アナリストレポートを用いた中長期株価動向推定

Stock Price Analysis in Mid-to-long Term Using Analyst Reports

堅木 聖也 [1*]　坂地 泰紀 [1]　和泉 潔 [1]　石川 康 [2]　笠岡 恒平 [2]
Toshiya Katagi[1]　Hiroki Sakaji[1]　Kiyoshi Izumi[1]　Yasushi Ishikawa[2]　Kohei Kasaoka[2]

[1] 東京大学大学院工学系研究科
[1] School of Engineering, The University of Tokyo
[2] 日興アセットマネジメント株式会社
[2] Nikko Asset Management Co., Ltd.

Abstract: In this paper, we propose a methodology to forecast the direction and extent of volatility in mid-to-long term excess return of stock price by applying natural language processing and neural networks on the context of analyst reports. Analyst reports are prepared by analysts in research department in stock brokerage firms and we consider the content of reports include usefull information to forecast movements in stock prices. First, our method extracts opinion sentences from analyst reports, while the remaining parts correspond to non-opinion sentences. Second, our method predicts stock price movements by inputting opinion sentences and non-opinion sentences to neural networks separately.

1　はじめに

現在，日本における株式投資の重要性が拡大している．日本取引所グループ (JPX) の調査レポート [日本 18] によると，日本における個人株主数は右肩上がりで増加しており，特に 2017 年度の個人株主数は，前年度比 162 万人増加して 5,129 万人となり，初めて 5 千万人を超えることとなった．

この流れは更に加速すると考えられる．直近ではアベノミクスの効果，及び 2020 年に予定されているオリンピックの効果により，多くの企業の株価は増加傾向にあり，実際，東京証券取引所の第一部上場銘柄を対象にした指数である，東証株価指数 (TOPIX) も右肩上がりとなっている．その結果として，日本株の魅力度は増大している．

投資にあたって，投資家達は，対象企業に対するあらゆる情報を調べることになる．しかし，その情報源は多角化しており，情報収集は困難を極める．企業のホームページを見ると，IR(Investor Relations) のページには決算短信，決算説明会資料，アニュアルレポート，有価証券報告書といった資料が乱列しており，検索エンジンで社名を検索すれば，多くのニュースがヒットし，株価の掲示板を見ると，過去の株価推移や投資家達の意見が存在する．さらに近年では Twitter や Facebook，Instagram といった SNS(ソーシャルネットワークサービス) における人々のコメントも投資家のセンチメントを反映していると言えよう．実際，迫村ら [迫村 13] の研究では，SNS の 1 つである Twitter の投稿における，強気比率や偏りが，その後の株式リターンに反映されている可能性が示されている．

このように，情報化が進んだ昨今では，株式投資という行為一つを取っても，参照すべき情報が溢れており，各投資家達にとっては，その取捨選択が困難を極めている．この状況下でより注目を集めているのがアナリストレポートである．アナリストレポートはその名前の通りアナリストが個別企業に対して書くレポートであり，金融のプロフェッショナルが，公表資料やニュース，会社取材，株価バリュエーション，マクロ経済動向などを全て考慮した上で，事実関係を整理し，自身の評価を下したものであるため，各情報源の上位互換とみなすことができるであろう．そこで本研究では，アナリストレポートの本文を解析し，株価動向を予測することを目指す．特に，株価の動向において特に重要となる，株価のボラティリティの大小及び株価の対市場超過リターンの符号の予測を今回は行う．更に，本文の中でも，文章を 2 通りに分類することを考える．アナリストレポートには，アナリストが集約した客観事実を示す文章 (非意見文) と，それに基づいてアナリスト自身がどう考えたのかを表す文章 (意見文) が存在する．これらの内，意見文は市場が瞬時には織り込み難いアナリストの意思を反映するため，株価ボラティリティの大小に対する予見性を持ち，残った非

