方法について述べる。今回は、3章で述べた株価予測の方法に沿って行い実際に株価予測を行う。前節で述べた期間から4社のニュース記事と日経平均株価を取得しデータを蓄積していく。最終的に蓄積させたデータのグラフと株価予測での正解率を算出して関係性があるかを確認していく。また相関係数を求めて良い結果である3つのパラメータのみを用いた株価予測も行った。

5. 実験結果

本章では、4章で述べた実験条件を元に行った実験の結果を示す。まず、ニュースサイトを感情解析した結果を示す。全期間で取得した4つのニュースサイトの感情解析をまとめた結果をグラフを図7に示す。グラフには、8つのデータの時系列変移が示されている。

また各ニュースサイトのそれぞれの感情値の結果を表3に示す。相関係数を求めて良い結果である3つの感情パラメータと相関係数を表4に示す。さらに株価予測の結果を表5に示す。

6. 考察

本章では、5章で示した実験結果を元に考察を行っていく。

まず実験全体についての考察を述べる。表3を見

表 5 感情解析の結果

ニュースサイト名	
総合	46.853
朝日新聞	39.459
毎日新聞	54.136
読売新聞	42.572
産経新聞	44.716
相関係数を利用した株価予測	正解率 (%)
高い相関性を示したパラメータを利用	63.825

ると[哀しみ]の感情値が多く総合では、学習期間は40807予測期間では23679となっており、他の感情値に比べて圧倒的に多く算出されていることがわかった。これは、感情辞書内の[哀しみ]の感情が含まれる単語の割合が多い彼ではないかと考えられ、実際に感情辞書での感情が含まれる単語の割合でみると実際の解析結果時の割合と概ね同じとなっていた。

また学習期間と予測期間の割合とそれぞれの期間での解析結果は、概ね等しいかった。これは、ニュースサイトで掲載されるニュース記事の量が等しいことが挙げられる。またニュースサイトによって感情値の量が異なるのは、一日に掲載される量の違いであり、より多くのニュース記事が掲載されているからであり、合計で最も掲載量が多かったのは毎日新聞である。

次にニュース記事の感情解析の合計値と日経平均株価の推移のグラフを見ると感情値の推移は、全ての感情値がほぼ同じような変化をしている。これは、一日の記事の掲載量の変化によって変動するからと考えられる。また日経平均株価が急に高くなる期間があるが、感情値の量はそれほど変化していないのがわかる。直接グラフを見ても感情値と日経平均株価の相関性が明確にわかる部分はなかったと言える。

表4での相関係数の結果を見ると、最も高い相関性を示したのも朝日新聞の[喜]である、0.28であった。これは、株価との間にやや相関があると言えるが、強い相関あるとは言えない。これら以外の結果もほとんど相関性があると言えるような結果ではなかった。次に表5の株価予測の結果を見る。なお今回の結果に正解率は、それぞれ1000回の試行回数を繰り返して行った平均の正解率である。感情解析の結果は、ほとんどのニュースサイトが50%を下回る結果となった。しかしながら相関係数の評価が高かった3つのパラメータを用いた株価予測の結果は最も良い結果となった。これにより相関係数が良ければ株価

表3 感情解析の結果

ニュースサイト名	期間	喜	怒	哀	楽	愛	憎	穏
総合	学習期間	21802	12405	40807	23012	15937	6952	6209
	予測期間	12810	7297	23679	13310	8939	3867	3397
朝日新聞	学習期間	5897	3170	10784	6316	4143	1769	1688
	予測期間	3581	1898	6468	3693	2426	1027	893
毎日新聞	学習期間	6488	3778	12267	6731	4839	2095	1800
	予測期間	3809	2183	7152	3908	2603	1172	1007
読売新聞	学習期間	2015	1357	4734	2480	1568	771	666
	予測期間	1222	812	2804	1518	861	389	362
産経新聞	学習期間	7402	4100	13022	7485	5387	2317	2055
	予測期間	4198	2404	7255	4191	3049	1279	1135

予測の結果が良くなることがわかる。今回の結果だけを見れば、関係性があるのではないかと考えられる。しかしながら結果としては、50%を下回っているのでこの原因を探る必要がある。下回ってしまった原因として考えられるのは、まず第一に相関係数が全体的に低かったことが挙げられる。実験から相関係数が高ければ予測結果が良くなることがわかる。つまり相関係数をより高まれば株価予測の精度は上がるのではないかと考えられる。相関係数を上げるためには、ニュース記事のカテゴリ分けや感情辞書の変更または、感情値以外の数値化を行う必要があるのではないかと考えられる。相関係数を高めてから再実験を行う必要があると考えられる。

7. まとめと今後の課題

本研究では、インターネット上のテキストデータと株価の変動の間に相関性があるのかを調査した。先行研究によってインターネット上のテキストデータと株価の間には何かしらの関係性があると示された。しかしながらこの先行研究は、日本国内の研究ではなくアメリカの研究であり、使用したテキストデータの言語が英語でさらにインターネットの使用目的が異なっているため日本でも同様の結果となるとは言いがたい。実際に日本国内の株価予測の研究でこれほど精度の高い正解率を出せる研究はほとんどない。そこで本研究では、日本国内でのインターネット上のテキストデータに対して感情解析を行い、この感情解析の結果を元に日経平均株価の変動を予測することでを目的として、日本国内でのインターネット上のテキストデータと株価の変動の間に相関性の調査を行った。実際の研究では、日本国内でのニュースサイトのニュース記事をテキストデータとして取得し、ニュース記事に対して感情解析を行い株価予測を行っ

た。この実験の結果、ニュース記事と日経平均株価の推移に明確な関係性があるとは言いがたい結果となったが、最も良い正解率は、相関係数で高い評価を得たパラメータでの株価予測で63.825%となった。このことから相関係数が高ければ株価予測の精度が上がるということがわかった。今後の展望として、株価予測の正解率をより高めるために6章で考察した相関係数を良くする手法を実際に行い再実験を行う必要がある。

参考文献

- [1] 佐藤謙太, 小高知宏, 黒岩丈介, 白井治彦, テキスト解析による株価変動予測の試み, 平成26年度電気関係学会北陸支部連合大会, 2014
- [2] Xiao-Jun, Zeng Johan Bollen, Huina Mao, Twitter mood predicts the stock market, *Journal of Computational Science*, pp1-8, (2010)
- [3] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto, Applying Conditional Random Fields to Japanese Morphological Analysis, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, (2004).
- [4] Francis Bond, Timothy Baldwin, Richard Fothergill, Kiyotaka Uchimoto, Francis Bond and Timothy Baldwin and Richard Fothergill and Kiyotaka Uchimoto, *The 6th International Conference of the Global WordNet Association*, (2012).
- [5] 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, 2001-03-31