2G4-OS-25a-4

ニュースのテキスト情報から株価を予測する

Estimating news articles' negative-positive by Deep Learning

五島 圭一 *1　高橋 大志 *2　寺野 隆雄 *3
Keiichi Goshima　Hiroshi Takahashi　Takao Terano

*1*3東京工業大学 大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

*2慶應義塾大学 大学院経営管理研究科
Graduate School of Business Administration, Keio University

This study analyses the relationship between textual information and financial markets in Japan, focusing on Headline News, a source of information that has immediate influence on the money market, and also which is regarded as an important source of information when making investment decisions. In particular we propose the objective way to estimate news articles' negative-positive by using Deep Learning.

1. はじめに

本研究は，日本株式市場における資産価格の決定要因について解明するため，ニュースのテキスト情報と株価の関連性について分析を行ったものである．とりわけ，Deep Learning[5] を用いたニュースのテキスト情報の極性（ポジティブ・ネガティブ）の推測を通して，ファイナンス分野でのテキスト分析における Deep Learning の有効性を検証する．

投資家は，新聞やテレビ，各企業のプレスリリース，ソーシャルメディアなど，様々なメディアからニュースを入手し，投資先となる企業を選定する．ニュースには数値情報だけでなく，テキスト情報も含まれており，それらを活用することで数値情報だけでは説明することが難しい資産価格の変動やマーケットメカニズムなどの分析や予測ができる可能性がある．そのため，2000 年代中頃から，資産価格の分野において，ニュースやソーシャルメディアといったテキストデータを，資産価格評価の分析に用いる試みが模索されている．例えば，Tetlock (2007) は Wall street Journal column から悲観度を抽出し，ダウ工業平均株価との関連性を見出している [7]．また，ソーシャルメディアと株価の関連性に言及している研究も存在する．Bollen et al. (2011) は，twitter の投稿内容を利用し，ダウ工業平均株価の変動を 87.6%の精度で予測できたとしている [6]．

このようにテキスト情報を用いることで，より正確な資産価格評価の試みがなされている．テキスト分析を行う際には，辞書の精度が重要となる *1．Loughran and McDonald (2011) では，ファイナンスの文脈に沿ったテキスト評価の重要性を指摘しており，彼らは金融用の辞書を作成し，より精度の高い結果が得られたと報告している [8]．

しかしながら一方で，資産価格分析における文脈に沿ったテキスト内容の評価を行う際には，人の手によって，経験的に行われることになり，評価者の主観に強く依存してしまう可能性がある．それに対する解決策の一つとして，実際の資産価格からニュース記事を評価する方法があり，Healy and Lo (2011) では，外国為替を用いてニュース記事の評価を行い，リスク指標の作成を試みている [4]．また，五島/高橋 (2015) は，日本語記事を対象に，個別銘柄の株価情報を用いて，SVR (Support Vector Regression) によってニュース記事のポジネガを推測することで，より客観的かつ資産価格分析の文脈に即したニュース記事内容の評価を試みている．

そこで本分析では，SVR をベンチマークとし，Deep Learning によるニュースのテキスト情報のポジネガを推測し，そのポジネガ情報を元にした株式投資戦略を構築し，本分析方法の有効性の検証を行った．次章は，データに触れ，3 章では分析方法，4 章では分析結果を記す．5 章は，まとめである．

2. データ

2.1 マーケットデータ

本分析では，個別銘柄の株価データについて，Thomson Reuters Datastrem から，トータルリターンの日次データを用いた．また，マーケットファクターのデータについては「日本上場株式　久保田・竹原 Fama-French 関連データ」からマーケットリターン (Rm)，リスクフリーレート (Rf)，バリューファクター (HML)，サイズファクター (SMB) の日次データを使用した．

2.2 ニュースデータ

ニュースデータについては，ロイターニュースを用いた．ロイターニュースは，トムソンロイター社の提供するニュースであり，本分析では，日本証券市場に関する日本語のニュース記事のみを分析対象とした．主に利用したタグ情報は，ニュースの発信日時・ニュースの見出し・各ニュースと関連する企業名（証券コード）を利用した．

