

要約文章生成と意味類似度による 株式市場におけるニュース記事の影響度の測定

Analysis the Effect of News Articles on the Stock Market
by Summarizing and Semantic 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

¹ Graduate School of Business Administration, Keio University

Abstract: Stock market fluctuations are caused by various factors. This study focuses on news articles to improve the accuracy of stock market predictions. The effect of news articles is bigger when information that was not predicted is public, earning surprise. We have summarized past news articles by GPT-2 to discover news trends. We also converted the news articles into vectors and link them to stock prices by machine learning. As a result of this analysis, we were able to improve the accuracy of stock market predictions.

1 はじめに

株式市場の株価はさまざまな要因によって変動する。株価の挙動を説明するために、これまで数多くの研究が行われてきた。しかし複雑な株式市場をモデル化することは困難であり、現在でも株価の動きを完璧に説明することはできていない。

株価の変動は、一般的にこれまでに市場では認識されておらず、予測しづらい情報が発表されたときに特に大きく反応する。これをアーニングサプライズという。アーニングサプライズに関連した研究として、アナリスト数や企業業績によって検証したもの[1]や、アーニングサプライズが将来どのように株価と関係するか検証したもの[2]などが存在する。

また、株価は景気や企業業績、市場の動きなどに影響を受けて変動する。これらの情報を受信する媒体として代表的なものに有価証券報告書や四季報、ニュースなどがある。とりわけニュースは誰でもアクセスできる情報源であり、個人のような小さい投資家から機関投資家までが活用している。ニュースと株価の関係性を扱ったものは数多く存在している[3]。また、近年は機械学習モデルを使った研究も盛んに行われている[4][5]。

そのため本研究では、ニュース記事を要約することで算出されるアーニングサプライズをモデルに組み込むことで、株価の説明モデルの精度の向上を目的としている。その手法として、近年の機械学習の

発展により自然言語処理で文章の意味を理解することが可能となったため、言語生成モデルでニュース記事の要約文章を生成し、文章のベクトル化によって類似度を求めて分析する。

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。それぞれ以下通りである。

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、2016年のシャープ株式会社に関連するニュース記事の発信日時、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ティックデータを用いた。ティックデータとは、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のことである。本研究では、2016年の取引時間と約定価格を使用した。

3 分析手法

本研究の分析手法の構造を、図1に示す。まず一定期間のニュース記事から要約文章を生成し、次の期間に発表されたニュース記事との類似度を比較した。その後、算出された類似度より株価との関連性を分析した。要約記事の作成には GPT-2 を用いて、比較のための文章ベクトルの生成は Bert を用いた。また、株価との関連性は、ニュースが発表される前後の株価の変動率を算出し、ラベル付を行った上で、NN を用いて分析を行った。

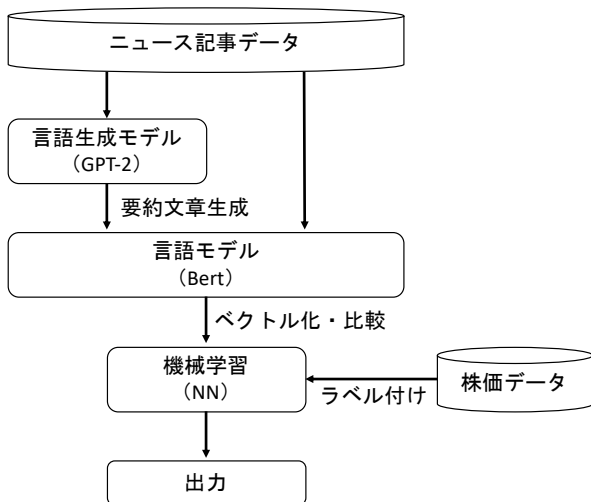


図1 分析手法の構造

3.1 GPT-2 を用いたニュース記事の要約

Generative Pretrained Transformer 2 (GPT-2)[6]は、汎用人工知能の研究をするアメリカの非営利団体である Open AI が開発した言語生成モデルである。世界中のインターネット上にある 800 万の Web サイトの文章を学習しており、読解や翻訳、要約などのタスクを得意とする。

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。要約する期間は一ヶ月毎として、その月のニュース記事の要約文章を生成した。

3.2 Bert を用いた文章のベクトル化

Bidirectional Encoder Representations from Transformers(Bert)[7]は、Google が開発した言語モデルである。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。

本研究ではニュース記事を比較するために、文章を多次元ベクトル化するのに用いた。

3.3 NN による分析

Neural Network(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。単一のノードにおいて、複数の入力($x_0, x_1 \dots x_n$)をすると、一つの出力(Y)をする(図2)。これらのノードを複数組み合わせることで最適な値を算出する。本研究では文章の類似度を比較したベクトルと株価との結びつけを行うのに用いた。

