F-009

大規模ティックデータを用いたニュース記事分類と文章トピック時系列変化の可視化
News Article Classification and Visualization of Text Topic Time-Series Changes Using Large-Scale Tick Data

西 良浩† 高橋 大志†
Yoshihiro Nishi Hiroshi Takahashi

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対し影響を与える重要な情報である．企業のファンダメンタルズや投資家のセンチメントなどが，資産価格に影響を与えるとの報告が行われており [4, 7] ，ニュース記事は時流に応じてそれらに関する情報を金融市場に配信し，株価に対して影響をもたらしていると考えられる．

自然言語データを機械学習モデルや深層学習モデルを用いて分析する際には，一般的にはデータ数が多いほど分析の精度は向上する傾向にある．しかしながら，金融市場において発信されたニュース記事には限りがあり，特に特定の期間内において指定のメディア内で発信された対象企業の分析を行う場合のように範囲が絞られる場合において獲得できるニュース記事の数には制約がある．近年では，自然言語生成技術により，分析に用いるデータを増やし，精度向上を試みる取り組みがなされているが，ニュース記事が主な話題としているトピックがなにかを時系列で考慮し自然言語生成を行い，ニュース記事のようなデータを拡充する取り組みはまだ十分に行われていない．

本研究では，ニュース記事が主として扱ったトピックの時系列変化を検証し，さらに変動があった単語分布の比較を行う． 本論文の構成は次の通りである．まず第 2 章で関連研究ついて述べる．第 3 章では分析対象とするデータについて述べる．第 4 章では分析手法について述べ，第 5 章では分析結果について述べる．最後に第 6 章で結論を述べる．

2. 関連研究

ニュースと株価変動の関連性を分析した研究は数多くある．例えば，ニュースの文章を用いてテキストマイニングを行い株価変動の分析した研究では，ニュースに含まれるファンダメンタルズやセンチメントの情報が株価に反映される可能性があることが報告されている[2, 3].

近年では自然言語処理技術の発展に伴い，分析に使用するデータを自然言語処理モデルにより生成し，分析データを増やす取り組みも行われている[1, 8, 9, 10]．ニュース記事が主なトピックとして扱う話題は時系列で変化することが考えられるため，そのトピックの移り変わりを考慮した自然言語生成を行うことでより精度の高い生成ができる可能性がある．しかしながら，ニュース記事が主な話題としているトピックがなにかを時系列で考慮し自然言語生成を行い，ニュース記事の株価変動への影響を評価する取り組みはまだ十分に行われていない．

3. データ

本研究では，分析の対象期間を 2014 年 4 月から 2017 年 3 月までとし，マーケットデータとニュースデータを用いて分析を行った．2014 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

マーケットデータとして，国内上場企業のうち，自動車企業のなかから主要 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）の株式価格を対象として分析を行った．分析対象期間は，2014 年 4 月から 2017 年 3 月とした． 2014 年から 2017 年までの 4 年間分の市場取引データ約 19 億件を取得し，対象企業・対象期間に絞り，分析に使用した．本分析に用いた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．また，本分析では，企業属性に関するデータも日経 NEEDS，Datastream から取得し分析に用いた．

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記事を用いた．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 4 月から 2017 年 3 月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 2,326 件を取得した．

4. 分析手法

本研究では対象のニュース記事データに対し，トピック時系列変化の分析を行い，その後，文章トピックに変化が生じている年を対象とし年・ラベルごとにニュース記事の単語分布を比較する．

