GPT-2 によるニュース記事生成を用いたニュース評価モデルの構築と共起ネットワーク分析

English Title (Style “英文タイトル”)

西　良浩[1]　菅　愛子[1]　高橋　大志[1]

Yoshihiro Nishi[1], Aiko Suge[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

Abstract: News articles have great impacts on asset prices in the financial markets. Many attempts have been reported to ascertain how news influences stock prices. However, the limitations in the number of available data sets usually become the hurdle for the model accuracy. In this study, we propose a news evaluation model utilizing GPT-2. A news evaluation model is a model that evaluates news articles distributed to financial markets based on price fluctuation rates and predicts fluctuations in stock prices. Reuter's news texts are classified based on each return through LSTM models. Using co-occurrence network analysis, we reviewed the overview of the news articles retrieved. News articles generated by GPT-2 was used with original news articles, and the model accuracy was examined. The results showed that created news articles are influential over the prediction of stock price fluctuation.

1. はじめに

金融市場において配信されるニュース記事は資産価格評価に影響を与える重要な情報である．ニュース記事が株価に与える影響に関してはこれまでに多くの取り組みが報告されており，金融市場の資産価格に大きな影響を与えることが示唆されている[3][4][6][9]．ニュースは有用な情報を多く含む一方で, ニューステキストのデータ取得には制約があり，データ量は限定的となる傾向がある．

近年は情報技術の進展に伴い，マシーンラーニングやディープラーニングを用い，非構造化データを自然言語処理により分析する手法が多く存在している．金融市場において発信されたニュースを用いた株式変動に関する分析をマシーンラーニングやディープラーニングを介して行うことで新たな示唆を得る手法が模索されている．

本研究では，金融市場において配信されたニュース記事の共起ネットワーク分析と企業の株価に与える影響を Long Short-Term Memory(LSTM)[7]を介したニュース分類分析により評価するニュース評価モデルの構築を行った．また，大規模言語生成モデルである GPT-2[12][13]を用いてニュース記事の生成を行い，ニュース評価モデルの正解率向上を試みた．

2. 関連研究

金融市場において配信されたニュースが株価の変動に与える影響に関して分析を行った取り組みは数多くある．例えば，ニュース記事のテキストマイニングにより株価変動を分析した研究では，ニュース記事に含まれるファンダメンタルおよびセンチメントに関する情報が株価に反映されている可能性が報告されている．ニューステキストをナイーブ・ベイズ分類器によって分類し，株価との関係について分析した取り組み[6]，ニューステキストを SVM により分析した取り組み[3][9][15]などの取り組みがこれまでに報告されている．

自然言語処理の分野においては，文書生成に関する研究は活発に行われている[2]．質問に対する回答文の生成を行った取り組み[1]，ニュース記事の生成を行った取り組み[11]，発話に対する回答文の生成を行った取り組み[16]など，いくつかの取り組みが報告されている．

3. データ

本研究では，分析の対象期間を 2014 年から 2016 年までとし，ニュースデータとマーケットデータを用いて分析を行った．2019 年時点で時価総額が最も高い上位 3 社（トヨタ自動車株式会社，日産自動車

株式会社，本田技研工業株式会社）を主要な自動車企業とし，分析対象とした．

3.1 マーケットデータ

マーケットデータには，取引成立価格や取引量などの株式取引に関する情報が含まれており，各行にマイクロ秒単位のタイムスタンプが付されている．2014 年から 2016 年までのトヨタ自動車株式会社，日産自動車株式会社，本田技研工業株式会社に関するマーケットデータ 14 億 990 万 1,961 件を取得した．

3.2 ニュースデータ

ニュースデータとして，トムソン・ロイター社が配信を行ったニュースを取得した．日本企業に関するニュースは主として英語もしくは日本語により配信されている．配信されたニュースのテキスト情報には，ヘッドラインと本文があり，ヘッドラインは本文内の重要な内容を要約したテキストデータである．ニュースには配信された日時のタイムスタンプが付されている．