本分析で用いるロイターニュースは，日本証券市場に参加している数多くの機関投資家がリアルタイムで閲覧するメディアであり，新聞やテレビニュースに比べ，イベントからニュース発信までのラグが小さく，ニュース発信時点において，資産価格に織り込まれていない情報を相対的に多く有すると考えられる．分析対象期間は 2009 年から 2010 年とし，分析対象企業は東証 1 部上場企業のみを分析対象とした．

連絡先: 五島 圭一　東京工業大学
大学院総合理工学研究科 知能システム科学専攻
〒 226-8502 神奈川県横浜市緑区長津田町 4259-J2-1705
E-mail: goshima.k.aa@m.titech.ac.jp

*1　本稿では，テキスト情報に極性（ポジネガ）を付与をするためのリストのことを辞書と呼んでいる．

3. 分析方法

3.1 分析手順について

ここでは，本分析の分析手順の概略を記す．図 1 は，分析の流れを図にしたものである．

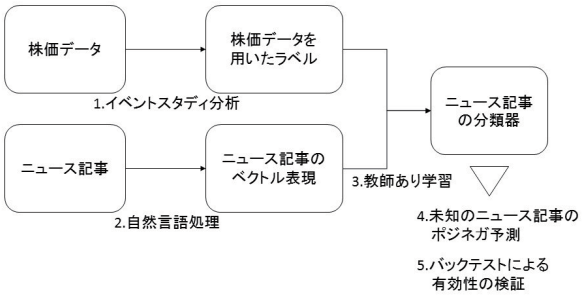

図 1: 本分析における手順の概略図

(1) はじめに，株価データを基に，ニュース記事にラベル（ポジティブ-ネガティブ）の付与を行った．株価データを基にした評価を行うことにより，客観的な記事評価を行うことが可能となる．本分析では，日本証券市場を対象としてイベントスタディ分析によって株価を教師情報としたニュース記事のラベルの生成を試みた．(2) 次いで，各ニュース記事を，bag-of-words に基づき，記事内容のベクトル表現を行った．(3) 更に，株価データからラベルを付与したニュース記事を訓練データとし，機械学習 (SVR，Deep Learning) によってニュース記事へのポジネガ付与を行う分類器を作成し，(4) テストデータとなるニュース記事へのラベル付与を行った．2009 年のニュース記事を訓練データとし，2010 年のニュース記事をテストデータとした．(5) 最後に，機械学習によって付与されたラベルを元に，株式投資戦略を構築し，バックテストを行い，考察を行った．以上の手順によって，本分析を進めた．次節以降において，それぞれの分析方法について詳細を記述する．

3.2 株価データからのラベル付与について

本分析ではイベントスタディ分析 [2] によって，株価データからニュース記事へのラベルの付与を試みた．

正常リターンを算出するためのモデルについては，Fama-French の 3 ファクターモデル [3] によって行った．また，モデルのパラメータを推定する際の推定期間に関しては，イベント日から 125 日前から 6 日前の 120 日間において推定を行った．イベントウィンドウに関してはニュース発信日の当日から 1 日後までの間とした．これは，ニュース記事が包含する情報を要因とした株価変動のみを抽出するためである *2．本分析で使用したニュースデータであるロイターニュースは報じられた日時が明確でイベント日を特定しやすいため，可能となると考えた．15 時以降に発信されたニュース記事については次の市場営業日に編入し，日付が市場休業日のニュースに関しても同様に，次の市場営業日に編入し，分析を進めた．

ここで，標準化を行い，ニュース発信日当日から 1 日後までの標準化された累積異常リターン $SCAR_i(0,1)$ を，当該ニュース記事が包含する情報を要因とした株価変動とし，ニュース記事の教師ラベルとした．