*連絡先：東京大学大学院工学系研究科
E-mail: m2017tkatagi@socsim.org

意見文は客観事実の中から株価形成に重要だとアナリストが判断した事象を多く含むため，超過リターンの正負に対する予見性を持つという仮説を持った．この仮説を検証し，アナリストレポートと実際の株価変動の関係性を示すことを本研究の目的とする．

加えて，リターンの方向に関しては，Bollen ら [Bollen 11] の先行研究において，極性値の情報を利用することで成果を挙げていることから，極性値の情報をも用いることで更なる性能の向上を目指した．

そして本研究は，実際に株式投資を行われている資産運用会社との共同研究となっているため，将来的に資産運用への応用を目指すことを念頭に置く．

2　全体の流れとデータセットの作成

本研究では，意見文の判別，株価動向の推定，の 2 段階で実験を行った．

2.1　意見文の判別のデータセットの作成

2017 年に発行された 10,100 本のアナリストレポートからランダムに 100 本を抽出した．さらにこの中からボディに相当する部分のみを抽出した．合計 2,213 文に対して，手動で意見文/非意見文のラベリングを行った．意見文はレーティングや企業の売上や純利益の次年度の予測値といったアナリスト自身の予測や，今後企業が取るべき施策，現在の業績となった背景などといった内容が含まれている文を指しているのに対し，非意見文は企業の過去の業績値といった事実に関する文を指している．その結果，意見文と判断されたものは 1,025 文となった．

2.2　株価動向の推定のデータセットの作成

2017 年に発行されたアナリストレポートの各レポートの発行日，本文を取得した．また，発行日と発行日から 2 週間後の株価及び，TOPIX を取得し，この期間での TOPIX に対する超過リターンを得た．超過リターンを採用したのは，2017 年が長期化した景気回復局面にあり，単純なリターンを用いると正のリターンに分布が偏ってしまうからだ．さらにベンチマークに対する相対パフォーマンスにより運用能力を評価される機関投資家にとっては，対市場超過リターンに対する予見性が重要であるという背景もある．分類の結果，2,041 個のデータを入手した．超過リターンの正負の予測のために，超過リターンの正のものに 1，負のものに 0 とラベリングを行った．また，ボラティリティの予測のために，超過リターンの絶対値が 2.745%より小さいものに 0，大きいものに，0 とラベリングを行った．この 2.745%は，レポートがちょうど半分に分割される閾値となっている．

3　意見文の判別

第 1 段階である意見文の判別に関しては，MLP(Multilayer Perceptron)，Bidirectional LSTM(Bidirectional Long short-term memory)，SVM(Support Vector Machine)，ランダムフォレストを用いて実験を行った．

MLP， SVM，ランダムフォレストについては，使用するアナリストレポートに含まれる単語を One hot 表現したベクトルを入力データとした．利用されている単語は 5,126 個であるので，ベクトルの次元も 5,126 次元となった．

また，Bidirectional LSTM への入力データは，Word2vec を用いて作成したベクトルとした．コーパスはアナリストレポートを使用した．その結果，金融に特有な単語を意味を保持したままベクトル化することができた．Bidirectional LSTM 層から得られた隠れ状態ベクトルは MLP 層へと伝播され，最終層でソフトマックス関数を用いて，意見文/非意見文となる確率を出力させた．モデルの様子を図 1 に示した．

それぞれの手法による結果は，表 1 のようになった．なお，結果においては意見文であるものを正としている．表 1 において，Pre とは Precision を意味する．表

表 1: MLP と LSTM による意見文の判別の結果

手法	Pre	Recall	F1
MLP	0.733	0.827	0.777
Bidirectional LSTM	0.846	0.772	0.808
SVM	0.790	0.618	0.694
ランダムフォレスト	0.757	0.643	0.695

1 から，Bidirectional LSTM が最も F1 が高いことが分かる．この理由は，入力データにあると考えられる．今回は単語の分散表現を入力させたが，単語の分散表現はすでにアナリストレポートを使って事前学習させたものとなっている．すなわち，アナリストレポートに出現する単語の順や位置関係に関わる情報がこのモデルの学習前からインプットされていることになる．

