一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
NLC2016-43(2017-02)

専門家記事と機械学習に基づくWebニュースからの日経平均株価予測

一瀬　　航†　　嶋田　和孝†

† 九州工業大学大学院 情報工学府 先端情報工学専攻
〒 820–8502 福岡県飯塚市川津 680–4
E-mail: †{k_ichinose,shimada}@pluto.ai.kyutech.ac.jp

あらまし　近年，機械学習を用いたテキストマイニング手法によって，テキスト情報と市場変動の関係性を発見し，市場分析に応用する研究が増えている．また，Web ニュースは企業の株価に少なからず影響を与えており，世に存在する個人投資家がこのニュース記事を参考にしていると考えると，Web ニュースから未来の株価が予測できる可能性がある．そこで本論文では，Web ニュースを対象とし，より多くの投資家が市場の分析に用いていると考えられる指標である日経平均株価の予測を目的とする．テキストを用いた金融予測では膨大なテキスト情報を用いて機械学習を行うことが一般的である．しかし，投資家は市場に影響を与える多様な情報を自ら取捨選択し，独自の着眼点にしたがって市場の分析を行っている．本研究では，この着眼点，つまり，分析にどのような情報が必要なのかという知識を専門家の分析記事から抽出し，これにより機械学習の精度が向上するかの検証と新素性の提案を行う．
キーワード　株価予測，サポートベクターマシン，テキストマイニング，Web ニュース，専門家記事

Stock market prediction from Web news using expert articles and machine learning

Ko ICHINOSE† and Kazutaka SHIMADA†

† Kyushu Institute of Technology, Graduate School of Computer Science and Systems Engineering　Kawadu 680–4, Iidukasi, Fukuoka, 820–8502 Japan
E-mail: †{k_ichinose,shimada}@pluto.ai.kyutech.ac.jp

Abstract　The market analysis is one of the important tasks for text mining. Many researchers have proposed methods using text information for analyzing the market. In this situation, Web news has an important role to predict stock prices. In this paper, we propose a method to predict the Nikkei Stock Average, which is one of the most important stock market indexes. We extract viewpoints for analyzing web-news from analysis's articles of an expert and apply the viewpoints and a machine learning technique into the method. Then, we classify the next day into “UP” or “DOWN” by using the articles of a day. The experimental result shows the effectiveness of extracting viewpoints from expert articles.
Key words　Stock price prediction，SVM，Text mining，Web news，Expert articles

1. はじめに

近年，ネット証券の誕生によって初心者でも株の取引が簡単にできるようになり，株式投資が活発になっている．活発な投資活動は経済成長に必要な成長資金の供給を助け，経済の効率化も望める．しかし，1 年以内で新規の個人投資家の 90%以上が退場するといわれるほど株式投資は初心者にとっては厳しい世界である．経済成長のために長期安定的な投資の定着を図るためにも，株式投資初心者が大きな損失を出すことを防ぎ，投資の継続を支援することは社会的にも大きな意義がある．

投資家が市場の分析に用いる情報には，経済指標やテクニカル指標と呼ばれる数値で表される情報と，市場に対して影響を持つ人物の発言や企業の動向，事件や事故についての記事といったテキスト情報の 2 種類がある．前者の情報の分析に関しては書籍や Web サイトなどでも数多く取り扱われているが，後者の情報の分析に関しては個人の経験や感性に拠るところが大きいため分析方法を解説したものはほとんどなく，投資初心者が独学で勉強するのは容易ではない．そのため，機械でテキスト情報を用いた市場の分析ができれば初心者の支援になると考えられる．

テキストマイニング技術を応用して，テキスト情報と市場変動の関係性を発見し，市場分析に応用する研究がある [1] [2]．そのような金融テキストマイニングの対象となるテキスト情報源には Twitter，新聞，インターネット掲示板，Web ニュースなど様々なものがある．この中でも Web ニュースは配信頻度が高く，即時性に優れている．辻ら [3] の研究では Web ニュース中に出てくる複数の企業に対応した株価予測をそのニュースのテキスト情報を用いて行っている．この研究のようにテキスト情報を用いて株価を予測する研究は個別企業に焦点を当てたものが多い．個別企業の株価の予測は，当然ながらその企業の株に注目している投資家にしか恩恵はない．また，限られた Web ニュースだけで全ての企業の株価を個別に予測することは現実的ではない．

