高頻度データおよびLSTMによる韓国株式市場におけるニュースの分析

(The Analysis on Korean Stock Market By High Frequency Data and Text Classification using LSTM)

尹　聖在(ユン　ソンジェ)[1]、菅　愛子[1]、高橋　大志[1]

Sungjae Yoon[1], Aiko Suge[1], Hiroshi Takahashi[1]

[1]慶應義塾大学大学院経営管理研究科

[1]Graduate School of Business Administration, Keio University

Abstract: This study analyzes the relationship between news aritcles and stock price changes. While several researches about stock return analysis using deep learning and frequency data have been reported, Korean cases haven't. In this study, we analyze the influence of news articles on Korean stock markets using high frequency data and text classification by LSTM network.

1. はじめに

情報と価格に関する研究は、資産価格理論において関心を集めている。投資家は数値データ、文字データをはじめとしたあらゆる情報に基づき、投資の判断を行う。近年、コンピューター能力の向上で、ビッグデータやこれまで扱うことが困難であったデータを用いて、従来の研究課題に対して詳細な分析が可能となった。その結果、ニュースなどのテキスト情報を用いて資産価格分析を行う研究の事例も増えている。

テキスト分析を用いた先行研究としては、例えばAntweiler & Frank(2004)[1]ではYahoo! Finance Raging Bullに掲示されたダウ指数に上場された45社の150万件以上のメッセージを研究した結果、市場の変動性の予測に役に立つことを報告している。Tetlock(2007)[2]ではWall Street Journal columnを用いてメディアコンテンツと株式市場の活動の相互作用について研究し、メディアの高い悲観度は市場価格の下落を誘導し、異常な悲観度は一時的な取引量の上昇を導くことを報告している。また、Tetlock (2008)[3]ではS&P 500の個別企業に対して、Wall Street JournalとDow Jones News Service研究した結果、記事の悲観的な言葉はファンダメンタルに関する測り難い様相を捉えているので売上やリターンの予測に有用であると報告している。

金融市場における投資家行動との関連性に関する報告も存在している。Engelberg (2012)[4]はDow Jones Newsを解析し、空売り筋の利益は利用可能な公開情報の優れた解析能力によって導かれると報告している。また、Dougal (2012)[5]は、Dow Jones Industrial AverageとWall Street Journalのコラムの関係から、フィナンシャル・ジャナーリストは投資家の行動に大きな影響を与えると報告している。

アメリカ以外の株式市場を対象とした研究報告も見られる。日本株式市場の場合は五島・高橋(2016)[6][7]は、深層学習（ディープラーニング）を用いて、Thomson Reutersのニュースを指標化し、株価との関連性を分析した報告が見られる。韓国株式市場の場合はKim & Willett (2014)[8]は2007年8月から2010年3月までの韓国の新聞記事と韓国総合株価指数（Korea Composite Stock Price Index, KOSPI）との関係性から世界経済危機下のニュースによる韓国株式市場の行動を報告している。Lee & Cho (2014)[9]ではNaver News Serviceを用いて、韓国株式市場のモメンタム効果について報告している。

深層学習を用いた株価分析研究は、以下の事例がある。宮崎・松尾(2017)[10]はTOPIX Core 30の過去データからのリターン予測にRNN（Recurrent Neural Network）を適用した結果を報告している。松井・汐月(2017)[11]はトヨタ自動車の1分足株価データにLSTM（Long Short-Term Memory、[14]）Networkを適用した結果を報告している。五島・高橋・寺野(2017)[12]はRNNをボラティリティー クラスタリングのモデル化に応用している。韓国株式市場における研究では、Chong & Han & Park (2017)[13]が韓国企業の株式リターン予測に深層学習を適用した結果を報告している。

このようにテキスト情報や深層学習を用いた様々な市場分析が存在している。しかし、株式取引データの観点からみると、多くの研究は日次単位や分単位での分析である。また、秒単位の高頻度データを用いた先行研究は五島・高橋以外の研究(2016)[6][7]による日本株式市場を対象としたものに限られており、韓国株式市場に関しては報告されていない。更に、韓国市場におけるテキストデータについて深層学習を用いた研究はこれまで報告されていない。よって、本研究では韓国株式市場における高頻度データ（ティックデータ、秒単位の株式取引データ）とニュースデータを用い、ニュースと株価変動の関係を深層学習を用いてモデル化し、実際の結果と比較する。

