

高頻度データ分析を通じたニュース情報が投資家行動に与える影響の解明

研究代表者 高 橋 大 志 慶應義塾大学 大学院経営管理研究科 教授

1 はじめに

近年、機械学習などの急速な進展および分析可能なデータの多様化・大規模化などを背景として、株式市場などを対象とした市場分析においても計算機を活用した分析手法が関心を集めている^{16 19 20 3 6}。ファイナンス分野においては、有価証券報告書やニュースなどテキストデータは重要な役割を果たしており、テキストデータから有用な情報を獲得する試みは高まっている^{1 2 12 13 18}。自然言語処理技術は、近年急速な進展を遂げている分野の一つであり、文章内の依存関係をアテンション機構により捉えるトランスフォーマー(Transformer)が提案され、トランスフォーマーを基盤として数多くのモデル(例えば、BERT、GPT など)が提案されている。本研究でもこれらの背景を受け、ニュース記事を対象とした分析を試みている。ニュース記事などのテキストを用いた分析は、情報が、どのようなプロセスで資産価格に反映されていくのかという資産価格理論における中心的な問いと密接に関連するものであり、その重要性は大きく、とりわけ近年、社会科学の分野においても急速に学術的関心は高まっている^{4 5 10}。本研究でもこれらの背景を受け、ニュース記事など文章データを分析対象とし資産価格に関する情報が投資家行動を通じ与える影響を明らかにすることを試みる^{8 9 14 17}。即時性の高いニュース記事等を含めた分析は、資産価格に関する情報が、どのようなプロセスで価格に反映されていくのかという資産価格理論における中心的な問いと密接に関連するものであり、その重要性は大きく、とりわけ近年、社会科学の分野においても急速に学術的関心は高まっている^{11 22 23 24}。

データ

本分析では、分析対象としてトムソンロイターが提供するニュースデータ(以下ロイターニュース)、有価証券報告書、特許文書を採用した。ロイターニュースは、メディアが発信する情報に分類されるものであるが、そのようなメディアが発信する情報は、各団体が発信する一次情報に比べると、各メディアの記者やアナリストによる情報の取捨選択が行われており、社会や市場に対して相対的に重要な情報が含まれていると考えられる。本データは、金融市場における価格付けにおいて主要な役割を果たしている機関投資家の基にも配信されている膨大な量のデータであり、即時性の高いニュースデータであることから、資産価格にも高い影響力を有することが期待されるデータである。更に、本分析では、ニュースデータに加え、企業属性に関する有効な情報獲得のため、企業活動に関する詳細な記述を含む有価証券報告書、特許文書などを用いた分析も実施している。

分析においては、ニュースデータに関しては2003年から2019年の期間の配信されたニュースを対象として分析を行った。ロイターニュースには、ニュースの属性に応じて、いくつかの分類が与えられているが、本分析では、即時性の高いニュース記事に対する資産価格の反応を分析するため、再送記事と訂正記事については分析から除外した。また、ニュース記事のテキスト情報に注目したため、決算情報のみのニュース記事についても、除外して分析を行うこととした。

3 分析結果

3-1 文章データ分析システム

図1は、本研究にて構築した文章データ分析システムの概略を示したものである。同図は大きく3つの部分に分けられる。図中の一番上は、分析対象となる「企業活動に関する文章データ」(テキストデータ)である。本研究にて取り扱うテキストデータとしては、ニュース、有価証券報告書、特許文書など複数のタイプの文章がある。「文章データのベクトル表現」の部分においては、与えられた文章データから自然言語処理技術を用いてベクトル化が行われる。ベクトル化にあたっては、自然言語処理において数多くの手法が提案されているが、本研究では、それら自然言語処理技術の研究成果を取り込み文章データのベクトル

ル表現を獲得する^{21 24 15}。図中の「文章評価」の部分、文章データをいくつか基準のもと評価測定するプロセスである。当プロセスにおいては、（１）深層学習を利用した教師無学習による分類評価、（２）ニュースに対する資産価格の変化を用いた評価など、いくつかの視点から文章評価を行った。図２は、資産価格の変化を基に文章の評価をする方法の概略を示したものである。本研究では、ファイナンス分野において広く用いられているイベントスタディと呼ばれる手法を、本研究の評価手法の一つに取り入れる⁷。図中の右上のアブノーマルリターンの部分にある図は、特定のトピックのニュースの発信に対する市場の反応を図示したものとなっている。このように、発信時間記録のあるニュースを用いることで市場の反応を測定することが可能である。

企業活動に関する文章データ（例：ニュース等）

News.1 : ESG is important theme for management.
News.2 : Company A has announced the acquisition of Company B.
News.3 : The long-term interest rate has increased by 0.1%.
News.4 : Company C has announced a successful research and development of a new drug.