そこで本論文では，より多くの企業の株価に影響を与えていると考えられる日経平均株価の予測を目的とする．日経平均株価は，東証一部に上場している約 1900 社の中から選ばれた，日本を代表する 225 社の株価の平均である．この指標は，日本の景気を判断する上での重要な指標と考えられており，投資家がよく市場の傾向の分析に用いる指標でもある．この指標を正確に予測することは投資判断に役立つと考えられる．

テキストを用いた金融予測では膨大なテキスト情報を用いて機械学習を行うことが一般的であるが，Web 上に溢れる多様なテキスト情報を人が全て処理するのは不可能であり，投資家は市場に影響を与える多様な情報を自ら取捨選択し，独自の着眼点に従って市場の分析を行っている．要するに，重要なのは市場に影響がありそうな情報だけであり，膨大なテキスト情報の大部分が不要な情報であると捉えることができる．そのため，頑健な予測モデルを作成するためにはどの情報が市場の分析に必要なのかを判断するための着眼点が必要だと考えられる．過去に我々 [4] は，独自にフィルタを作成することで機械学習に必要なニュースと必要でないニュースを仕分け，結果的に株価の予測精度を向上させることに成功している．しかし，我々のフィルタは株の専門家ではない者がヒューリスティックで作成したものであるため，そのフィルタが本質的に適当かどうかは不明である．そこで，我々はネット上で配信されている専門家が分析した株や経済に関する記事に注目する．これらの記事は専門家の独自の視点に基づいて分析が行われたものであり，専門家という視点での分析の着眼点を多く含んでいると考えられる．そこで本研究では，この着眼点という分析に必要な情報がどのようなものなのかを専門家の分析記事から抽出し，これにより機械学習の精度が向上するかの検証を行う．あわせて，専門家の記事を使った機械学習のための新素性の提案も行い，その有効性の検証を行う．本システムのイメージを図 1 に示す．

2. 関 連 研 究

テキスト情報を用いた市場分析に関する研究として前川ら [5] の研究がある．前川らは，大量の記事データから投資家が反応している「言葉」を探索することで，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持つかどうかの検証を行った．Lee ら [6] は，企業が提出する金融報告書に注目し，報告された金融イベントに応じて企業株価の変化を予測するシステムを開発した．これにより，株価予測にテキストを利用することでテキストを用いない場合よりも相対的に 10%も精度が上がることを示した．他にも，インターネット掲示板に投稿されたメッセージの分析に基づき，そこから得られる掲示板指標と株式指標の関係性を明らかにしようという丸山ら [2] の研究や，経済の専門家や金融機関が Web 上に発行するマーケットリポートから日本国債市場の月次の動向を推定する手法を開発した和泉ら [7] の研究がある．さらに，Twitter からツイート内容に含まれる単語をベースとしたテキストの特徴量とグラフ表現した際のグラフ特徴量の 2 つを抽出することで，Twitter 上の膨大な情報の中から，経済動向の分析に有益な情報を得ることで短期的な経済動向の分析を行った迫村ら [8] の研究がある．Twitter を用いたものは他にも Bollen ら [9] の研究がある．Bollen らはツイートにおける気分に注目し，6 つの心的状態を表す指数を抽出することでダウ平均株価の予測を行っている．

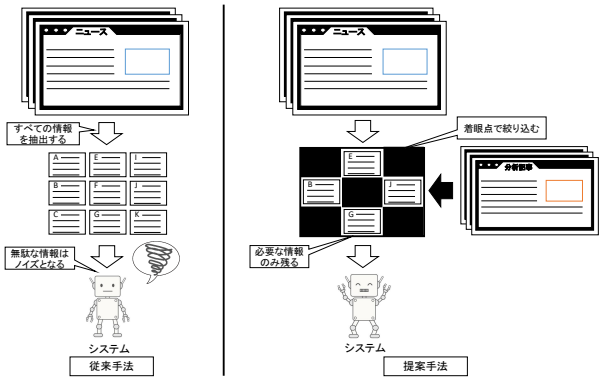

図 1　システムのイメージ図
Fig. 1　Illustration of our system

このように様々なテキスト情報を対象とした，金融テキストマイニングの研究が行われているが，専門家の書いたテキスト情報からそのノウハウや知見を抽出するような試みを行っている研究はない．そこで，本研究では Web で配信されている専門家の分析記事から市場の分析のための着眼点を抽出することによって，Web ニュースを情報源とした日経平均株価予測を行う．