2. データ

株式取引データは Thomson Reuters 社から取得した。本研究では 2013-2014 年の韓国株式市場の高頻度データより、韓国有価証券市場の時価総額上位 5 社の企業（表 1、2012 年末基準）を対象にし、株式価格・取引量・取引時間を抽出し分析を行った。韓国有価証券市場の時価総額上位 5 社の企業は有価証券市場の時価総額の割合の 30%を占めている。

表 1. 韓国有価証券市場の時価総額上位 5 社の企業（2012 年末基準）

	企業名	時価総額 (十億 KRW)	時価総額の割合
1	サムスン電子	224,189	19%
2	現代自動車	48,130	4%
3	POSCO	30,428	3%
4	現代モービス	28,035	2%
5	起亜自動車	22,903	2%
	合計	353,685	30%

ニュースデータは Thomson Reuters 社のニュースを用いた。Thomson Reuters 社は金融市場において最も信頼されているニュースソースの一つである。分析対象は表 2 の上位 5 社についての英語ニュースで、韓国株式市場開場時間のみ（平日 9-15 時、UTC+9）を抽出した。データのタグ情報には、ニュースの発信日時と 5 社の証券コード、ニュースの属性 (アラート)を利用した。

表 2：2013-2014 年度上位 5 社の英語ニュース数（韓国株式市場開場時間のみ）、

年度	アラート
2013	281
2014	274

ニュースの属性は 2 種類で、アラートと本文ニュースがある。今回の分析に用いたアラートには緊急情報の発信が含まれている。タイトルが付与されており、記事本文は含まれない。ニュース内容によっては、アラートが存在せずに本文ニュースのみが発信される場合もある。本文ニュースにはタイトルと詳細内容が記録されている記事本文が含まれる。また、アラートニュースが存在する場合、本文ニュースはアラートの発信から 0〜30 分程度経過後、発信される。今回の分析はアラートのみ研究対象とした。

図 1. アラートと本文ニュースの関係

3. 分析方法

本研究では深層学習モデルとして RNN（Recurrent Neural Network）の一種である LSTM（Long Short-Term Memory、[13]）Network を用いた。LSTM Network は従来の RNN が持っていた長期的な情報をうまく取り込むことができない問題を解決した手法であり、1997 年に初めて発表された後、改善[15]がなされてきた。この手法は特に自然言語処理や音声認識の分野によく活用されており、最近は株価変動予測の研究（[10][11]）にも適用されている。

図 2 は今回の研究に適用する深層学習モデルの模式図である。大きく、入力層、中間層、出力層で構成されている。まず、入力層に文字データのベクトル表現が入り、中間層に移る。その後、中間層で学習と分類が行われ、出力層に移動し、出力データが出て来る。中間層は LSTM Network を構成する LSTM 層、分類の重みを学習する全結合層、全結合層のデータからソフトマックス関数を用いて各クラスに所属する確率を求めるソフトマックス層からなる。

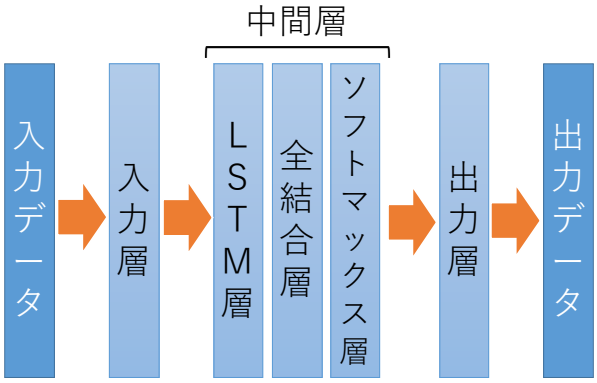

図 2.　本研究に適用する深層学習モデルの模式図

深層学習による分類分析は以下のプロセスに従う。

(1) まず、2013 年度の全ての英語ニュースに対し、高頻度データを用いて以下のように定義したリターンを求める。すなわちニュース発信を 0 分として、そのニュース対象企業の 30 分～16 分前と 16 分～30 分後の株式価格差をリターンとし、比率に直した。

$$\text{リターン}(\%) = \frac{(\text{30 分～16 分前の平均価格}) - (\text{16 分～30 分後の平均価格})}{(\text{30 分～16 分前の平均価格})} \times 100$$

