テキストマイニングによる アナリストレポートを用いた株価動向予測

鈴木 雅弘 *1　堅木 聖也 *2　坂地 泰紀 *2　和泉 潔 *2　石川 康 *3
*1 東京大学工学部, *2 東京大学大学院工学系研究科
*3 日興アセットマネジメント株式会社
{msuzuki@g.ecc., {sakaji, izumi}@sys.t.}u-tokyo.ac.jp
m2017tkatagi@socsim.org, yasushi.ishikawa@nikkoam.com

1　はじめに

東京証券取引所の調査によると[1]，近年の日本における個人投資家の数が増加している．投資家は投資のために株価だけでなく企業の売上高，収益，経営状況や為替など多くの情報を調べる必要がある．しかしながら，情報源は多様化し，また全てについての情報を収集することは困難である．企業のウェブサイトを参照すると，投資家向け情報ページには財務諸表や決算説明資料，年次報告書，証券レポートがある．検索エンジンで企業名を検索すれば，多くのニュース記事を見つけることができる．金融市場についてのインターネット掲示板には，金融の情報や株価動向に関連した様々な投資家の意見が載っている．さらに，近年では Twitter や Facebook，Instagram など，ソーシャルネットワークサービス (SNS) における人々の投稿やコメントも投資家の感情を反映しうる．Bollen らはツイートに現れる感情がダウ・ジョーンズ工業平均の予測に有用であることを示した [1]．Sakaji らは統計的手法を用いて新聞記事から経済のトレンドを示す根拠表現を自動的に抽出する手法を提案した [2]．Kitamori らは，決算短信から業績予想や経済予測を示す文章を抽出し分類する手法を提案した．この分類手法は半教師ありの手法を用いたニューラルネットワークをもとにしている [3]．

以上の手法はニュース記事や決算短信など事実のみについて書かれた文書に適用しているが，これらの情報を投資家が将来の投資に直に活かすことは難しい．アナリストレポートはこの状況において注目を集めている．アナリストレポートは，ニュースやプレスリリース，株価の評価，マクロ経済のトレンドなどを考慮に入れ，それぞれの銘柄に対する評価を専門家であるアナリストがまとめたものである．そのため，アナリストレポートはニュース記事や決算短信など，各投資の情報源の上位互換とみなせる．本研究では株価動向の推定のためにアナリストレポートの本文を分析する．株価動向において特に重要である，超過リターン (株価の市場に対するリターン) の正負と株価のボラティリティの大小を予測することを主眼に置く．さらに，アナリストレポートは，その形式や内容が発行した証券会社ごとに異なる可能性があるため，証券会社ごとに分類してその有効性を検証する．また，アナリストレポートのテキスト分析のために適した言語モデルを構築するために，様々なリソースから Word-Embedding のベクトル辞書を作り，それぞれの組み合わせの有効性を試す．

2　手法

アナリストレポートはアナリスト自身の予測と客観的事実の両方を含む．そのためアナリストの意見文が株価予測に有効であるという仮定のもと，まずアナリストレポートを意見文と非意見文に分ける．そして意見文，非意見文，両方の別々の入力などによって株価動向を予測する．これにより意見文，非意見文といった文脈により株価動向予測の有効性が異なる可能性を検証する．以下ではまず本研究で用いたアナリストレポートなどのデータセットを述べた後，本研究で用いた手法について述べる．

[1] https://www.jpx.co.jp/markets/statistics-equities/examination/01.html

2.1 意見文抽出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の中から無作為に 100 本を抽出した．そこに含まれる 2,213 文について意見文と非意見文に手作業で分類した．意見文の典型例を以下にあげる．

- 2Q 実績を踏まえ，業績予想を下方修正する．
- 収益性低下の要因としては研究開発投資が考えられる．

意見文はレーティング (株価動向の予測) や企業の次年度の売上，純利益といったアナリスト自身の予測や，現在の業績に至った背景等の内容が含まれている．一方非意見文の典型例は以下である．