以後の実験において，意見文を抽出するものに関しては，この Bidirectional LSTM の学習モデルを使って抽出することとした．

4　株価動向の推定

株式投資を行う際に，投資家が意識するポイントは 2 つある．1 つ目は超過リターンの方向が正か負のどちら

かであるかという点であり．もう 1 点は，株価の変動が大きいかどうかという点である．これらの情報を組み合わせることで．各銘柄の動きをアナリストレポートのレーティングのように，Strong buy，Buy, Sell，Strong Sell と分類でき，株の取引に応用されることが期待される．そこで，本研究では，この 2 点，超過リターンの正負，超過リターンのボラティリティの大小に対する判別を行った．

手法としては，意見文の判別において成果を上げた Bidirectional LSTM を用いた．Bidirectional LSTM への入力の仕方については，意見文と非意見文がそれぞれボラティリティと方向性に寄与するのではないかという仮説の下，以下の 4 つを試行した．

- 全文を入力
- 意見文のみを入力
- 非意見文のみを入力
- 意見文と非意見文を分割した後，別々に入力

全文を入力，意見文のみを入力，非意見文のみを入力の 3 つの入力方法に関しては，図 2 のネットワークを用いた．

また，意見文と非意見文を分割した後，別々に入力する場合に関しては，図 3 のネットワークを用いた．上記 3 つと異なり，2 つの Bidirectional LSTM を用いているため，隠れ状態ベクトルの次元は合計 160 次元として，MLP 層に入力した．

また，リターンの方向性に関しては，極性値の利用を検討した．極性値の取得方法についていは，東京工業大学高村大也教授の PN Table[1]及び，金融に特化した極性辞書 [伊藤 17] の利用を検討した．利用辞書選定のために，意見文の判別の際に用いた，2,213 文に含まれる意見文に手動で，Positive かどうかをラベリングし，辞書によって判別させた．判別は文に含まれる単語の極性を辞書を参照し獲得した後，平均をとることで行った．その際，極性値が正のものを Positive，負のものを Negative と定義した．これらから Positive を正として，Precision，Recall，F1 を計算すると表 2 のようになった．これより，3 指標においてスコアが勝った