本研究では，ニュース配信の前後 1 分間に取引があった英語のヘッドラインを用いて分析を行う．2014 年から 2016 年までのトヨタ自動車株式会社，日産自動車株式会社，本田技研工業株式会社に関するニュース 2,259 件を取得した．表 1 は取得したニュースの内訳を各社ごとに示したものである．取得した 2014 年から 2016 年までにトムソン・ロイター社が配信を行ったトヨタ自動車株式会社に関するニュースは 1,065 件，日産自動車株式会社に関するニュースは 587 件，本田技研工業株式会社に関するニュースは 607 件であった．

表 1：取得したニュースの件数

	件数
トヨタ自動車	1,065
日産自動車	587
本田技研工業	607
合計	2,259

4. 分析手法

4.1 ニュース記事の共起ネットワーク分析

金融市場で配信されるニュース記事には，期間中における該当企業に関するトピックが主として含まれている．共起ネットワーク分析では抽出されたエッジを使用して，類似した発生パターンを持つ単語を接続した[5]．共起ネットワークを用いて，取得したニュース記事の主要なトピックに関する情報の取得を行った．分析期間中に 50 回以上配信された記事に含まれる語彙を抽出し，Jaccard 係数を使用して各語彙の共起関係を計算した[10]．

4.2 GPT-2 によるニュース記事生成を用いたニュース評価モデルの概観

本研究では，GPT-2 を用いてニュース記事の生成を行った．図 1 に構築した 3 つのニュース評価モデルのアーキテクチャを示す．本研究において構築したニュース評価モデルは Labeling，Vectorization，Classification Layer の 3 つを通じて，ニュース評価を行った．

図 2 は，従来の研究と本研究の比較を表している．従来の研究は，オリジナルのニュース記事のデータとマーケットデータを用いて分析を行う事が主流であった．しかしながらオリジナルのみを分析データとして用いる場合，取得できるデータの数に制限があり，データ数の制限はニュース評価モデルの精度の制限となっていた．本研究では，オリジナルのニュース記事だけでなく，GPT-2 により作成したニュース記事をデータベースに追加し，ニュース評価モデルの正解率向上構築を試みた．

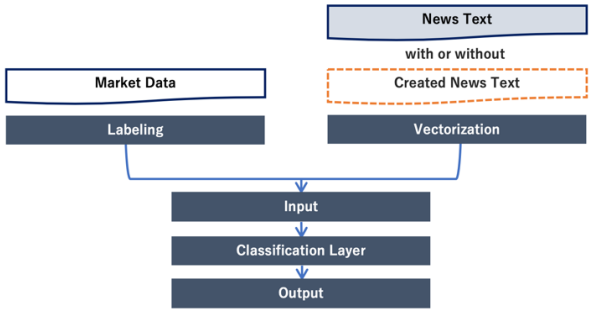

図 1：ニュース評価モデルのアーキテクチャ

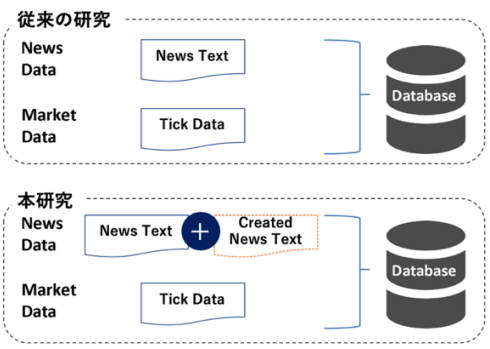

図 2：従来の研究と本研究の比較

4.3 株式変動率に基づくニュース記事のラベル付け

Labeling では，取得したニュース記事にラベル付けを行った. ラベル付けには(1)の定義式を用いた. 図 3 はラベル付けの方法について示している．ニュース配信前後のマーケットデータを取得し，（1）の定義式により株価変動率を求め, 2014 年から 2016 年までのトヨタ自動車株式会社, 日産自動車株式会社, 本田技研工業株式会社に関するニュース記事 2,259 件にラベル付けを行った．ラベルは Positive と Negative の二値とし，$\alpha > 0\%$ の場合は Positive，$\alpha < 0\%$ の場合は Negative とし, ラベル付けを行った.