(2) 図 3 のようにリターンを次の 3 つのクラスに分類し、教師データとする。

ポジティブ：リターン　＞ α
ニュートラル：$-\alpha$ ＜ リターン ＜α
ネガティブ：リターン　＜ $-\alpha$

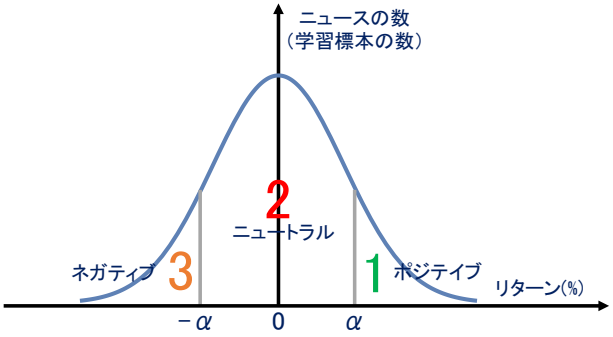

図 3.　3 つのクラスに分類の模式図

ここで α は色々な設定ができるが、今回は α =0.5%に設定し、教師データを作成する。

(3) ニュースの深層学習を行うために、英語ニューステキストに対して以下の前処理を加える。

1. 各種の句読点(Punctuation mark)を削除する。
2. 記録されている URL アドレスを削除する。
3. アラートタイトル全体を小文字化する。
4. アラートタイトルの先頭と末尾に存在する空白部分を削除する。
5. アラートタイトルを単語ごとに分割（トークン化、tokenize）する。

(4) 図 2 の深層学習モデルを以下の条件に設定する。

- 入力データ：(2)で作成されたリターンデータを教師とし、(3)の前処理後のニュースデータを変数とする。
- 入力層では Skip-gram Model[16]を用いて単語ごとに分割された各ニュースに対し、ベクトル表現への変換を行う。
- 中間層 ：3 種類 3 層(1 LSTM 層、1 全結合層、1 Softmax 層)

(5) 2013 年度英語ニュースとリターンの分類結果で学習を行う。ここでモデルの精度を検証するため、学習済みモデルに 2013 年度英語ニュースを代入して評価する。評価には、クラス分布の一致度を用いた。こうした 10 回分の評価結果を合算し、比率を求めることで深層学習モデル検証とした。

(6) 2013 年のデータによる深層学習モデルの作成・検証後、2013 年度英語ニュースに対する期待リターン分析も行う。その後、同様の前処理を施した 2014 年度英語ニュースの 3 クラス分類の分類分析を 10 回実施し、2014 年度英語ニュースに対する期待リターン分析も行う。期待リターンは分類分析クラスごとに分類された銘柄の実際リターン(2013 年又は 2014 年)を算出し平均したものである。

4. 分析結果

4.1 教師データ

表 3 は 3. 分析方法に記載した方法により、2013 年度英語ニュースに対するリターンを計算し、教師データを作成した結果である。

表 3. リターンクラスの分布
(r=リターン)

ニュース種類 / リターンクラス	アラート
r > 0.5 （ポジティブ）	33 (11.7%)
-0.5 < r < 0.5 （ニュートラル）	177 (63.0%)
r < -0.5 （ネガティブ）	71 (25.3%)
合計	281 (100.0%)

4.2 深層学習モデルの検証

2013 年度英語ニュースに対するリターンに基づいて作成された学習データ適用の深層学習モデルに対し、2013 年度英語ニュースの分類分析を行った結果、おおよそ 80%以上の精度が確認された。これによって分類アルゴリズムの構築ができた。

表 4.分類分析結果と実際の一致度
(2013 年度のアラート)

	一致	不一致	合計
一致度	80.2%	19.8%	100.0%

4.3 2013 年度ニュースの期待リターン

2013 年度英語ニュースの学習に対し、2013 年度英語ニュースの期待リターン分析を行った結果、ポジティブなニュースの場合はプラスのリターンを示しており、ネガティブなニュースの場合はマイナスのリターンを示している。こうしてある程度の精度をもった分類器が作成されたことを確認した。