3. 対象データ

本研究に用いる 2 種類のデータについて説明する．

3.1 Web ニュース

本研究では訓練事例となる実験データに Web ニュースと日経平均株価の終値を用いる．記事の獲得には Yahoo ファイナス(注1)という Web サイトを利用する．この Web サイトは複数のニュース配信サイトで配信される記事のリンクを「日本株」，「FX」，「市況・概況」，「経済総合」のように記事のカテゴリ毎に毎日まとめている．このサイトで 2015 年 1 月 5 日〜12 月 13 日の間に配信されたものを対象に，以下の 3 つの条件を満たす

(注1)：http://finance.yahoo.co.jp/

記事を収集する．
(1) 「日本株」に分類されているもの
(2) 株式市場が閉じた後に配信されたもの
(3) 平日に配信されたもの

(1) の条件を付けた理由は「株式カテゴリ」の記事が市場に一番影響を与えると考えたからであり，(2) の条件も市場が開いている間に配信された記事よりもその後に配信された記事のほうが次の日の日経平均に影響すると考えたからである．また，(3) の条件は平日にしか市場は開かないために付けた．

前　処　理

集めた Web ニュースはそのまま使用せず，前処理として以下の処理を適用する．

Step1　全 Web ニュースを配信日毎に分割する

Step2　各配信日毎の全 Web ニュースから，日経平均株価を構成する 225 銘柄の企業名と「日経平均」を含んでいる 1 文のみをそれぞれ抽出する

Step3　抽出された各文を各銘柄にまとめて 1 文書とする

この前処理を行った Web ニュースを学習データ作成の対象とする．

3.2　専門家の分析記事

本研究では市場の分析のための着眼点を専門家の分析記事から抽出する．記事の収集には Yahoo ファイナンスの株価予想(注2)のサイトを利用する．このサイトでは「日本株」,「日経平均」,「NY ダウ」,「米国ドル」などを対象として今後その対象がどのように変化するのかといった予想を様々な専門家が独自の分析を行い，不定期に配信を行っている．このサイトで 2015 年 6 月 30 日～2016 年 5 月 24 日の間に配信されたもので，「日経平均」を対象とした記事を収集する．

4.　提 案 手 法

本研究では，専門家の分析記事から株価予測のために必要な情報がどのようなものであるか，いわゆる市場の分析に必要な着眼点を抽出し，その着眼点に従って Web ニュースから学習データを作成することにより機械学習による株価予測の精度が向上するかの検証を行う．また，本研究で行う株価予測は教師あり学習をもとにしたものであり，予測の対象は翌日の日経平均株価が前日の終値と比較して上昇するか下落するかの 2 値である．この予測に直接用いる情報は前日に配信された Web ニュースのテキスト情報とする．

専門家の分析記事から着眼点を抽出する方法と，それに基づいた素性について以降で説明する．

4.1　テキスト情報を用いた素性

機械学習でニュース記事を情報源として扱うには，テキストを固定長のベクトルとして表現する素性が必要である．本研究では，図 2 のように，あらかじめ用意しておいた索引語を用いて，その索引語に登録されている単語がテキストに含まれているかどうかをベクトルで表現する方法を素性として採用する．ここでは，その一般的な手法であり本研究のベースラインとなる Bag-of-Words と本研究で提案する専門家の分析記事を用いた索引語の作成について説明する．また，以下の 4 種類の手法（1）～（4）は図 2 中の A の索引語の作成方法に該当し，（5）は B の重みに該当する．

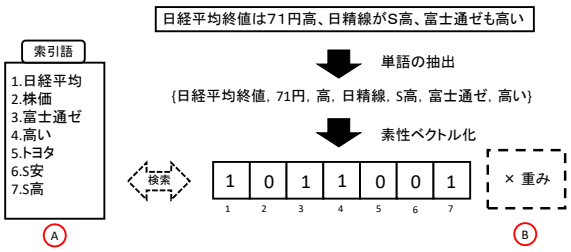

図 2　テキストデータのベクトル化の流れ
Fig. 2　Vectorization of text

(1) Bag-of-Words (BoW)

