要約文章と機械学習による株価変動の分類モデルの構築

Classification and Prediction of Stock Price Using Summarization and Machine Learning

野矢淳[1]　高橋大志[1]
Jun Noya[1], and 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

要約:　情報技術の発達に伴い，株式市場分析においても情報の重要性が注目されてきている．本研究は，ニュース記事に着目し，情報技術を用いて投資家が判断や意思決定を支援するモデルの構築を目指す．その方法として，TOPIX Core30 の構成銘柄のニュース記事の要約文章を生成し，分散表現に変換して機械学習で分析を行った．

1 はじめに

近年，あらゆるモノがインターネットに繋がるInternet of Things（IoT）が急速に広がっている．IoTによって，センサーやデバイスの量が増え，またそれらがインターネットで相互に繋がることから，生成されるデータの量が爆発的に増えている．株式市場においても同様のことが言え，どのように必要な情報を抽出するかが課題となっている．また，株式市場の株価は業績や景気の変動などといった様々な情報によって変動する．公開されている情報は投資家に平等に与えられるが，投資家がその情報を入手し，読み取り判断に要する時間は一人ひとり異なる．そして，この間にも株価が変動する可能性がある．そのため投資家は情報を素早く受け取り，膨大な量の情報から必要な情報のみを取り出して的確に判断する必要がある．

そこで機械学習や自然言語処理を用いてこの問題を解決することを試みた．近年の機械学習や自然言語処理の発達は目覚ましく，膨大な情報を扱うことが出来るようになり，処理速度や予測精度も向上している[1-3]．このように機械学習を用いてニュースと株価の関連性を分析した研究は数多く行われている[4-9].

そのため本研究は，投資家が判断や意思決定するまでの時間を短縮化することを目指し，要約文章の生成，及びそれらを株価の説明モデルへ適用することを目的とする．また，要約文章が過去の情報を含んだ文章であるとすると，新しい情報が発表されたときに，過去に発表されていない情報は株価への影響度が大きく，発表された情報は株価への影響度が小さいと仮定することができる．これらを株価の説明モデルへ組み込むことができれば，株価の予測精度の向上の寄与する可能性がある．

2 データ

本研究では，要約文章生成のためにニュース記事と株価の結びつけに株式市場データを使用する．

分析の対象は，TOPIX Core30 の構成銘柄（日本たばこ産業，セブン＆アイ・ホールディングス，信越化学工業，武田薬品工業，アステラス製薬，日立製作所，パナソニック，ソニー，キーエンス，デンソー，ファナック，村田製作所，日産自動車，トヨタ自動車，ホンダ，キヤノン，三井物産，三菱商事，三菱 UFJ フィナンシャル・グループ，三井住友フィナンシャルグループ，みずほフィナンシャルグループ，東京海上ホールディングス，三井不動産，三菱地所，JR 東日本，JR 東海，日本電信電話，KDDI，NTT ドコモ，ソフトバンク）とした．また分析の期間は 2016 年 1 月から 2017 年 12 月までの２年間とした．

2.1 ニュース記事データ

ニュース記事データは，トムソン・ロイター社が提供するロイターニュースを用いた．トムソン・ロイター社は世界最大級のマルチメディア通信社であり，日本においても幅広いニュースを提供している．特にトムソン･ロイター社の報道スピート，正確性，信頼性は高い評価を得ており，数多くの投資家が活用する情報源である．本研究では，分析対象に関連する英語のニュース記事の発信日時及び本文を使用した．

2.2 株式市場データ

株式市場データは，東京証券取引所における株式の約定価格や約定数量の推移を示した時系列の取引データであるティックデータを用いた. 本研究では，分析期間における分析対象の約定時刻と約定価格を使用した．この株式市場データは，ニュース記事が公開された時刻の前後 1 分の価格を得るために用いた．また，1 分の間に 1 回以上取引が記録されている場合は，その取引金額の平均値を分析に用いた．なお．取引時間外のニュース記事については分析の対象外とした.

3 分析手法

本研究の分析手法の構造を，図 1 に示す．まず，対象期間における対象企業のニュース記事を取得し，これをニュース記事データとした．また，ニュース記事が発表された前後 1 分の株価を取得し，株価の変動率を算出した後にラベル付けを行った．これを株価データとした．次に，ニュース記事データから自然言語処理モデルを用いて, 要約文章を生成した．その要約文章及び，次の期に発表されたニュース記事を自然言語処理モデルを用いて，それぞれ分散表現に変換した．それら分散表現を用いて，２文章の類似度を算出した．この類似度及び分散表現，株価データを機械学習に入力することで，ニュースと株価の関連性を分析した.