$$\text{株式変動率}(\%) = \frac{(\text{ニュース配信１分後の平均株価}) - (\text{ニュース配信１分前の平均株価})}{(\text{ニュース配信１分前の平均株価})} \times 100$$

$$\text{Positive: } \alpha > 0\%$$
$$\text{Negative: } \alpha < 0\% \qquad (1)$$

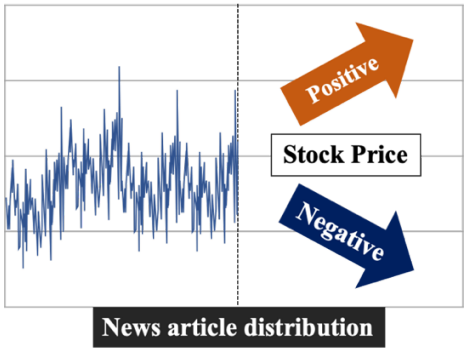

図 3: ラベル付けの方法

4.4 GPT-2 を用いたニュース記事の生成

GPT-2 とは多くのタスクで SOTA を達成した最新の大規模言語生成モデルである[12]．本研究において使用する GPT-2 のモデルは，800 万の Web ページ（計 40GB）を 24 層のネットワークで，およそ 3 億 5,000 万個のパラメータを用いて学習している[12][13]．取得した Positive なニュースと Negative なニュースを元に，ラベルごとに新たにニュース記事を生成した． Positive なニュース記事と Negative なニュース記事を 1,000 件ずつ生成し, 計 2,000 件のニュース記事の生成を行った.

4.5 Word2vec を用いたニュース記事のベクトル化

Vectorization では, ニュース記事のベクトル化を行った．ニュース記事の単語ベクトル学習には Word2vec を用いた. ニュース記事のベクトル化には, 最も広く用いられている word2vec の Skip-gram を用いた[8]．文書中の中心の単語から周辺の単語を予測するモデルを Skip-gram という．Skip-gram は，$W_1, W_2, \ldots, W_t$の順で単語が現れる場合に，(2) の定義式を用いて確率変数の対数の項を最大化するベクトルを学習により探索を行う.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(W_{t+j}|W_t) \qquad (2)$$

$p(W_{t+j}|W_t)$は，Hierarchical Softmax を用いて計算を行なっている．Hierarchical Softmax は，頻度の高い単語の順にハフマン木を作成し，階層ごとにロジスティック回帰を用いて全結合ソフトマックスに近似させる手法である[14].

4.6 LSTM によるニュース記事の分類分析

Classification Layer では，取得したニュースと付与したラベルを元に，LSTM を介してニュース分類分析を行った．LSTM は時系列データを学習する RNN の一種である．LSTM は RNN を拡張しており，長期的な依存関係の学習を可能としている[7]．分類分析に LSTM モデルを使用し，精度検証にはクロスバリデーションスコア（正解率）を用いた.

本研究では Keras の Sequential モデルを用い，LSTM による分類モデルを構築した．Classification Layer に LSTM を使用し，比較検証にはクロスバリデーションスコア（正解率）を用いた．各 Layer を Keras の. add（）メソッドにより追加し，二値分類用に Compile を設定した．loss='binary_crossentropy', optimizer='rmsprop', metrics=['accuracy']とし，分析を行った.