一般的にテキストをベクトル化するのによく用いられる素性である Bag-of-Words を利用する．全ての Web ニュースに対して形態素解析を行い，品詞が「名詞」,「動詞」,「形容詞」である単語から索引語の集合を作る．同じように，次は学習データ用の各 Web ニュースに対しても形態素解析を行い，品詞が「名詞」,「動詞」,「形容詞」である単語から検索語の集合を作る．そして，索引語の集合に登録されている全ての単語を一つずつ見ていき，検索語の集合にも登録されていれば 1，されていなければ 0 としてベクトルを作成する．このように，索引語の集合の大きさが作成するベクトルの大きさとなり，検索語の集合を使ってそのベクトルの各値を 1 か 0 かに決定する．なお，形態素解析器には MeCab(注3)を用いる．

(2) 専門家の分析記事を用いた Bag-of-Words (E-BoW)

この素性は (1) Bag-of-Words の索引語の作成に用いるデータを Web ニュースから専門家の分析記事に変更したものである．すなわち，専門家が使っている語彙のみを索引語に利用する．これにより，専門家の着眼点を取り入れる．具体的には，全ての専門家の分析記事に対して形態素解析を行い，品詞が「名詞」,「動詞」,「形容詞」である単語から索引語の集合を作る．次に学習データ用の各 Web ニュースに対して形態素解析を行い，品詞が「名詞」,「動詞」,「形容詞」である単語から検索語の集合を作る．以降は (1) Bag-of-Words と同様である．

(3) 専門家の分析記事を用いた Bag-of-Keywords (E-BoK)

この素性は Peng ら [10] が使用している Bag-of-Keywords (BoK) を参考にしている．彼らは分散意味表現と Bootstrap 法を利用することにより，ニュースのテキストデータから，初期値として少数の重要な単語を選択するだけで数多くの重要そうな単語を集めてくることを可能としている．彼らの BoK を応用することで，専門家の分析記事から着眼点を抽出しつつ，BoW の大きな問題点である単語スパース問題を回避することが期待できる．これを本研究に適用するために Peng らの BoK の手法の一部の変更を行う．次に，具体的なアルゴリズムについて説明する．

(注2)：http://info.finance.yahoo.co.jp/kabuyoso/

(注3)：http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

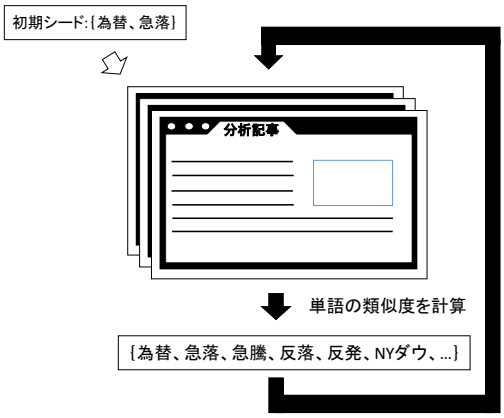

図 3　索引語獲得のイメージ
Fig. 3　Extracting of word indexes

まず，本研究では最新の分散意味表現手法の一つである Word2Vec [11] と Bootstrap に基づいて単語の選択を行う．Word2Vec のアルゴリズムは高い精度で意味表現ベクトルを作成できることで知られている．学習データとしてテキストデータを与えると，そのテキストに出現する各単語を，その周辺に現れる単語の情報を用いることで，高々数百次元のベクトルとして表現する．このように単語を意味表現ベクトルとして表すことで，単語同士の類似度を測ることができるようになる．

この Word2Vec を用いて以下のステップで E-BoK を構築する．

（1） 専門家の分析記事から Word2Vec のモデルを学習する
（2） 初期シードとなる単語を人手で選択する
（3） ニュース記事に含まれている全単語とそれぞれのシードとの意味表現ベクトルの類似度を計算する
（4） 類似度が高い各 TOP10 をシードに追加する
（5） 3 と 4 を繰り返す