表 2: 極性辞書を使用した極性判別の結果

使用辞書	Pre	Recall	F1
一般辞書	0.556	0.009	0.018
金融特化辞書	0.640	0.739	0.686

金融特化辞書を採用することにした．このような差が

[1] http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html

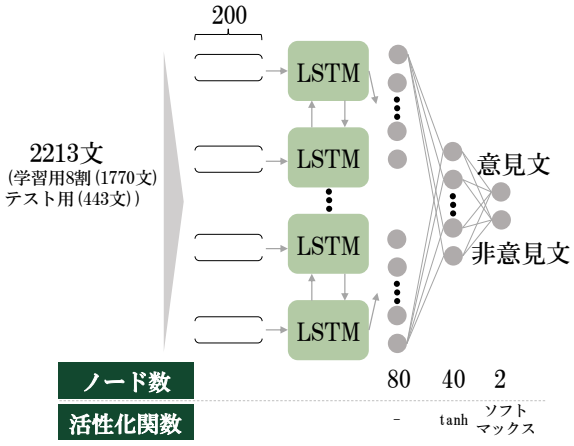

図 1: 意見文判別のモデル

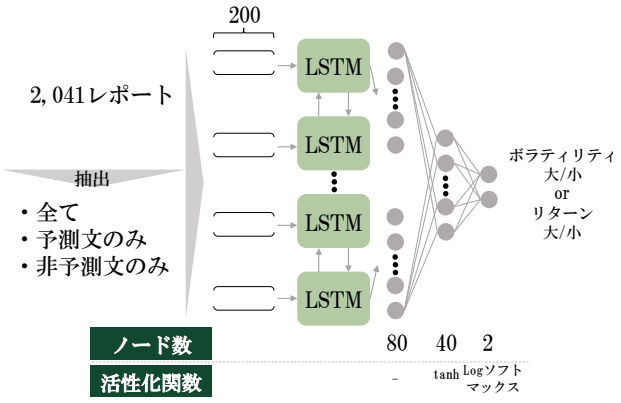

図 2: 全文・意見文のみ・非意見文のみを入力する場合のモデル

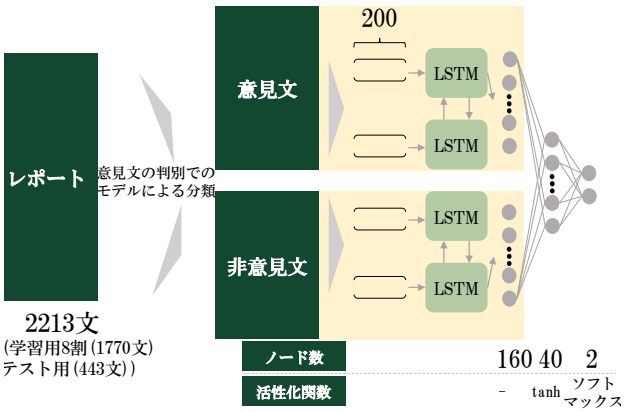

図 3: 意見文と非意見文を分割した後，別々に入力する場合のモデル

出た理由は，そもそも一般辞書の極性が負に偏っていたことや，金融特化辞書では「自社株買い」のような一般的ではないが，極性に大きく影響する単語までスコアが振られていたことが挙げられる．金融特化辞書により獲得された極性値の入力方法は，図 4 に示した 2 種類を検討した．一つは左図に示した，文章の極性値を入力する方法である．文章に含まれる各単語の極性値を平均した後に，MLP 層に入力した．もう一方は右図に示した，各単語の極性値を入力する方法である．各単語の極性値を入力として用いている Word2vec のベクトルに追加した後に Bidirectional LSTM 層に入力した．

さらに，今回はアナリストレポートの本文という同一の入力から，超過リターンの方向性とボラティリティという複数の対象を予測することから，マルチタスク学習 [Caruana 97] の利用を試みた．図 5 に示したように，同一の入力とネットワークにより，同時に超過リターンの方向とボラティリティを推定することを目指した．なお，極性値を用いた推定とマルチタスク学習を用いた推定では，入力テキストは文章全体とした．