† 慶應義塾大学大学院 経営管理研究科
Graduate School of Business Administration, Keio University

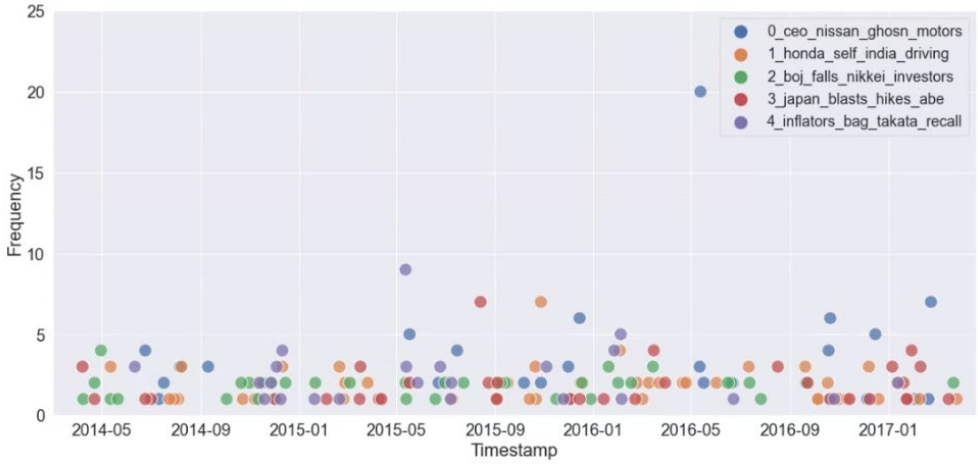

図 1　2014 年 4 月から 2017 年 3 月におけるニュース記事のトピックトレンド結果

4.1 ニュース記事のトピックトレンド分析

トピックモデルとは，教師無しでテキスト中に扱われている共通のテーマを明らかにするために有効な手法である．本研究ではトピックモデルによるニュース記事トレンドの分析に BERTopic を用いる[5]．従来のトピックモデリング技術はクラスタの重心に近い単語がそのクラスタやトピックを代表とすると仮定しているが，実際にはクラスタが常にクラスタ重心を中心とした領域の中にあるとは限らないため，トピックモデルにより算出したトピック表現の結果が誤解を招く可能性がある．BERTopic はクラスタリング技術と TF-IDF のクラスベースでのバリエーションを活用して異なるトピック表現を生成するモデルであり，より柔軟なトピックモデリングを可能としている．

本研究では 2014 年 4 月から 2017 年 3 月のニュース記事に対して BERTopic を用い 5 つのトピックを抽出し，ニュース記事トピックトレンドの変化が生じているかを検証する．

4.2 株価変動率に基づくラベル付け

次に，(1)の定義式により 2014 年 4 月から 2017 年までのニュース 記事にラベル付けを行う．該当のニュース記事が配信された前後 1 分間の平均株価を取得し，ニュース記事ごとに株価変動率を算出する．本研究では大規模な株式価格ティックデータを用いて株価変動を算出し，株価変動状況の実態をもとにして各ニュース記事のラベルを算出する．

$$\text{Stock price fluctuation rate } (\alpha) = \frac{(\text{Average stock price 1 minute after news}) - (\text{Average stock price 1 minute before news})}{(\text{Average price 1 minute before news})} \times 100$$

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

4.3 年・ラベルごと単語分布の比較

各年のラベル別単語分布の可視化および比較には Scattertext を用いる[6]．Scattertext とはラベル間の単語分布の差異を可視化するライブラリである．テキスト中に含まれる単語を発生頻度に基づき 2 次元上に散布図で出力し，傾向に違いがあるかを捉えることができる．

本研究では株価変動に基づきラベル付けしたニュース記事を用い，Positive な影響を与えたニュース記事と Negative な影響を与えたニュース記事の単語分布を可視化し，年ごとに変化が生じているかを検証する．

5. 分析結果

本研究では対象データに対し，BERTopic を介したニュース記事のトピックトレンド分析，株価変動率によるラベル付け，Scattertext による年・ラベルごと単語分布の比較を行った．

5.1 ニュース記事トピックトレンドの分析結果

BERTopic を介したニュース記事トピックトレンド分析の結果，抽出された 5 つのトピックは 0_ceo_nissan_ghosn_motors，1_honda_self_india_driving，2_boj_falls_nikkei_investors，3_japan_blasts_hikes_abe，4_inflators_bag_takata_recall であった．縦軸は期間内における該当トピックの発生回数，横軸は時間を示している．特にトピック 0 は日産自動車株式会社のカルロス・ゴーン氏に関するトピックであり，2016 年 5 月ごろに話題としてピークを迎えた．また，トピック 4 はタカタ株式会社のエアバッグ・リコール問題に関するトピックであり，2015 年の 5 月ごろに話題としてピークを迎えた． 他いずれのトピックにおいても各トピックの出現頻度は時系列によって変化している．

5.2 株価変動率に基づくラベル付け結果

株価変動率に基づく 2014 年 4 月から 2017 年 3 月までのニュース記事 2,326 件へのラベル付けの結果，2014 年 4 月から 2017 年 3 月にロイターニュースが配信したポジティブ・ネガティブなニュース記事は大きな偏りはなく，2014