反復処理を行い，最終的に得られた単語群を E-BoK の索引語とする．また，類似度の計算は単語空間内の二つの単語の類似度を表すコサイン類似度を用いる．この反復処理のイメージを図 3 に示す．

(4) Bag-of-Tuple (BoT)

BoW や BoK はニュース記事に特定の単語が出現しているかしていないかの情報しか持っていない．しかし，専門家の分析記事は分析が主な内容であることから，例えば，「海外の動きが少ないですし～」，「日経は強い動きが続いていますので～」のような文章から分析の着眼点を単語の出現だけで捉えきることは難しい．この例での抽出したい着眼点とは「海外の動きが少ない」や「強い動きが続く」といった情報であり，「続く，海外の動き，強い動き，少ない」といった個々の単語の出現情報ではない．そこで，分析に必要なのは起点となる動詞とその動作主であると仮定し，専門家の記事から分析の着眼点として「動作主+動詞」のペアを抽出して使用する Bag-of-Tuple(BoT) を提案する．次にアルゴリズムについて説明する．

まず，全ての専門家の分析記事に対して係り受け解析を行い，全ての「動作主+動詞」のペアで索引語の集合を作る．同じように，次は学習データ用の Web ニュース記事に対しても係り受け解析を行い，全ての「動作主+動詞」のペアから検索語の集合を作る．そして，索引語の集合に登録されている全てのペアを一つずつ見ていき，検索語の集合にも登録されていれば 1，されていなければ 0 としてベクトルを作成する．このように，索引語の集合の大きさが作成するベクトルの大きさとなり，検索語の集合を使ってそのベクトルの各値を 1 か 0 かに決定する．なお，係り受け解析器には CaboCha(注4) を用いる．

(5) ニュース記事の影響度 (ES)

この素性はあるニュース記事がどれくらい日経平均に影響を与えるかを測るための指標である．投資家は日々配信されるニュース記事を投資するかどうかの判断材料の一部として利用していると考えられる．そのため，日本の相場動向を測る指標として最も用いられる日経平均株価はニュース記事によって少なからずの影響を受けていると考えられる．しかし，全てのニュース記事が一様に影響度を持っているわけではなく，当然ながら記事の内容によって持っている影響度が異なるはずである．それぞれの記事がどのくらいの影響度を持っているかを調べることは難しい．そこで本研究の予測対象である日経平均株価を構成する 225 社の企業に注目する．日経平均株価を構成するこれらの企業に関する記事は比較的大きな影響度を持っていると考えることができ，特定の企業の株価に注目することで，日経平均株価との株価の連動度合を求めることができる．よって，本素性はニュース記事の日経平均株価に与える影響度を，日経平均株価と 225 銘柄の株価の相関係数を求めることによって表す．よって ES は [-1,1] の範囲の値をとる．今回は，2015 年 1 月 5 日～2015 年 12 月 30 日までの過去の株価から，日経平均株価と 225 銘柄間の相関係数をそれぞれ求める．そして，各銘柄ごとに作成される素性ベクトルに対応する銘柄の相関係数を重みとして掛けることで，素性ベクトルにニュース記事の影響度を反映させる．

4.2　分類器の作成

本研究では日経平均株価の予測を機械学習によって行う．ここでは機械学習を用いた分類器の作成について説明する．分類器には SVM [12] を，機械学習を行うツールには Weka(注5) を使用する．3. 節の Web ニュースに 4.1 節で説明した素性を適用して学習データを作成する．具体的には，4.1 節の（1），（2），（3），（4）の方法で作成した索引語を使って，3. 節の前処理が行われた後の Web ニュースをベクトル化する．Web ニュースのベクトル化のイメージを図 4 に示す．図 4 のように，各日毎に，各銘柄のニュースデータを 4.1 節のそれぞれの索引語でベクトル化し，適宜 4.1 節の（5）ES で重みづけし，得られたベクトルを単純な加算で合成することによりその日のニュースデータの素性ベクトルを 1 つ得ることができる．この素性ベクトルに教師値を付与することで学習データができる．なお，合成前に得られるベクトルはその日のニュース記事に出現する 225 銘柄の企業数に左右されるため，日によっては 1 個～226 個（0 の