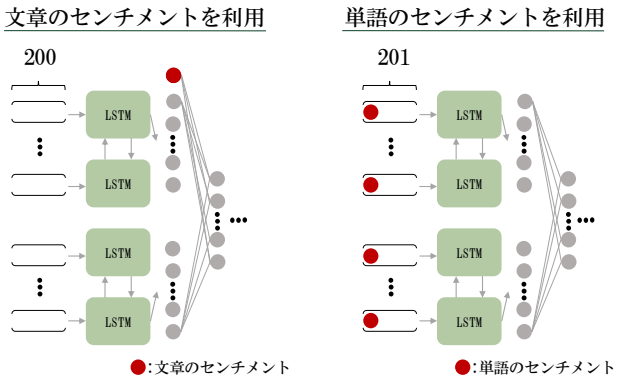

図 4: センチメントの Bidirectional LSTM への入力方法

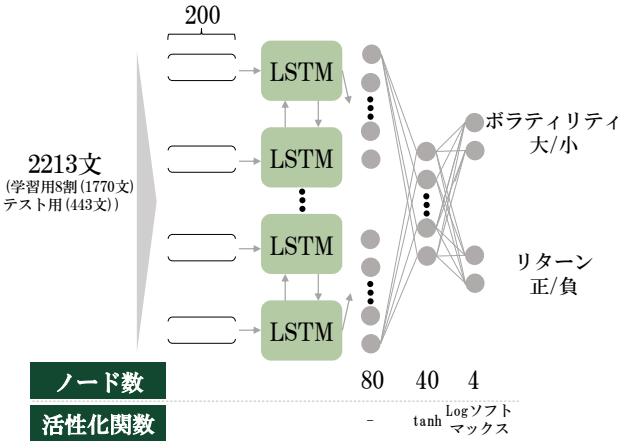

図 5: マルチタスク学習のネットワーク

5 結果と考察

以下，各実験において，リターンが正の時，ボラティリティが大の時を正とした場合の，各評価指標を示した．

5.1 超過リターンの正負の推定

これより，F1 を比較すると，非意見文のみを入力→意見文と非意見文を別々に入力→全文→意見文のみを入力，の順で性能が良いことが分かる．

5.2 超過リターンのボラティリティの推定

これより，F1 を比較すると，意見文のみを入力→意見文と非意見文を別々に入力→全文→非意見文のみを入力，の順で性能が良いことが分かる．

以上 2 つのリターンの方向性の実験とボラティリティの大きさの実験から，リターンの方向性の推定には非意見文が，ボラティリティの大きさの推定には意見文が重要となっていることが確認された．このような結果になった理由として，市場の投資家の投資スタイルの観点から検討をする．株のトレーディングにおいては，リスクを取り，短期間で売買を行う短期トレーディングと，リスクを減らし，長期間でトレーディングを行う長期トレーディングが存在する．前者を行う投資家は比較的少額の金額を日々取引していくので，日々の株価のトレンドを形成していくと考えられる．後者は，資産運用会社が取る手法であり，取引頻度は小さ

表 3: 超過リターンの方向の判別の評価指標

入力の種類	Pre	Recall	F1
全文	0.564	0.591	0.577
予測文のみ	0.490	0.517	0.503
非予測文のみ	0.614	0.659	0.636
予測文と非予測文	0.589	0.603	0.596

表 4: 極性辞書を使用したボラティリティ判別の結果

入力の種類	Pre	Recall	F1
全文	0.552	0.520	0.538
意見文のみ	0.567	0.587	0.577
非意見文のみ	0.532	0.509	0.520
意見文と非意見文	0.553	0.553	0.553

いが，取引額が大きいため，市場にインパクトを与えやすいと考えられる．

ここで，それぞれの投資家が参考にする情報を考える．短期トレーディングを行う投資家は，短期保有を前提とするため，将来的な業績予測よりも，すでに公開されている事実や，直近での株価の変動の様子を参考に投資判断を下すと考えられる．一方で長期トレーディングを行う投資家は，長期的な保有を前提とするため，まだ確定しない将来の業績に目を向ける．そのような情報が盛り込まれているのがアナリストレポートであるが，アナリストレポートは膨大な量が発行されているため，全てを一言一句読み込むということは時間的に難しい．実際の運用担当者は，アナリストにより違いが出易い意見文のみを精読する場合も多い．これらをまとめると，日々のトレンドを作っていく短期トレーディングを行う投資家は既に公開された事実を考慮し取引を行うのに対し，取引額が大きいため株価の変動を引き起こしやすい長期トレーディングを行う投資家はアナリストの意見文のように市場に十分織り込まれていない情報を考慮し取引を行うことになる．そのために，超過リターンのボラティリティの大小はアナリストレポートの意見文の情報から予測され，超過リターンの方向性はアナリストレポートの非意見文の情報から予測されたのだと考えられる．

5.3 極性値を用いた推定

表 5: 極性値を利用したリターンの方向性推定の評価指標

使用する極性辞書	Pre	Recall	F1
文章全体	0.573	0.607	0.590
単語	0.556	0.584	0.570

極性値を用いない場合の結果である表 3 の全文と比較すると，文章全体の極性値を利用した場合では 3 つの指標全てにおいて改善がみられ，一方で単語のセンチメントを利用した場合では，3 つの指標全てにおいて悪化していることが分かる．