- 今期の売上は過去最高となった．
- 期末配当は 150 円を予定している．

非意見文は企業の過去の業績値などの事実に関する文を指している．2,213 文のうち 1,188 文が意見文に，残りの 1,025 文が非意見文と分類された．

2.2 株価動向予測のデータセットの作成

本研究では，日本の大手証券会社 (A, B, $\cdots$) から発行された 58,010 本のアナリストレポートを使用した．まず全てのアナリストレポートの発行日を取得し，各レポートについて発行日からその 10 営業日 (約 2 週間) 後までの株価と TOPIX の値を取得した．各指標は終値を用いる．これらの指標を用いて超過リターンを計算する．レポート発行日当日の株価を C_0，TOPIX を T_0，10 営業日後の株価を C_{10}，TOPIX を T_{10} として，超過リターンは式 (1) によって計算される．

$$\frac{(C_{10}-C_0)}{C_0}-\frac{(T_{10}-T_0)}{T_0} \tag{1}$$

単純なリターンではなく超過リターンを用いる理由としては，2017 年頃の日本が長期的な経済回復傾向にあるためである．単純なリターンを用いるとリターンが正に偏ってしまう．更に，個人投資家にとってはベンチマーク比で評価されることから超過リターンの予測可能性のほうが重要である．正の超過リターンのレポートには 1 を，負の超過リターンのレポートには 0 をそれぞれラベリングした．全てのアナリストレポートの中で 29,430 レポートが 1 に，28,580 レポートが 0 にラベリングされた．また，各レポートの銘柄についてヒストリカル・ボラティリティも計算した．レポート発行日当日から 10 営業日後までの株価を $C_0, C_1, \cdots, C_{10}$ とする．ボラティリティは式 (2) のように，ある日の株価とその前日の株価の分数をそれぞれ求めた配列の標準偏差である．

$$\mathrm{SD}\left(\frac{C_1}{C_0}, \frac{C_2}{C_1}, \cdots, \frac{C_{10}}{C_9}\right) \tag{2}$$

ボラティリティの値が中央値より高いものに 1 を，中央値より低いものに 0 をラベリングした．

2.3 入力ベクトルの作成

アナリストレポートの文章を提案手法に入力するために，200 次元の分散表現 (ベクトル) を構築した [4]．分散表現の構築にあたって，文章を単語ごとに分けることと，単語ごとに分散表現に変換するという 2 つの前処理を行った．前者の処理では，MeCab[2]とその辞書として mecab-ipadic-NEologd[5] を，後者では Global Vectors for Word Representation (GloVe)[3]を用いた．GloVe において以下の 5 つのコーパスを分散表現の構築のために用いた．

- アナリストレポート 1: 意見文抽出に使用
- アナリストレポート 2: 株価動向予測に使用
- ロイターの日本語記事
- Wikipedia の日本語記事
- 日本経済新聞の記事

2.4 意見文抽出

意見文抽出においては，ニューラルネットワークを用いたモデルを用いた．リカレントニューラルネットワーク (RNN) の中でも特に，Long Short-Term Memory (LSTM)[6][7] と Gated Recurrent Unit (GRU)[8] は高い成果をあげている．そのため，本研究ではこれらの中で，特に両方向の手法を採用した．一般的な片方向の LSTM や GRU では，過去の情報しか学習に用いられないのに対し，両方向の (bidirectional) LSTM や GRU では過去の情報に加え未来の情報を用いて学習することが可能である．

本研究の概要図を図 1 に示す．まず分散表現によって得られる LSTM や GRU への入力ベクトルを GloVe によって作成する．LSTM や RNN に 1 度に入力する

[2] https://taku910.github.io/mecab/
[3] https://nlp.stanford.edu/projects/glove/