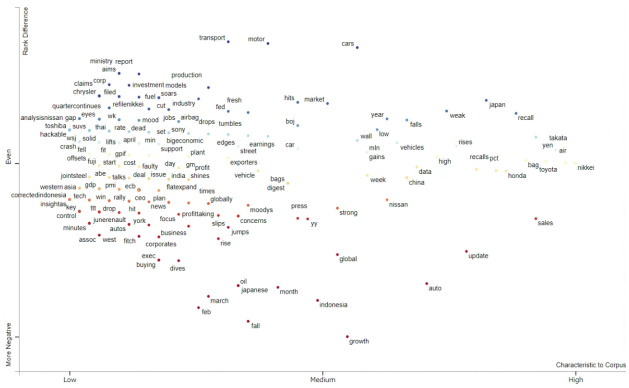

図 2　2014 年 4 月から 2015 年 3 月における
ニュース記事の単語分布

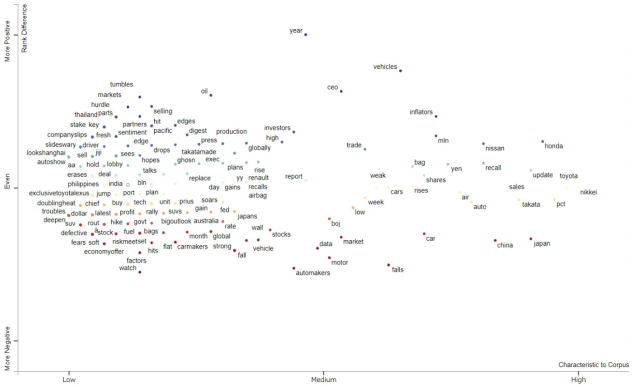

図 3　2015 年 4 月から 2016 年 3 月における
ニュース記事の単語分布

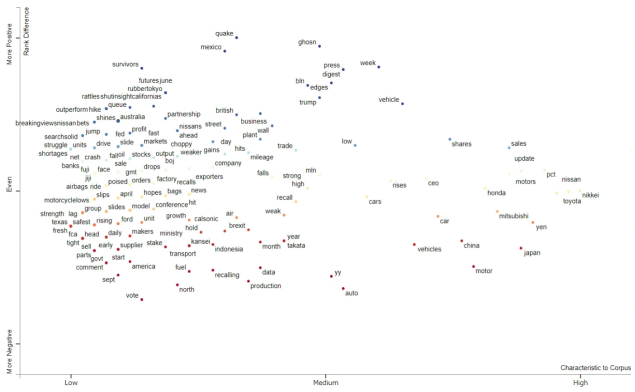

図 4　2016 年 4 月から 2017 年 3 月における
ニュース記事の単語分布

年 4 月から 2015 年 3 月における Positive なニュース記事は 450 件，Negative なニュース記事は 430 件，2015 年 4 月から 2016 年 3 月における Positive なニュース記事は 358 件，Negative なニュース記事は 377 件，2016 年 4 月から 2017 年 3 月における Positive なニュース記事は 365 件，Negative なニュース記事は 346 件であった．

5.3 単語分布の比較結果

Scattertext を介して 2 次元上に各年のニュース記事の単語分布を出力した．縦軸には密なランキング差異（Dense ranking difference）を示している．ここでの「密なランキング」とはランキング化した際にタイのランクをスキップしない連続するランキングのことである．正規化されたカテゴリ固有の密なランク差を示している．横軸はデータセット固有の表現であるかの度合いを示している．比較の結果，単語分布は各年で変化している．特徴的な単語として 2014 年 4 月から 2015 年 3 月までのデータを用いた算出結果において，Negative な単語として growth が算出されているが，該当年度においては自動車業界において” FITCH: ROBUST CHINA GROWTH FOR JAPANESE AUTOS, BUT RISKS REMAIN”のような成長に伴うリスク示唆もしくは他諸国の成長について報じられたニュース記事が多くみられた為であると考えられる．

6. 結論

本研究では，2014 年 4 月から 2017 年 3 月までの日本の自動車業界大手 3 社である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のニュース記事データとティックデータを用いてニュース記事の自動分類とニュース記事における文章トピックの時系列変化の分析を行なった．文章トピックの時系列変化の分析を行なった結果，ニュース記事の文章トピックには時系列の変化が見られた．また，単語分布の各年比較結果においても同様に変化が確認された． 本研究において，ニュース記事がテーマとして扱うトピックは時系列で移り変わっており，分析対象年度によっては特徴的なトピックが存在することが確認できた．

今後の研究として，半期・四半期等のより細かい粒度での検証をし，株価変動評価モデル[10]にトピックの時系列変化を考慮したニュース記事生成システムを採用し，新たな株価変動評価モデルを構築することを予定している．

参考文献

[1] Bird J.J, Pritchard M, Fratini A, Ekárt A and Faria D.R.: “Synthetic Biological Signals Machine-Generated by GPT-2 Improve the Classification of EEG and EMG Through Data Augmentation”, IEEE Robotics and Automation Letters, 6, 2, pp. 3498-3504, 2021.
[2] 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.
[3] 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.
[4] Fama E.F.: ”Efficient Capital Markets: A Review of Theory and Empirical Work”， Journal of Finance, 25，pp.383-417，1970.
[5] Grootendorst M.: “BERTopic: Neural topic modeling with a class-based TF-IDF procedure”, arXiv:2203.05794, 2022.
[6] Kessler J.: “Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ”, arXiv:1703.00565, 2017.
[7] Shleifer A.: “Inefficient Markets: An Introduction to Behavioral Finance”, New York: Oxford University Press，2000.
[8] Nishi Y, Suge A, Takahashi H.: “Construction of News Article Evaluation System Using Language Generation Model”, Agents and Multi-Agent Systems: Technologies and Applications 2020, 186, pp. 313-320, 2020.
[9] Nishi Y, Suge A, Takahashi H.: “News Articles Evaluation Analysis in Automotive Industry Using GPT-2 and Co-occurrence Network”. New Frontiers in Artificial Intelligence, 12331, pp. 103-114, 2020.
[10] Nishi Y, Suge A, Takahashi H.: “Construction of a news article evaluation model utilizing high-frequency data and a large-scale language generation model”, SN Bus Econ, 1, 104, 2021.