(注4)：http://taku910.github.io/cabocha/
(注5)：http://www.cs.waikato.ac.nz/ml/weka/

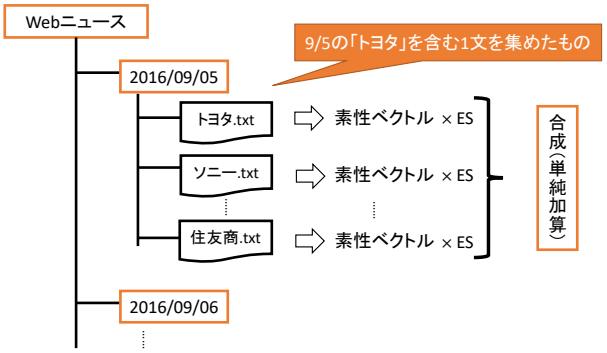

図 4　Web ニュースのベクトル化のイメージ
Fig. 4　Vectorization of web-news

表 1　専門家の分析記事の有効性の実験結果
Table 1　The results of classifiers using expert articles

分類器		適合率	再現率	F 値	正解率
（1）BoW	Minus	36.0	20.7	26.3	-
	Plus	57.4	74.4	64.8	-
	Ave	48.6	52.4	49.0	52.3
（2）E-BoW	Minus	48.4	35.6	41.1	-
	Plus	62.2	73.6	67.4	-
	Ave	**56.5**	**58.0**	**56.6**	**58.0**

時はベクトルを作らない）の差異がある．

教師値にはその記事の次の日の日経平均株価の終値が上昇した (Plus) か下落した (Minus) かの 2 値を用いる．具体的には以下のようにする．

$$\text{教師値} = \begin{cases} Plus & (\text{次の日の終値} > \text{前日の終値}) \\ Minus & (\text{次の日の終値} <= \text{前日の終値}) \end{cases}$$

この式を用いて教師値を決定する．

5. 実　　験

5.1　専門家の分析記事の有用性の検証

まずは，4.1 節の（1）Bag-of-Words (BoW) と（2）専門家の分析記事を用いた Bag-of-Words (E-BoW) の素性と 3. 節の Web ニュースで作成した学習データで機械学習を行い，分類実験を行った．評価には Leave-one-out 交差検定を用いた．その結果を表 1 に示す．評価項目は Plus クラスと Minus クラスの適合率と再現率と F 値と正解率である．F 値とは，適合率と再現率の調和平均である．また，表中の Ave は Plus と Minus の加重平均である．表 1 の結果を見てみると，ベースラインの（1）BoW の F 値と比較して（2）E-BoW は 7.6 ポイント上昇していることがわかる．またその他の全ての項目でもベースラインを上回っている．この結果から，専門家の分析記事中の単語のみを Web ニュースのベクトル化のための索引語に利用することが有効であるということがわかった．

5.2　専門家の分析記事を応用した素性の実験

次に，4.1 節の専門家の分析記事をさらに応用した素性である（3）E-BoK と（4）BoT，それと，ニュース記事の影響度を考慮した重み（5）ES を用いた分類器の評価実験を行った．評価には Leave-one-out 交差検定を用いた．また，（3）E-BoK の索引語を作成する条件として，今回は初期シードとして“リスク，経済指標，急落，シグナル，売る，買う，底堅い，上昇トレンド，下降トレンド，期待感”の 10 単語を人手で選択して与え，最終的に 1202 単語を索引語として獲得した．実験結果を表 2 に示す．まず，専門家記事を用いた単純な BoW である（2）E-BoW をベースラインとして F 値で比較すると，（3）E-BoK は 0.5 ポイント，（4）BoT は 1.2 ポイント上昇している．この結果から，（3）と (4) はより効果的に専門家の分析記事から分析の着眼点を抽出できていることがわかる．

次に，（5）ES の重み付けによる精度の変化を比較する．表の結果から分かるように，重み付けを行った結果，重み付けを行う前より F 値が（2）E-BoW は 6.2 ポイント，（3）E-BoK は 1.6 ポイント下落している．一方で，（4）BoT は 1.3 ポイント上昇している．このことから，ニュース記事の影響度による重み付けはどんな素性でも精度が向上するわけではないが，一部の素性では精度向上の効果があることがわかった．