シーケンスの長さを揃えるために，最も長いシーケンス長になるよう 200 次元の 0 ベクトルでパディングを行った．LSTM や GRU からの出力には Additive Attention をかけることで，予測モデルのどの部分が重視されているかを表現できるようにしてより正確な予測を可能にした．Attention の出力は MLP に入力され，MLP の最終層で 1 と 0 のラベルの確率がそれぞれ予測される．より確率の高い方のラベルが結果として採用される．

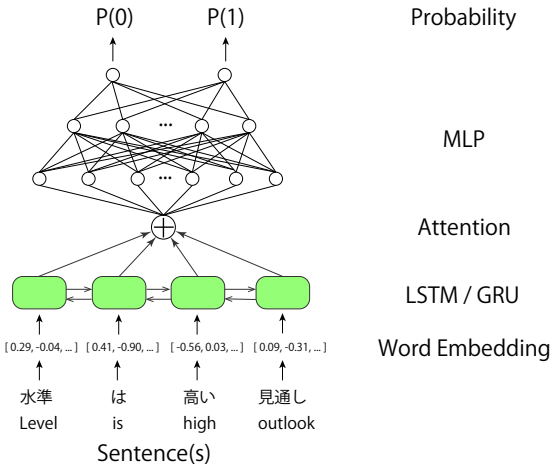

図 1: 提案手法におけるネットワークの概要

2.5 株価動向推定

株価動向推定では 4 つの入力方法について実験を行った．アナリストレポートの全文を入力するもの，意見文のみを入力するもの，非意見文のみを入力するもの，意見文と非意見文を分類した上で別々に入力するものである．はじめの 3 つの入力方法についてのネットワークの概要図は図 1 となる．最後の意見文と非意見文を別々に入力する際のネットワークの概要図を図 2 に示した．

3 実験

意見文抽出タスクにおいては，全 2,213 文を 7:1:2 に分割し，それぞれを Train, Valid, Test とし，エポック数や隠れ層の数などパラメータを変えながら Valid データでの F1 が最も小さくなるようにパラメータサーチを行った．最も良い性能を出したパラメータを用いて，株価動向予測に用いる 58,010 レポートの文について判別を行った．

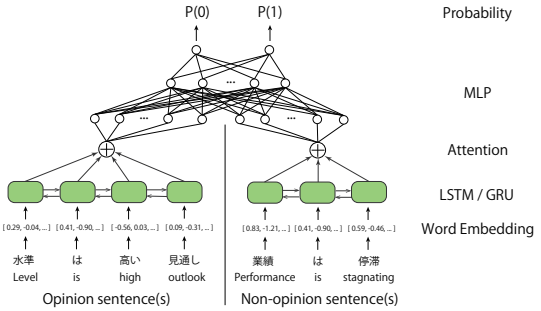

図 2: 株価動向予測で意見文と非意見文を別々に入力する際のネットワークの概要

株価動向予測タスクでは超過リターンの正負とボラティリティの高低の 2 つを予測した．2.5 節で述べた 4 つの入力方法で実験を行った．また，アナリストレポートの発行会社ごとに分けて入力した．意見文抽出タスクと同様，全 58,010 レポートを 7:1:2 に分割しパラメータサーチを行った．

意見文抽出と株価動向予測の両タスクで，比較手法は SVM と Random Forest を用いた．入力は文 (レポート) に含まれる各単語の One-hot 表現のベクトル和とした．

4 結果

意見文抽出でバリデーションデータでの F1 の上位 5 つの結果を表 1 に示した．表中の発行は発行元の証券会社，コーパス列のレポート 1，レポート 2 はそれぞれアナリストレポート 1，アナリストレポート 2 を指す．最も良かったパラメータ (表 1 の最上部) をテストデータに適用した際の F1 は 0.813 となった．SVC と Random Forest のテストデータでの F1 はそれぞれ 0.797 と 0.664 となった．

