

自動車メーカーの株価変動に基づいた機械学習による 分類モデルの構築と要約文章を用いた分析

Construction and Analysis of News Evaluation Model for Automobile Manufacturers Using Summarization and Similarity

野矢淳¹ 高橋大志¹

Jun Noya¹, and Hiroshi Takahashi¹

¹ 慶應義塾大学大学院経営管理研究科

¹Graduate School of Business Administration, Keio University

要約: 株式市場の変動は様々な要因によって引き起こされる。とりわけ、ニュースは最も一般的な情報源であり、投資家にとって重要な役割を果たしている。本研究では、ニュースのサプライズに注目して分析を行った。分析の手法として、過去のニュース記事を要約し、次期に発表されたニュースと比較することによってアーニングサプライズを株価の説明モデルに組み込んだ。要約文章生成には PEGASUS、分類モデルの構築にはランダムフォレストを採用した。

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、投資家はその情報を入手し、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いく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、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、自動車メーカー大手3社、日産自動車株式会社(日産)、トヨタ自動車株式会社(トヨタ)、本田技研工業株式会社(ホンダ)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ド、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。

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。

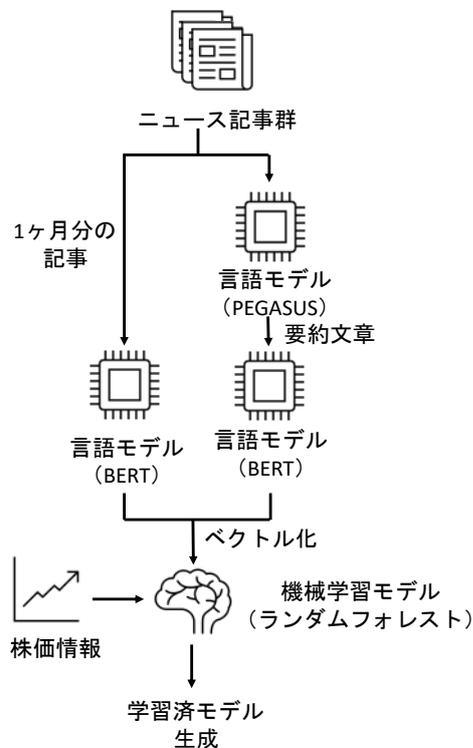


図1 分析手法の構造

3.1 PEGASUS によるニュース記事の要約

Pre-training with Extracted Gap-sentences 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。ニュース記事データを銘柄、月ごとに分け、該当する記事を全てPEGASUSに入力し、要約文章を生成した。対象銘柄が3社、対象期間が24ヶ月あるため、合計72個の文章を出力した。

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 ランダムフォレストによる分析

Random forest (ランダムフォレスト)とは、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。本研究では、分散表現及び文章の類似度と株価データとの関連性の分析にランダムフォレストを用いた。本研究は、分散表現及び類似度をラベル化した株価データと結びつける分類タスクである。分散表現のような説明変数が多数存在していても学習が可能で、予測精度が高いランダムフォレストを用いることとした。

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

まず初めに、有用な要約文章の生成が可能であるかを確認する。自然言語処理モデルを用いて、過去のニュース記事から要約文章を生成した。次に、過去のニュース記事を要約した文章が、株式市場分析において有効であることを確認した。

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974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339件、0.05以下である Low ラベルは635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	339
Low	635
合計	974

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、分散表現のみを機械学習で分析した Model 1 (既存手法)、要約文章との類似度のみを機械学習で分析した Model 2 (提案手法)、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 (既存手法+提案手法)

である。

分析の結果は図3の通りである。正答率は、Model 1が63.07%、Model 2が57.85%、Model 3が64.87%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、次に分散表現のみを用いた Model 1、要約文章との類似度のみを用いた Model 2 となることを確認した。

この結果より、提案手法である要約文章との類似度のみを用いた手法では十分に株価の変動を説明できていない。一方で、要約文章との類似度と分散表現を組み合わせる手法は、既存の分散表現のみを用いた手法より高い正答率となった。

要約文章との類似度のみを用いた場合は十分に株式市場を説明できていない要因について以下のように考えた。まず要約文章との類似度を用いてアーニングサプライズが適切に測れていなかった可能性がある。例えば、PEGASUS による要約文章の生成時に全ての情報を含むことができなかった可能性や、BERT による分散表現への変換が上手く機能しなかった可能性がある。これは現状の自然言語処理の技術の限界である。

また、本研究で測定したアーニングサプライズと株式市場の関係性が薄い可能性がある。これは、アーニングサプライズ以外の要因、例えば文章の語句や意味の方が株価への影響が大きいと、適切に反映されなかったのではないかと考える。そのため、分散表現と要約文章との類似度の両方を用いた場合の正答率が大きかったのであると考えた。

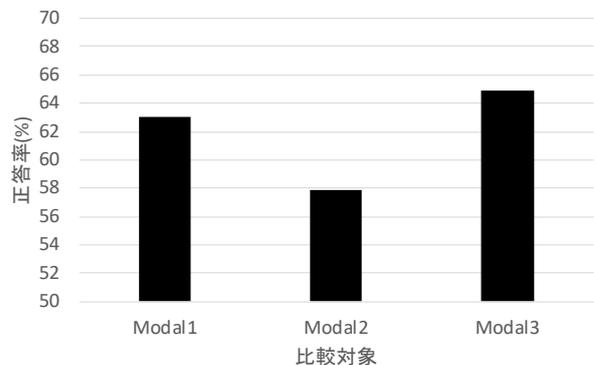


図3 要約文章と既存手法の比較の結果

次に、過去のニュース記事との類似度と過去の要約文章の比較を行った。これは要約文章が過去の個別のニュース記事を用いて分析するより、株式市場分析において優位であることを確認するためである。比較対象は、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 1 (要約文章を用いた手法)、個別記事との類似度及び分散表現を機械学習した

Model 2 (個別記事を用いた手法) である。

分析の結果は図4の通りである。正答率は、Model 1が64.87%、Model 2が64.25%となった。要約文章との類似度を用いたModel 1の正答率が、個別の記事との類似度を用いたModel 2より僅かながら正答率が高いことを確認した。

つまり、要約文章が個別記事と同等もしくは多少ではあるがよりよく株式市場を説明できている。この要因について以下のように考えた。要約文章は個別記事のまとめたものであり、個別記事より多くの情報を含んでいる。そのため、要約文章を用いた手法の方が、個別の記事を用いた手法より正答率が高くなっている可能性がある。詳細な分析は今後の課題である。

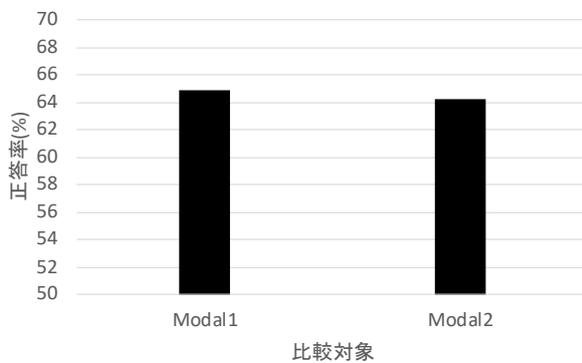


図4 要約文章と個別記事分析との比較の結果

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。

今後の課題として、さらなる検証を行うために、対象企業数や対象期間を増やしてデータ数を増やすことが挙げられる。また、ニュース記事の他、企業のホームページや有価証券報告書から分析を行う方法も検討している。

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, Snehanstu 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, 178 2016. [オンライン]. Available: <https://www.webcg.net/articles/-/34941>.