この実験では（5）ES の全ての値を重みとして利用したが，ES は相関係数であることから，0 に近い値ほどニュース記事は日経平均株価に影響を与えないということである．このため，ES が低い値のニュースデータを学習データに用いるのは好ましくない可能性がある．そこで，使用する ES に閾値を設定し，（4）BoT ＋ （5）ES の分類器の精度がどのように変化するかを実験してみた．その結果を図 5 に示す．図の横軸が ES の絶対値の閾値であり，例えば，閾値が 0.1 ならば各銘柄の相関係数の絶対値が 0.1 以下となる銘柄に関するニュース記事を学習データに用いない．そして，縦軸がその閾値で学習を行った場合の分類器の F 値を表している．また，黄色のグラフはその閾値を設定したときに学習データに使用するデータとして残った銘柄の累計を表している．例えば，閾値が 0 の時は 212 日間で累計 11,002 銘柄，1 日平均 52 銘柄を使用して学習データを作成している．この図から分かるように，閾値を大きくしていくほど F 値は右肩下がりになっていき，同時に使用銘柄数も右肩下がりになっていることがわかる．特に，閾値が 0.9 の場合に使用するデータ数は最大数の 4.8%程にまで激減しているため，健全な学習が行われているとは言い難い．この結果から，ES は閾値を設けずに全てを重みとして使用するのが一番妥当であるということが分かった．

5.3　今後の課題

5.1 節の実験結果から，専門家の分析記事から Web ニュースの分析のための着眼点を抽出することの有効性を検証することができた．また，5.2 節の実験によって，より効果的に分析の着眼点を捉えることができるいくつかの分類器を作成することができた．この節ではこれらの結果を踏まえて，分類器の予測精度をさらに向上させるための手法について考える．

まず考えられるのが 5.2 節のいくつかの分類器を組み合わせることである．それぞれの分類器の特徴を考慮して，予測しやすい日や失敗しやすい日などの傾向を把握することができれば，互いの長所をいかした予測システムを構築することができる可能性がある．ここで，組み合わせとして考えられるのが最も分

表 2　専門家の分析記事を応用した素性の実験結果

Table 2　The result of classifiers using new features applied expert articles

分類器		適合率	再現率	F 値	正解率
(2) E-BoW	Minus	48.4	35.6	41.1	-
	Plus	62.2	73.6	67.4	-
	Ave	56.5	58.0	56.6	58.0
(3) E-BoK	Minus	48.1	42.5	45.1	-
	Plus	63.0	68.0	65.4	-
	Ave	56.8	57.5	57.1	57.5
(4) BoT	Minus	49.3	41.4	45.0	-
	Plus	63.3	70.4	66.7	-
	Ave	**57.6**	**58.5**	**57.8**	**58.5**
(2) E-BoW+(5) ES	Minus	38.6	25.3	30.6	-
	Plus	58.1	72.0	64.3	-
	Ave	50.1	52.8	50.4	52.8
(3) E-BoK+(5) ES	Minus	45.8	43.7	44.7	-
	Plus	62.0	64.0	63.0	-
	Ave	55.4	55.7	55.5	55.7
(4) BoT+(5) ES	Minus	51.4	42.5	46.5	-
	Plus	64.3	72.0	67.9	-
	Ave	**59.0**	**59.9**	**59.1**	**59.9**