5. 分析結果

5.1 共起ネットワーク分析

図 4 は，取得した 2014 年から 2016 年までのニュース記事を用いて生成したトヨタ自動車株式会社,

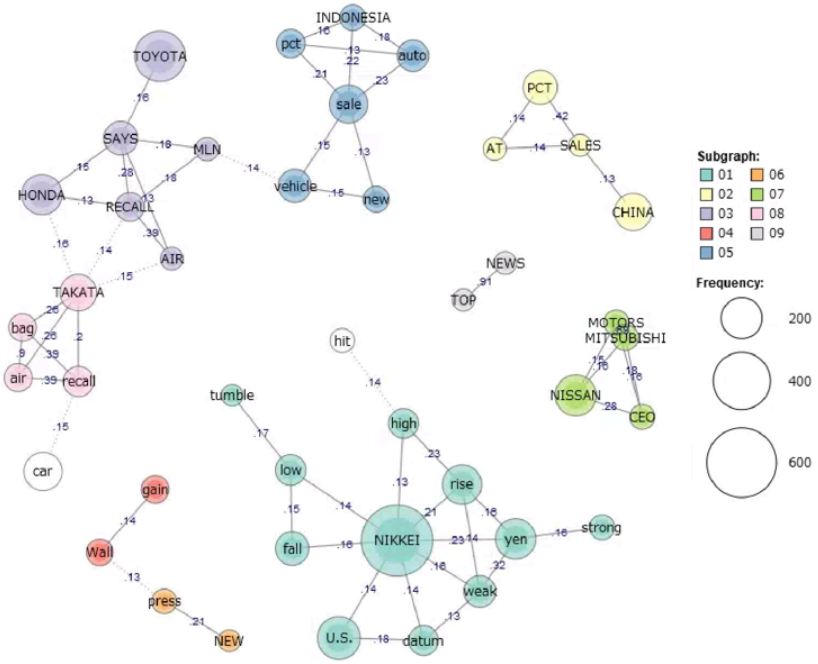

図 4: 取得したニュース記事の共起ネットワーク

日産自動車株式会社，本田技研工業株式会社に関するニュース記事の共起ネットワークである．円の色は、各語彙を含むグループを示している．円のサイズは、各語彙の出現回数を示している．円を繋ぐ点線はグループ間の共起関係を示しており，実線はグループ内の共起関係を示している．Jaccard 係数により，各語彙の共起関係を計算し線上に表示している．Jaccard 係数が高いほど，共起関係が強いことを示している．

共起ネットワーク内の”NIKKEI”は，主として日経平均株価を意味する語彙である．共起関係を見ると，市場に関して良いニュースと悪いニュースの両方を配信していることが分かる．また，タカタ株式会社のエアバッグのリコール問題に関するニュース記事は，2014 年から 2016 年の分析期間中に頻繁に話題にされており，本田技研工業株式会社とトヨタ自動車株式会社が，リコール問題に主として関与していたと推測できる．

5.2 株式変動率に基づくラベル付け

（1）の定義式により，ニュースデータのラベル付けを行った．ラベル付けの結果，2,259 件のニュースは Positive なニュースが 1,137 件，Negative なニュース 1,122 件となった．表 2 はラベル付けの結果を示したものである．

表 2: ラベル付けの結果

	ニュース記事の数	ニュース記事の例
Positive	1,137	TOYOTA TO START SELLING NX COMPACT CROSSOVER SUV IN U.S. IN NOV, AIMS TO SELL 42,000 NX SUVS ANNUALLY IN U.S. - EXEC
Negative	1,122	TOYOTA MOTOR SAYS NO TRUTH TO REPORT ABOUT TIE-UP TALKS WITH SUZUKI MOTOR

5.3 ニュース記事の生成

GPT-2 を用いてニューステキストの生成を行なった． GPT-2 を用いて Positive なニュースと Negative なニュースを各 1,000 件，計 2,000 件生成し，新たなデータとしてモデル 2 のデータセットに追加した．図 5 に生成されたニューステキストの例を示す．読解可能なニューステキストが生成されており，可読性は高い．

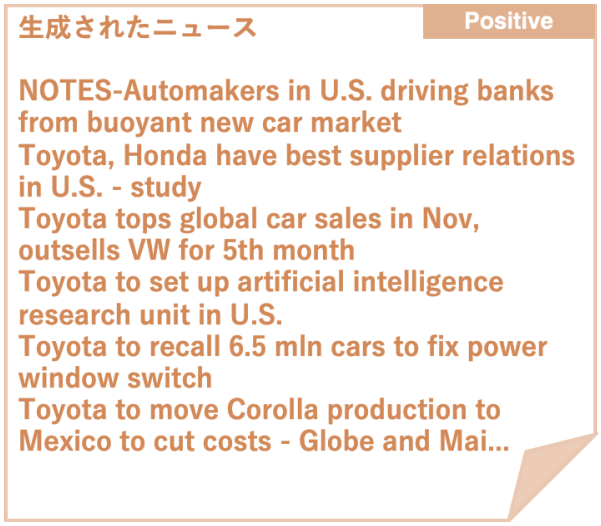

図 5: 生成されたニュース記事の例

5.4 ニュース分類分析による比較評価

Word2vec の Skip-gram モデルを用いてニュース記事をベクトル化し，LSTM を介したニュース分類分析を行った．表 3 は，各モデルの教師データ数とテストデー10 タ数を示したものである．モデル 1 とモデル 2 のデータはともに scikit-learn の train_test_split 関数を用いて，教師データとテストデータに分けた．テストサイズはどちらも 0.1 とした．

