5C-03

大規模言語生成モデルによるニュース生成を用いたニュース評価モデルの構築

西　良浩†　菅　愛子†　高橋　大志†

慶應義塾大学大学院経営管理研究科†

1 はじめに

ニュースは金融市場の資産価格に大きな影響を与える．ニュースと株価変動の関係性を分析し，ニュースを評価する取り組みはこれまでに多く行われており，ニュースと株価変動の間には関連性があると報告されている[1][2]．高精度なニュース評価モデルを構築することで，金融市場において配信されたニュースが企業の株価にポジティブな影響を与えるか，ネガティブな影響を与えるかを判断する事ができる．しかしながら，ニュースや株価変動率算出のための取引成立価格など，取得できるデータの数には制限がある．この制限は，通常，ニュース評価モデルの精度の制限となる．

本稿では，大規模言語生成モデルにより生成したニュースを分析用のデータとして付加的に用い，ニュース評価モデルの高精度化を提案する．評価実験には大規模言語生成モデルである GPT-2 を用いた[3][4]．実験の結果，ニュース評価モデルの精度が向上した．

2 提案手法

マーケットデータとニュースデータを用いて，ニュース評価モデルを構築し，そこへ大規模言語生成モデルにより生成したニュースを分析用のデータとして付加的に用いたモデルの提案を行う．マーケットデータとは，取引成立価格や取引量などの株式取引に関する情報のことである．ニュースデータとは，金融市場において配信されたニュース文書のことである．図 1 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，大規模言語生成モデルを用いて分析データの数を増大させる．オリジナルのニュースデータのみだけでなく，文書生成により作成したニュースをデータベースに追加し，より高精度なニュース評価モデルを構築する手法の提案を行う．

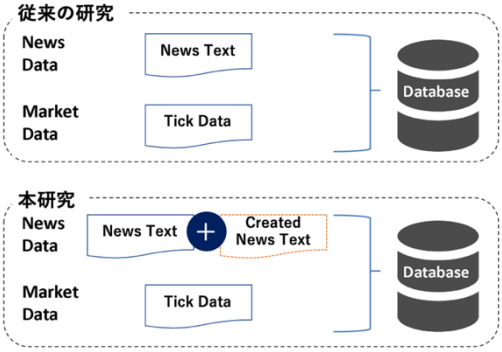

図 1: 従来の研究と本研究の比較

2.1 株式変動率を用いたラベル付け

マーケットデータとニュースデータを用いて，（1）の定義式により，株価変動率を求め，ニュースにラベル付けを行う．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，ラベル付けを行う．

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

$$Positive: \alpha > 0\%$$
$$Negative: \alpha < 0\%$$

（1）

2.2 大規模言語生成モデルの活用

ラベル付けしたオリジナルのニュースを元に，大規模言語生成モデルを用いて新たなニ

Construction of News Evaluation Model using News Generation by Large-Scale Language Generation Model
† Yoshihiro Nishi, Aiko Suge, Hiroshi Takahashi
† Graduate School of Business Administration, Keio University

ュースステキストの生成を行う．生成したニュースには，オリジナルのニュースに付与されたラベルと同じラベルを付与する．

図 2 は提案するニュース評価モデルのアーキテクチャを表している．ラベルの付いたオリジナルのニュースと生成したニュースをベクトル化し，分類分析を行う事でニュース評価モデルを構築する．

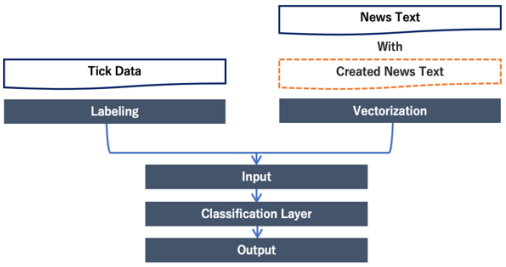

図 2: ニュース評価モデルのアーキテクチャ

3 評価実験

提案手法の有効性を示すため，トムソン・ロイター社より 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，既存手法と提案手法の比較評価を行った．取得したニュース 2,259 件のうち，Positive なニュースは 1,137 件，Negative なニュースは 1,122 件であった．ニュースの生成には，大規模言語生成モデルである GPT-2 を用いた．

3.1 GPT-2 を用いたニュース生成

GPT-2 とは 10 ベンチマークで SotA を達成した大規模言語生成モデルである．800 万の Web ページ（計 40GB）という大量の文章データを学習する事で，あらゆるジャンルの文書生成に Zero-shot で対応している．

実験に用いる GPT-2 のモデルは， 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．ラベル毎にニュースを 1,000 件ずつ生成し，Positive なニュース 2,137 件，Negative なニュース 2,122 件をデータセットとするモデル 2 を作成した．生成を行った． 例として，Positive なニュースを元に生成したニュースを図 3 に示す．人間も読む事ができる可読性の高い文書が生成されていた．

図 3: 生成されたニュースステキストの例

3.2 分析結果

ニュースをベクトル化し，ニュース分類を行った．ベクトル化には Word2Vec の Skip-gram モデルを用い，LSTM を介して分類を行った．オリジナルのニュースのみを用いたモデル 1 より，生成したニュースを加えたモデル 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．

表 1: 分類分析の結果

	モデル 1 （既存手法）	モデル 2 （提案手法）
正解率	0.615	0.784

4 おわりに

本稿にて，生成したニュースを分析用のデータとして付加的に用いるニュース評価モデルの提案を行った．評価実験の結果，提案した手法を用いたニュース評価モデルの精度が 16.9 ポイント向上した．

参考文献

[1] 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)

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

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

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