単語の極性値を利用した場合で悪化した理由として，単語の辞書に含まれていない単語の存在があげられる．今回の実験で使ったレポートに含まれる単語数は 693,562 単語であった．この単語において，辞書に含まれる単語に関しては，そのセンチメントを入力ベクトルの極性値用の成分に代入した．一方で，辞書に含まれていない単語の該当成分には 0 を代入した．693,562 単語の内，辞書に含まれていなかったものは 262,508 単語存在した．すなわち 40%程の単語の極性値が 0 となってしまっている．加えて，極性値は単語の分散表現ベクトルに追加しただけであり，201 次元の中の 1 次元のみしかしめない．これらの理由で，LSTM への全体の入力データに占める極性値の要素が非常に小さく，学習の際に極性値に適切な重みがつけられなかったと考えられる．

5.4 マルチタスク学習を用いた推定

表 6: マルチタスク学習を利用した超過リターンの推定の評価指標

	Pre	Recall	F1
マルチタスク学習	0.569	0.543	0.556

このようにマルチタスク学習では，リターンの方向性，ボラティリティの大きさともに改善させることができなかった．これは，この 2 つの情報に相関がないためではないかと考えられる．マルチタスク学習では，複数の学習対象のパラメータを共有することで性能の向上が期待される．学習対象が同じような特徴量を重要視するのであれば，複数の学習によってその特徴量がより抽出されるであろうが，そうでない場合はそれぞれの学習を邪魔してしまうであろう．今回の目標である超過リターンの方向とボラティリティは一見すると似た学習対象ではあるが，上述の通り，それぞれ非意見文と意見文といった相異なるセンテンスと深く紐づいている．このことから，これら 2 つを決定づける特徴量は明確に異なるものであるとの推察ができる．そのため，マルチタスク学習はうまく機能しなかったのではないかと考えられる．

6 まとめ

本研究では，株式投資を見据え，投資家が重要視する個別銘柄の TOPIX に対する中長期の超過リターンの方向，及びボラティリティの大きさをアナリストレポートの本文を用いて推定することを目指した．アナリストレポートを構成する各センテンスをアナリストの意見文，非意見文に分類した後にこれらの推定を行った結果，超過リターンの方向は非予測文を，ボラティリティの大きさは意見文を用いることで高い性能で推定できることが判明した．なお，分類に用いる手法は，時系列データの学習に強みを持つニューラルネットワークである，Bidirectional LSTM に優位性が見られた．株価動向の予測の向上を目指し，近年金融情報学の分野で注目を浴びているセンチメントの利用と，汎化性能の向上が期待されるマルチタスク学習の利用を検討

した．センチメントの利用に関しては，文章全体センチメントを利用する方法と単語のセンチメントを利用する方法を検討したが，欠損値が生じない前者では性能が向上し，本研究でもセンチメントの有用性が確認できた．一方でマルチタスク学習については，学習対象に強い相関がないためか，性能を向上させることはできなかった．

参考文献

[Bollen 11] Bollen, J., Mao, H., and Zeng, X.: Twitter mood predicts the stock market, *Journal of computational science*, Vol. 2, No. 1, pp. 1–8 (2011)

[Caruana 97] Caruana, R.: Multitask learning, *Machine learning*, Vol. 28, No. 1, pp. 41–75 (1997)

[伊藤 17] 伊藤友貴, 坪内孝太, 山下達雄, 和泉潔 F テキスト情報から生成された極性辞書を用いた市場動向分析, 人工知能学会全国大会論文集 2017 年度人工知能学会全国大会 (第 31 回) 論文集, pp. 2D21–2D21 一般社団法人 人工知能学会 (2017)

[日本 18] 日本取引所グループ F 日本取引所グループ調査レポート, *https://www.jpx.co.jp/markets/statistics-equities/examination/01.html* (2018)

[迫村 13] 迫村光秋, 和泉潔 FTwitter テキストマイニングによる経済動向分析, 第 9 回人工知能学会 ファイナンスにおける人工知能応用研究会 発表論文 (2013)