表 4 は，分類分析の結果を示したものである．LSTM による分類分析の結果，モデル 1 よりもモデル 2 の方が，クロスバリデーションスコア（正解率）が 16.9 ポイント高かった．有意水準を P＜0.05 とし，Fisher's exact test を行った結果，モデル 1 とモデル 2 は P＜0.0001 で有意であった．

表 3: 分析に用いたデータセット

	モデル 1（オリジナル）	モデル 2（文書生成）
教師データ	2,033	3,833
テストデータ	226	426
合計	2,259	4,259

表 4: LSTM によるニュース分類分析の結果

	モデル 1（オリジナル）	モデル 2（文書生成）
正解率	0.615	0.784

6. まとめと今後の課題

本研究において，金融市場において配信されたニュース記事を用いた共起ネットワーク分析とニュース評価モデルの構築を行った．共起ネットワーク分析の結果，期間中における該当企業に関する情報を得ることができた．また，GPT-2 を用いて分析に使用するニュース記事を新たに生成し，LSTM によるニュース分類分析の比較評価を行った結果，生成したニュース記事を追加したモデルの正解率は，オリジナルのニュース記事のみを用いたモデルの正解率よりも高かった．検定の結果，モデル 1 とモデル 2 の正解率には有意差があり，大規模言語生成モデルである GPT-2 をニュース記事生成に用い，分析に用いるデータ量を増大させることで，ニュース評価モデルの精度を向上させる可能性を見出した．

オリジナルのニュースデータ数を追加，金利や景気変動といったマクロの株価変動要因データを追加した分析は今後の課題としている．

参考文献

[1] Aishwarya, A., Jiasen, L., Stanislaw A., Margaret, M.C., Lawrence, Z., Dhruv, B., Devi, Parikh.: VQA: Visual Question Answering. In Proceedings of the International Conference on Computer Vision, (2015)

[2] Ehud, Reiter., Robert, Dale.: Building Natural Language Generation Systems. Cambridge University Press, (2000)

[3] Fung, G.P.C., Yu, J.X., Lam, W.: News Sensitive Stock Trend Prediction. In Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 481-493, (2002)

[4] Fung, G.P.C., Yu, J.X., Lam, W.: Stock Prediction: Integrating Text Mining Approach Using Real-time News. In Proceedings of the IEEE International Conference on Computational Intelligence for Financial Engineering, pp. 395-402, (2003)

[5] Fruchterman, T., Reingold. E.M.: 1991. Graph Drawing by Force-directed. replacement. Software—Practice and Experience, (21), pp. 1129–1164, (1991)

[6] Gidófalvi, G.: Using News Articles to Predict Stock Price Movements. Department of Computer Science and Engineering, Technical Report University of California, (2001)

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

[8] Mikolov T., Sutskever I., Chen K., Corrado G., and Dean J.: Distributed Representations of Words and Phrases and their Compositionality, In Proceedings of the NeurIPS, (2013)

[9] Mittermayer, M.A.: Forecasting Intraday Stock Price Trends with Text Mining Techniques. In Proceedings of the 37th Hawaii International Conference on System Sciences, (2004)

[10] Newman, M.E.: Modularity and Community Structure in Networks. Proceedings of the National Academy of Sciences, 103(23), pp. 8577–8582, (2006)

[11] Nishi Y., Suge A., Takahashi H.: Text Analysis on the Stock Market in the Automotive Industry through Fake News Generated by GPT-2, In Proceedings of the Artificial Intelligence of and for Business, (2019)

[12] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I.: Improving Language Understanding by Generative Pre-Training. Technical Report OpenAI, (2018)

[13] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I.: Language Models are Unsupervised Multitask Learners. Technical Report OpenAI, (2019)

[14] Rong, X.: word2vec parameter learning explained. arXiv preprint arXiv:1411.2738, （2014）

[15] Schumaker, R.P., Chen, H.: Textual Analysis of Stock Market Prediction using Breaking Financial News. In Proceedings of the ACM Transactions on Information Systems, 27, pp. 1-19, (2009)

[16] Zhang, R., Guo, J., Fan, Y., Lan, Y., Xu, J., Cheng, X.: Learning to Control the Specificity in Neural Response Generation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 1, pp. 1108-1117, (2018)