*2 正確には，ニュース発信後のみの株価変動を教師情報とすべきであるが，本分析では日次リターンを用いてニュース記事へのラベル付与を試みており，場中に裁定取引済みのニュース記事へのラベル付与も行えるよう，当日のリターンも含めている．時間単位での分析は，今後の課題である．

3.3 ニュース記事のベクトル表現について

テキスト分析をする際には，文書をベクトル表現することが求められる．本分析では，bag-of-words で表現を行うため，形態素解析，tf-idf 法，正規化を行った．そして，名詞，動詞，形容詞の 3 つの品詞に注目し，抽出した．また，数値情報に関する名詞は除去をし，テキスト情報のみをベクトルの素性としている．

3.4 機械学習によるポジネガ付与

学習データのニュースへのラベル付与については，SVR と Deep Learning によって試みた．SVR のパラメータチューニングについては，グリッドサーチによってハイパーパラメータの最適化を行っている．Deep Learning について，活性化関数は Rectifier 関数，隠れ層は 3 層，各隠れ層のユニット数は 100 とした．また，過学習を防ぐために Dropout 率を 10%としている．

3.5 バックテストのルールについて

最後に，株式投資戦略を構築し，バックテストを行うことで，本分析方法の有効性の検証を行った．前節の SVR と Deep Learning によって，各ニュース記事に付与されたラベルの値に対して，標準正規分布 *3 を仮定し，$z_{0.975}$ を超えたとき，ニュース記事によってもたらされた情報によって有意にリターンがプラスになると予測できると考えた．機械学習によって，ラベルが付与されたニュース記事を対象として，ラベルの値が $z_{0.975}$ を超えたニュース記事に付随する銘柄を当日の終値で購入し，1 日後の終値で売却するというロングポジションを取ることで，インデックスを作成した．同日に複数のニュース記事のラベル値が $z_{0.975}$ を超えたときは，個別銘柄のトータルリターンを単純平均することによってインデックスを算出している．一方で，該当銘柄が存在しないときは売買は行っていない．

また，同様に，$-z_{0.975}$ を下回ったとき，ニュース記事によってもたらされた情報によって有意にリターンがマイナスになると予測できると考えた．ラベルの値が $-z_{0.975}$ を下回ったニュース記事について，付随する銘柄を当日の終値で空売りを行い，1 日後の終値で買い戻すというショートポジションを取るものとして，インデックスを作成した．ショートポジションについても，同日に複数のニュース記事のラベル値が $-z_{0.975}$ を下回ったときは，個別銘柄のトータルリターンを単純平均することによってインデックスを算出し，該当銘柄が存在しないときは売買は行っていない．

これら，2 つのインデックスを平均したものをロングショート戦略によるインデックスとし，本分析手法の有効性の検証を行った．

4. 分析結果

前章にて記したロングショート戦略によって，バックテストを行った結果を記述する．ロングショート戦略のよって算出されたインデックスを Fama-French の 3 ファクターモデルによって，パフォーマンスの測定をし，考察を行った．

表 1 は SVR 及び Deep Learning によって，2010 年のニュース記事の超過リターンを予測し，株式投資戦略によって作成したインデックスをファクターモデルを用いてパフォーマンスを測定した結果を示したものである．

まず，SVR によって，ポジティブあるいはネガティブだと予測したニュース記事からロングショート戦略によって作成し

*3 厳密には，自由度 $n-4$ のステューデントの t 分布に従うが，本分析では推定期間が 120 日と十分に長く，標準正規分布への近似をしている．

たインデックス (Rt - Rf) については，α が有意確率 5%水準で 0.20 となり，マーケットファクター (Rm - Rf)，サイズファクター (SMB)，バリューファクター (HML) を考慮してもなお，超過収益を獲得していることを確認できる．

次に，Deep Learning によって，ポジティブあるいはネガティブだと予測したニュース記事からロングショート戦略によって作成したインデックスについても，同様に超過収益を獲得できることが示された．α が有意確率 5%水準で 0.17 となり，超過収益を獲得していることを確認できる．