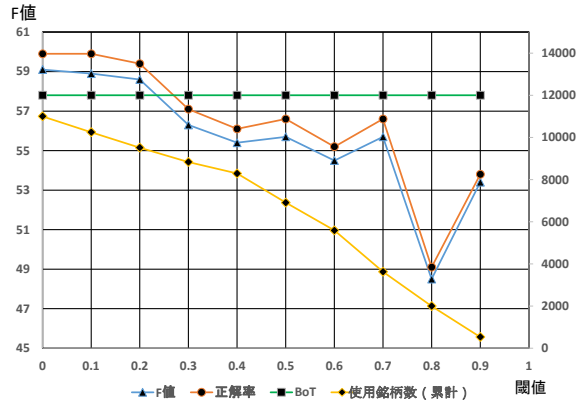

図 5　ES の閾値による精度の変化

Fig. 5　The changes of values by threshold level for ES

類精度が高い（4）BoT+（5）ES （BoT-ES）と 3 番目に高い（3）E-BoK である．2 番目に高い（4）BoT はその素性の構成が BoT-ES と非常に似ているため組み合わせても意味がないことが予想できる．同じ理由でその他の組み合わせも最良とならない．次に，組み合わせた場合のシステムに精度向上の見込みがあるのかを検証するために E-BoK の分類器と BoT-ES の分類器の予測が正解する日にどのような差異があるのかを調べた．その結果を表 3 に示す．表の 1，2，3 はそれぞれ，E-BoK と BoT-ES が共に正解した日の数，E-BoK と BoT-ES のどちらかまたはどちらもが正解した日の数，E-BoK と BoT-ES のどちらか一方のみ正解した日の数を表している．表の 1 と 2 から，二つの分類器は正解できる日が 38.2%ほどしかかぶっておらず，適切にどちらかを選択することができれば正解率の上限が 78.8%の予測システムを作ることができることを示している．また，このようなシステムを作るためには，二つの分類器のうちどちらか一方しか正解できない 86 日のデータに二つを区別できるような傾向があるのかなどを分析していく必要がある．

表 3　正解日の差異

Table 3　The difference of the day that classifiers predict correctly

No.	演算	正解日の数	割合 (正解日/212 日)
1	E-BoK ∩ BoT	81 日	38.2%
2	E-BoK ∪ BoT	167 日	**78.8%**
3	E-BoK ⊕ BoT	86 日	40.6%

6. おわりに

本研究では，専門家の分析記事を利用して Web ニュースの分析のための着眼点を抽出することで機械学習による日経平均株価予測の精度が向上するかの実験を行った．また，専門家の分析記事からより効率的に着眼点を抽出するための素性をいくつか提案し，実際に分類器を作成して実験することでより精度の良い分類器が作れることを確かめた．

今後の方針としては，予測の精度をさらに上げるために，今回作成して精度の良かった分類器を組み合わせたシステムの開発を目指す．そのためにも組み合わせる分類器の分析を行い，特徴や予測が正解する日にどんな差異や傾向があるのかを重点的に調べていく必要がある．

文　献

[1] 高橋宏圭，関和広，上原邦昭，“株価回帰と Web ニュース記事分析を組み合わせた株価動向推定”，電子情報通信学会，信学技報，pp. 103-108，2012.

[2] 丸山健，梅原英一，諏訪博彦，太田敏澄，“インターネット株式掲示板の投稿内容と株式市場の関連性”，金融情報学研究会，第 2 回研究会，2013.

[3] 辻洋平，古宮嘉那子，小谷善行，“Web ニュース中の複数企業に対応した株価予測”，電子情報通信学会，信学技報，pp. 109-113，2011.

[4] 一瀬航，嶋田和孝，“フィルタリングと機械学習に基づく Web ニュースからの日経平均株価予測”，電子情報通信学会，信学技報，Vol. 115，No. 70，2015，pp. 91-96

[5] 前川浩基，中原孝信，岡田克彦，羽室行信，“大規模ニュース記事からの極性付き評価表現の抽出と株価収益率の予測”，日本オペレーションズ・リサーチ学会，58(5)，pp. 281-288，2013.

[6] Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, Dan Jurafsky, “On the Importance of Text Analysis for Stock Price Prediction”, Proceedings of the Ninth International Conference on Language Resources and Evaluation, 2014

[7] 和泉潔，後藤卓，松井藤五郎，“経済リポートのテキスト分析による金融市場動向推定”，電子情報通信学会，信学技報，pp. 107-111，2011.

[8] 迫村光秋，和泉潔，“twitter テキストマイニングによる経済動向分析”，金融情報学研究会，第 9 回研究会，2013.

[9] Bollen Johan, Huina Mao, Xiao-Jun Zeng, “Twitter mood predicts the stock market”, Journal of Computational Science，2(1), pp. 1-8，2011

[10] Yangtuo Peng, Hui Jiang, ”Leverage Financial News to Predict Stock Price Movements Using Word Embeddings and Deep Neural Newtworks”, NAACL-HLT 2016, pp. 374-379.

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, ” Distributed representations of words and phrases and their compositionality”, NIPS 2013, pp. 3111-3119.

[12] Vladimir Vapnik, “The Nature of Statistical Learning Theory”, Springer-Verlag, 1995.