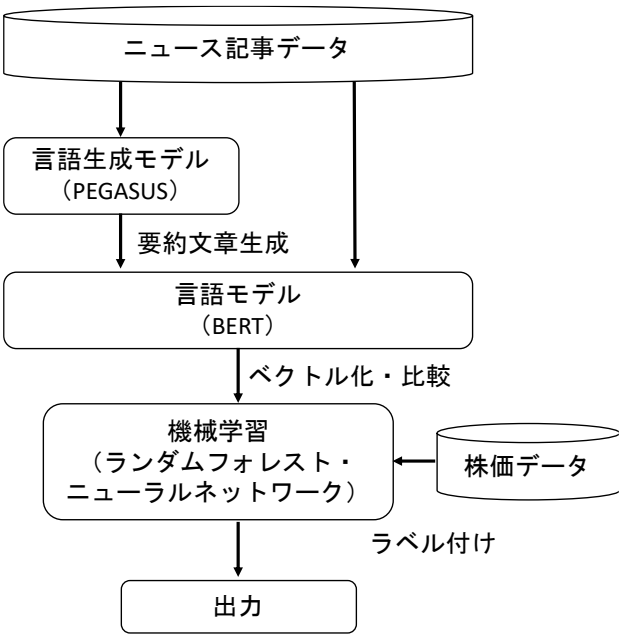

図 1　分析手法の構造

3.1 PEGASUS によるニュース記事の要約

Pre-tranning with Extracted Gap-sentenses for Abstractive Summarization (PEGASUS) [10] は，Transformer がベースの自然言語処理モデルであり，文章要約に特化している．2020 年初頭に要約文章生成タスクにおいて State of the Art (SOTA) を達成した自然言語生成モデルである.

本研究では関連銘柄のニュース記事を読み込み，要約文章を生成するのに用いた.

3.2 BERT による分散表現

Bidirectional Encoder Representations from Transformers(BERT)[11]は，Google が開発した自然言語処理言語モデルである．事前学習した後に転移学習によって様々なタスクを解くことができる．このモデルの最大の特徴は文章の文脈を理解することであり，文章の意味比較や．他の文章が続く可能性などを計算できる.

本研究では，対象銘柄に関連したニュース記事及び要約文章を，文章ごとに分散表現へ変換した．また, BERT を使用する際には， 512 単語（トークン）以下にする制約がある. そのため, 513 以上のトークンがある場合は，512 以下のトークンになるように切り捨てた.

3.3 機械学習による分析

本分析では，Random Forest（RF）と Neural Network(NN)を用いた．RF は，複数の機械学習アルゴリズムを組み合わせたアンサンブル学習の一つであり，複数の条件分岐の集まっている，木構造のアルゴリズムである決定木を組み合わせたアルゴリズムである．NN は脳のように動作する機械学習モデルであり，複数のノード（ニューロン）が結合して構成されているアルゴリズムである.

本分析では，対象銘柄に関連したニュース記事及び分散表現を入力とし，株価データを出力として学習することで，分析モデルを構築した.

4 分析結果

4.1 要約文章の生成

PEGASUS を用いて，各社のニュース記事を 1 ヶ月毎に要約した．その結果の例を図 2 に示す．本要約文章は，2016 年 8 月の日産の記事を要約した文章である．この月に日産は，効率の高い新しいエンジンを開発した [12]. 要約文章もその内容に即していることが分かる．この結果から，可読性の高い要約文章を生成できたことを確認できる.

Nissan Motor Co Ltd ?? 7201.T> has come up with a new type of gasoline engine .<n>The new engine uses variable compression technology .<n>Could replace some of today's advanced diesel engines .

図 2　生成した要約文章の例

4.2 要約文章を使った株価の分類モデル

本分析では，株式市場の価格変動を用いてニュースのラベル付けを行った．その結果は，表 1 のようになった．ニュース記事合計 2045 件のうち，ニュース記事が発表された時刻の前後 1 分の株価の変化率が 0.05 以上である High ラベルは 948 件，0.05 以下である Low ラベルは 1097 件であった．

表 1 ラベル付けの結果

ラベル	ニュース数（件）
High	948
Low	1097
合計	2045

これをもとに，要約文章と株式市場の分析を行った．本分析では，過去のニュースを要約した文章と既存の手法の比較を行った．比較対象は，BERT で分類学習を行った Model 1，要約文章との類似度を機械学習で分析した Model 2，要約文章との類似度及び分散表現を機械学習で分析した Model 3 である．Model 2 と Model 3 はそれぞれ，NN と RF の 2 種類の分析を行った．

分析の結果は図 3 の通りである．正答率は，Model 1 が 53.02%，Model 2 (RF)が 58.80%，Model 2 (RF) が 54.73%，Model 3 (RF)が 61.17%，Model 3 (NN)が 56.78%となった．要約文章との類似度及び分散表現を用いた Model 3 の正答率が一番高く，次に要約文章との類似度を用いた Model 2，BERT を用いた Model 1 となることを確認した．