これらの結果は，機械学習によるニュース記事の分析を通じ，超過収益の獲得をできる可能性を示すものであり，特に，ニュースのテキスト情報の極性（ポジティブ・ネガティブ）の推測について，Deep Learning の有効性を示すものである．本分析では，Deep Learning におけるすべてのハイパーパラメータを網羅しておらず，細緻なパラメータチューニングを行うことによって，より正確なニュースのテキスト情報のポジネガ推測を行える可能性がある．より精緻な分析については，今後の課題である．また，取引コスト等を考慮するなど，現実の投資条件を考慮した分析についても，今後の課題である．

5. まとめ

本研究では，投資家の意思決定ルールを解析するための分析対象として，ニュースのテキスト情報と株価の関連性について取り上げた．資産価格分析における文脈に沿ったテキスト内容の評価を行う際には，人の手によって，経験的に行われることになり，評価者の主観に強く依存してしまう可能性がある．本分析では，個別銘柄の株価情報を用いることで，より客観的かつ資産価格分析の文脈に即したニュース記事評価分析方法を提示した．とりわけ，Deep Learning を用いたニュースのテキスト情報の極性（ポジティブ・ネガティブ）の推測を通して，ファイナンス分野でのテキスト分析における Deep Learning の有効性の検証を行った．分析の結果，機械学習によるニュース記事の評価を通して，将来の株価予測ができる可能性を見出した．また，ニュースのテキスト情報の極性（ポジティブ・ネガティブ）の推測について，SVR と同様に，Deep Learning が有効であることを示す結果となった．今後の課題としては，Deep Learning のより細緻なパラメータチューニングや分析期間および分析対象資産の拡大などが挙げられる．

参考文献

[1] Bishop, Christopher M.: Pattern Recognition and Machine Learning, Springer (2006).

[2] Campbell, J. Y., A. W. Lo, and A. C. MacKinlay.: The Econometrics of Financial Markets, Princeton University Press (1997). 祝迫・大橋・中村・本多・和田訳: ファイナンスのための計量分析, 共立出版 (2003).

[3] Fama, E. F. and K. R. French.: Common risk factors in the returns on stock and bonds, *Journal of Financial Economics*, Vol. 33, pp. 3–56 (1993).

[4] Healy, Alexander and Andrew W. Lo,: Managing Real-Time Risks and Returns: The Thomson Reuters NewsScope Event Indices. In: Mitra, G. and Mitra L. (eds.), The Handbook of New Analytics in Finance, John Wiley & Sons, West Sussex, UK (2011).

[5] Hinton, G. E., Osindero, S. and Teh, Y. : A fast learning algorithm for deep belief nets, *Neural Computation*, Vol. 18, pp. 1527–1554 (2006).

[6] John Bollen, Hunia Mao and Xiaojun Zeng.: Twitter mood predicts the stock market, *Journal of Computational Science*, Vol. 2, No. 1, pp. 1–8 (2011).

[7] Paul C. Tetlock.: Giving Content to Investor Sentiment:The Role of Media in the Stock Market, *The Journal of Finance*, Vol. 62, No. 3, pp. 1139–1168 (2007).

[8] T. Loughran and B. McDonald.: When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, *The Journal of Finance*, Vol. 66, No. 1, pp. 35–65 (2011)

[9] 五島圭一, 高橋大志：株価を用いたニュース記事評価に関する研究, 第 23 回日本ファイナンス学会 (2015) (to appear)

表 1: SVR 及び Deep Learning によるロングショート戦略

	SVR によるロングショート戦略	Deep Learning によるロングショート戦略
	Rt - Rf	Rt - Rf
α	0.20**	0.17**
	(2.10)	(2.01)
Rm - Rf	0.16	0.05
	(1.49)	(0.51)
SMB	0.17	-0.02
	(0.65)	(-0.07)
HML	0.15	-0.03
	(0.53)	(-0.13)
adj.R	-0.002	-0.01
Obs	244	244

両側確率 : *** p < 0.01, ** p < 0.05, * p < 0.1