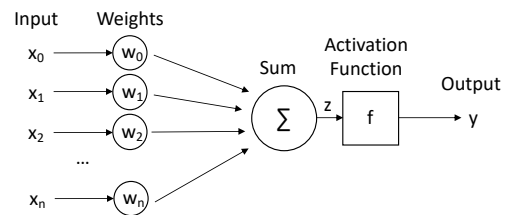


図2 NNのノードに対する入力と出力

4 分析結果

本研究では、文書をベクトル化した後に機械学習で分析した Model 1、要約文章との類似度のみを機械学習で分析した Model 2、要約文章との類似度及び文書をベクトル化した両方を機械学習で分析した Model 3 を用いて、正答率を比較した。

4.1 要約文章の生成

GPT-2 を用いて、一年間のニュース記事を一ヶ月毎に要約した。その結果の例を、図3に示す。これより可読性の高い要約文章を生成できたことが分かる。

The plan to raise Sharp's capital, saying it could raise 2 trillion yen (\$18.85 billion) with investors. Mr. Murakami, former manager of the fund, was also on the call. The government is unlikely to help Sharp's plans, some experts say that is a surprise, but the government should be helpful for Sharp. The government should offer Sharp and take Sharp Sharp and Hon Hai and SoftBank.

図3 生成した要約文章の例

4.2 文章のベクトル化と分析

Bert を用いて文章をベクトル化して要約文章との比較を行い NN で株価との関連性の分析を行った。

それぞれのモデルの結果は図4の通りである。正答率は、Model 1 が 55%、Model 2 が 35%、Model 3 が 57%となった。これより、要約文章との類似度を組み込んだモデルである、Model 3 の結果が一番高い正

答率であることが確認できた。これらの結果は、類似度の情報が資産価格評価に貢献できる可能性を示すものである。

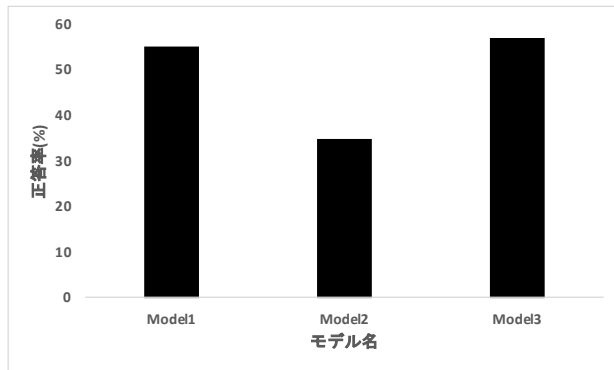


図4 NNによる分析結果

5 まとめ

本研究は、ニュース記事の要約文章生成によって、株価の説明モデルを構築した。GPT-2で要約文章を生成した後、Bertでベクトル化してNNで分析を行った結果、類似度と文章ベクトルを組み込んだモデルの精度が最も高いことを見出した。これらの結果は、ニュース記事と要約文章の類似度を通じサプライズを組み込むことで、株価の説明モデルの精度の向上の可能性を示すものである。

今後はデータ数を増やして結果をさらに検証する必要がある。また要約文章を生成するにあたってGPT-2の後継モデルであるGPT-3を使うことを検討している。

参考文献

- [1] Abraham, Rebecca, Harrington, Charles: Predictors of the Degree of Positive Earnings Surprises, *Open Journal of Accounting*, Vol. 05, pp. 25-34, (2016)
- [2] Panos Patatoukas, Hongjun Yan: "The Impact of Earnings Surprises on Stock Returns: Theory and Evidence," Yale School of Management, (2009)
- [3] Gidófalvi, Győző: Using news articles to predict stock price movements, University of California, San Diego, (2001)
- [4] Joshi, Kalyani, N, Bharathi, Rao, Jyothi: Stock Trend Prediction Using News Sentiment Analysis, *International Journal of Computer Science and Information Technology*, Vol. 8, pp. 67-76, (2016)
- [5] M. G. Sousa, K. Sakiyama, L. d. S. Rodrigues, P. H. Moraes, E. R. Fernandes, E. T. Matsubara: BERT for Stock

Market Sentiment Analysis, 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1597-1601, (2019)

- [6] Radford Alec, Wu Jeffrey, Child Rewon, Luan David, Amodei Dario, Sutskever Ilya: Language models are unsupervised multitask learners, *OpenAI Blog*, Vol. 1(8), pp. 9, (2019)
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *NAACL-HLT 2019*, (2018)
- [8] Yoshihiro Nishi, Aiko Suge, Hiroshi Takahashi: Text Analysis on the Stock Market through "Fake" News Generated by GPT-2, *INFORMS Annual Meeting*, (2019)
- [9] Yoshihiro Nishi, Aiko Suge, Hiroshi Takahashi: Construction of News Article Evaluation System Using Language Generation Model., 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)