この結果より，提案した要約文章との類似度を組み込んだ手法が，株式分析の精度向上に寄与する可能性があることが分かった．さらに，要約文章との類似度と分散表現を組み合わせる手法は，要約文章との類似度のみを用いた手法より高い正答率となった．分散表現を組み合わせた方が，アーニングサプライズ以上の情報を含ませることができるため，このような結果になったと考えられる．

5 まとめ

本研究は，投資家が判断や意思決定するまでの時間を短縮化することを目的とし，要約文章の生成及びそれを株価の説明モデルへの適用を行った．その方法として，自然言語処理を用いて，要約文章を生成した．さらに，要約文章が過去の情報を含んだ文章であるとして，新しい情報が発表されたときに，過去に発表されていない情報は株式市場への影響度が大きく，発表された情報は株式市場への影響度が小さいと仮定し，株式市場分析に含めて株 s 価の説明モデルの精度の向上を試みた．

本研究の結果は，ニュース記事を要約した結果，可読性の高い文章が生成できることを示し，またそれらを株価の説明モデルに組み込むことで精度の向上の可能性を示すものである．

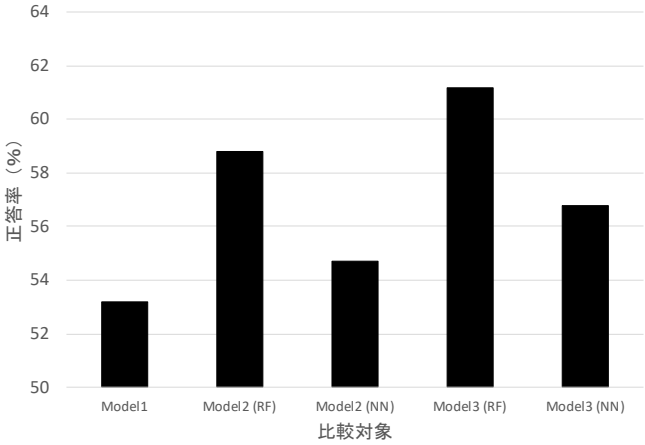

図 3 分析手法の比較の結果

参考文献

[1] Q. &. M. T. Le, “Distributed representations of sentences and documents,” International conference on machine learning, pp. 1188-1196, 2014.

[2] I. V. O. &. L. Q. V. Sutskever, “Sequence to sequence learning with neural networks.,” neural information processing systems, pp. 3104-3112, 2014.

[3] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I, “Language models are unsupervised multitask learners.,” OpenAI blog, 2019.

[4] Kara, Yakup Acar Boyacioglu, Melek Baykan, Ömer Kaan, “ Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange,” Expert Systems With Applications, pp. Vol.38(5), pp.5311-5319, 2011.

[５] Muh-Cherng Wu, Sheng-Yu Lin, Chia-Hsin Lin, “An effective application of decision tree to stock trading,” Expert Systems with Applications, pp. 270-274, 2006.

[６] Luckyson Khaidem, Snehanshu Saha, Sudeepa Roy Dey, “Predicting the direction of stock market prices using random forest,” 2016.

[７] Yoshihiro Nishi, Aiko Suge, Hiroshi Takahashi: News Articles Evaluation Analysis in Automotive Industry Using GPT-2 and Co-occurrence Network, In: Sakamoto, M., Okazaki,N., Mineshima, K., Satoh K. (eds) New Frontiers in Artificial Intelligence, Lecture Notes in Computer Science, pp.103-114, 2020.

[８] Yusuke Matsumoto, Aiko Suge, Hiroshi Takahashi: Capturing Corporate Attributes in a New Perspective through Fuzzy Clustering, In: Kojima K., Sakamoto M., Mineshima K., Satoh K. (eds) New Frontiers in Artificial Intelligence, Lecture Notes in Computer Science, vol 11717. Springer, pp.19-33, 2019.

[９] Shohei Fujiwara, Yusuke Matsumoto, Aiko Suge, Hiroshi Takahashi: Constructing a Valuation System Through Patent Document Analysis, In: Jezic G., Chen-Burger J., Kusek M., Šperka R., Howlett R., Jain L. (eds) Agents and Multi-agent Systems: Technologies and Applications 2020. Smart Innovation, Systems and Technologies, vol 186. pp.355-366, Springer, 2020.

[１０] Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu, “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” 著: International Conference on Machine Learning, 2020.

[１１] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., “Bert: Pre-training of deep bidirectional transformers for language understanding.,” 2018.

[１２] webCG, “日産が世界初の量産型“可変圧縮比エンジン”「VC-T」を開発,” webCG, 17 8 2016. [オンライン]. Available: https://www.webcg.net/articles/-/34941.