表 5. 2013 年度ニュースの期待リターン

ニュースの属性	r > 0.5 (ポジティブ)	-0.5 < r < 0.5 (ニュートラル)	r < -0. 5 (ネガティブ)
アラート	0.56	-0.07	-0.97

4.4 2014 年度ニュースの期待リターン

2013 年度英語ニュースの学習に対し、2014 年度英語ニュースの期待リターン分析を行った結果、ポジティブとネガティブのリターンを再現することが出来なかった。しかしながら、期待リターンはポジティブ分類の際に大きく算出された。

表 6. 2014 年度ニュースの期待リターン

ニュースの属性	r > 0.5 (ポジティブ)	-0.5 < r < 0.5 (ニュートラル)	r < -0. 5 (ネガティブ)
アラート	0.47	0.15	0.20

5. まとめ

今回、韓国株式市場における企業ニュースと株式リターンの関係について分析を試みた。企業のニュース(アラート)がリターンにもたらす結果を LSTM network を用いてモデル化し、分類分析評価を行った。

その結果、LSTM による分類アルゴリズムの構築ができた。未だ検討中ではあるが、ポジティブ分類ニュースにおいて期待リターンが高く出る傾向を得た。

謝辞

本研究の一部は、柏森情報科学振興財団の研究助成を受けたものである。

参考文献

[1] Antweiler, W., and M. Z. Frank “Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards," Journal of Finance 59(3), pp. 1259-1293. (2004)

[2] Tetlock, P. C. “Giving Content to Investor Sentiment : The Role of Media in the Stock Market," Journal of Finance 62(3), pp. 1139-1168. (2007)

[3] Tetlock, P. C., M. Saar-Tsechansky and S. Macskassy “More Than Words : Quantifying Language to Measure Firms ’Fundamentals," Journal of Finance, 63(3), pp. 1437-1467. (2008)

[4] Engelberg, J., A. V. Reed, and M. C. Ringgenberg “How Are Shorts Informed? Short Sellers, News, and Information Processing," Journal of Financial Economics 105(2), pp. 260-278. (2012)

[5] Dougal, C., J. Engelberg, D. Garcia, and Parsons C. A. “Journalists and The Stock Market," Review of Financial Studies 25(3), pp. 639-679. (2012)

[6] 五島圭一・高橋大志「ニュースと株価に関する実証分析-ディープラーニングによるニュース記事の評判分析-」『証券アナリストジャーナル』(2016)

[7] 五島圭一・高橋大志「ティックデータを用いたニュースと日本株式市場との関連性分析」人工知能学会 2016 年度全国大会 (2016)

[8] Kim Y.M., Willett T.D. “News and the Behavior of the Korean Stock Market during the Global Financial Crisis”, Korea and the World Economy, Vol. 15, No. 3 (December 2014) 395-419 (2014)

[9] Lee D.W., Cho J.H. “Stock Price Reactions to News and the Momentum Effect in the Korean Stock Market”, Asia-Pacific Journal of Financial Studies (2014) 43, 556–588

[10] 宮崎邦洋・松尾豊 [2017] “深層学習を用いた株価予測分析” 人工知能学会 2017 年度全国大会

[11] 松井藤五郎・汐月智也 [2017] “LSTM を用いた株価変動予測” 人工知能学会 2017 年度全国大会

[12] 五島圭一，高橋大志，寺野隆雄 [2017] “リカレントニューラルネットワークによるボラティリティ・クラスタリングのモデル化,” 情報処理学会 第

79 回全国大会 (2017)

[１３] E.Chong, C.Han, F.C. Park [2017] “Deep learning networks for stock market analysis and prediction : Methodology, data representations, and case studies,” Expert Systems With Applications 83, 187-205 (2017)

[１４] S. Hochreiter, J. Schmidhuber “Long Short-Term Memory,” Neural Computation 9(8) pp.1735-1780 (1997)

[１５] F. Gers, J. Schmidhuber, F.Cummins [2000] “Learning to forget continual prediction with LSTM,” Neural Computation 12, 2451–2471 (2000)

[１６] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean “Efficient estimation of word representations in vector space” 2013, ICLR