表 1: 意見文判別の上位 5 結果

F1	モデル	コーパス	エポック
0.836	GRU	アナリストレポート 2	130
0.835	GRU	アナリストレポート 2	125
0.835	LSTM	アナリストレポート 1	50
0.835	GRU	アナリストレポート 2	120
0.834	GRU	アナリストレポート 2	135

超過リターンの正負の予測におけるバリデーションデータでの F1 の上位 5 つの結果を表 2 に示した．最も良かったパラメータ (表 2 の最上部) をテストデータに適用した際の F1 は 0.558 となった．SVC と Random Forest のテストデータでの F1 はそれぞれ 0.506 と 0.533 で，発行した証券会社はどちらも A であった．

表 2: 超過リターンの正負の予測の上位 5 結果

F1	入力	発行	モデル	コーパス	エポック
0.574	全文	A	LSTM	レポート 2	140
0.573	全文	A	LSTM	レポート 1	140
0.573	全文	A	LSTM	レポート 1	120
0.573	全文	A	LSTM	レポート 1	135
0.573	全文	A	LSTM	レポート 1	145

ボラティリティの大小の予測におけるバリデーションデータでの F1 の上位 5 つの結果を表 3 に示した．最も良かったパラメータ (表 3 の最上部) をテストデータに適用した際の F1 は 0.635 となった．SVC と Random Forest のテストデータでの F1 はそれぞれ 0.571 と 0.627 で，発行した証券会社はどちらも A であった．

表 3: ボラティリティの大小の予測の上位 5 結果

F1	入力	発行	モデル	コーパス	エポック
0.667	全文	B	GRU	レポート 2	135
0.666	全文	B	GRU	レポート 2	145
0.666	全文	B	GRU	ロイター	115
0.665	全文	B	GRU	レポート 2	125
0.665	全文	B	GRU	レポート 2	130

5 考察

意見文抽出においては 0.8 以上の F1 を達成した．表 1，2，3 にある 15 の結果のうち，14 がアナリストレポート由来のコーパスを用いたものだった．アナリストレポートは新聞記事や Wikipedia の記事と異なり，事実や情報だけでなくアナリストの意見も含むことから，アナリストレポート特有の表現が存在し，その結果アナリストレポートに基づくコーパスがより良い結果となったと言える．一方で，表 2 と 3 では，意見文と非意見文に分けることが株価動向予測には効果的ではなかった．そのため我々の仮説は誤りとなった．

しかしながら，アナリストレポート 2 は意見文判別で高いパフォーマンスを出したが，アナリストレポート 1 はボラティリティの大小の判別に高いパフォーマンスを出すなど，アナリストレポートごとに有効な条件が異なった．株価動向推定についてアナリストレポートごとに異なる特性があると考えられる．そのため分析の内容に応じて異なるアナリストレポートを使用すべきである．

6 まとめ

アナリストの意見文が株価動向予測に効果的であるという仮定とともに，我々はアナリストレポートを意見文と非意見文に分割した．本実験環境では，この仮定は有効ではなかった．一方で，アナリストレポート由来のコーパスは株価動向予測に効果的だった．これは有益な情報である．今後は，意見文と非意見文について他の条件下で実験を行い，株価動向予測における自然言語処理手法の有効性を確かめたい．

参考文献

[1] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, Vol. 2, No. 1, pp. 1–8, 2011.

[2] Hiroki Sakaji, Hiroyuki Sakai, and Shigeru Masuyama. Automatic extraction of basis expressions that indicate economic trends. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pp. 977–984, 2008.

[3] Shiori Kitamori, Hiroyuki Sakai, and Hiroki Sakaji. Extraction of sentences concerning business performance forecast and economic forecast from summaries of financial statements by deep learning. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–7, 2017.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. CoRR, 2013.

[5] Sato Toshinori. Neologism dictionary based on the language resources on the web for mecab, 2015.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[7] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, Vol. 18, No. 5, pp. 602–610, 2005.

[8] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734. Association for Computational Linguistics, October 2014.