文章データのベクトル表現

	capital companies	consumption	consumption	corporate	due	export	fall	falling for	have	import	increased	increase	lead	management
News.1	0.00	***	***	***	***	***	***	***	***	***	***	***	***	0.00
News.2	***	***	***	***	***	***	***	***	***	***	***	***	***	***
News.3	***	***	***	***	***	***	***	***	***	***	***	***	***	***
News.4	0.00	***	***	***	***	***	***	***	***	***	***	***	***	0.00

文章評価



図 1 文書データ分析システムの概略（基本モデル）

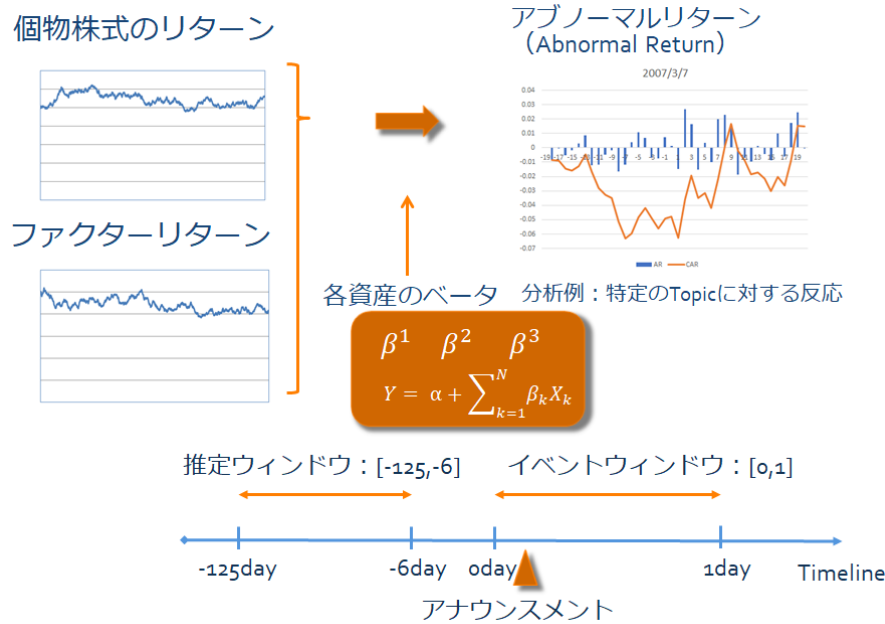


図 2 資産価格変化を用いた評価方法（イベントスタディ）

図3は、複数文書を対象とした分析の概略を示したものである（Okazaki/Takahashi [2023, 2024]）。複数文章を対象とした分析では、図1に示した文書データ分析システムを各文章にあてはめ比較分析を進める。本分析では、金融市場においてタイムリーに発信され即時性の高い文章データ（例：ニュースデータ）と年に一度程度と相対的に低い頻度で発信される文書データ（例：有価証券報告書）を対象とし両者の関連性について比較を行う。ニュース記事は、資産価格にインパクトを与える可能性の高い内容を含むものが多く、そのような背景から文章が発信された日時がサブミクロンのオーダーまで記録されている記事である。そのため、高頻度データと組み合わせた分析も可能なデータとなっている。一方、有価証券報告書は、金融分野において最も広く知られている文章であり、毎年同じ時期に定期的に発信される文書データである。本分析では、自然言語処理技術の利活用を通じ、両文章データからの有効な知識獲得、両者の関連性について分析を行い一部成果については国際学会などにおいて発表を行っている（Okazaki/Takahashi [2023, 2024]）。

本研究においては自然言語処理の最先端の成果と取り込んだ分析を行うが、近年の深層学習を用いた分析の課題としてモデルがブラックボックスとなる点などは広く指摘されている。そのような課題意識を背景とし、モデルの説明可能性に関する議論も広く関心を集めている。本分析では、説明可能性に焦点をあてた分析も実施しており、研究成果の一部は国際学会にて発表を行っている（Nishi/Takahashi [2023]）。

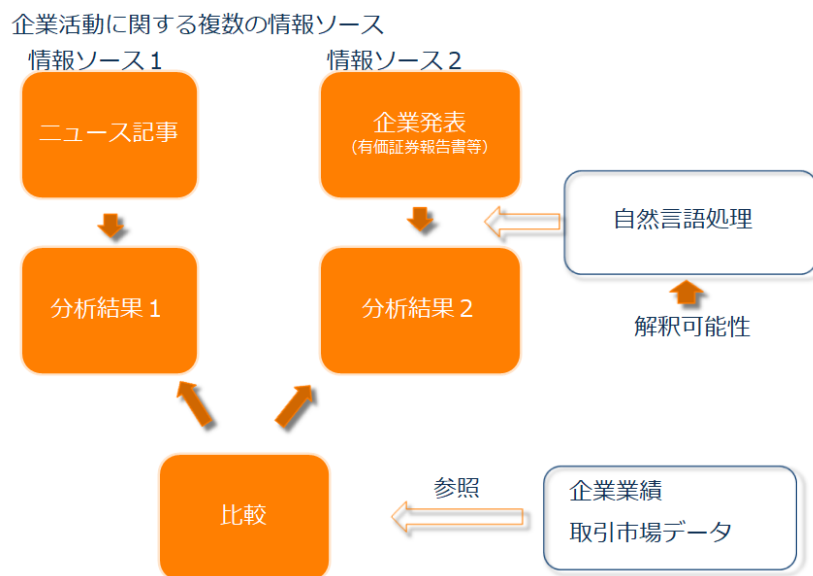


図3 複数文書を対象としたデータ分析システムとその周辺

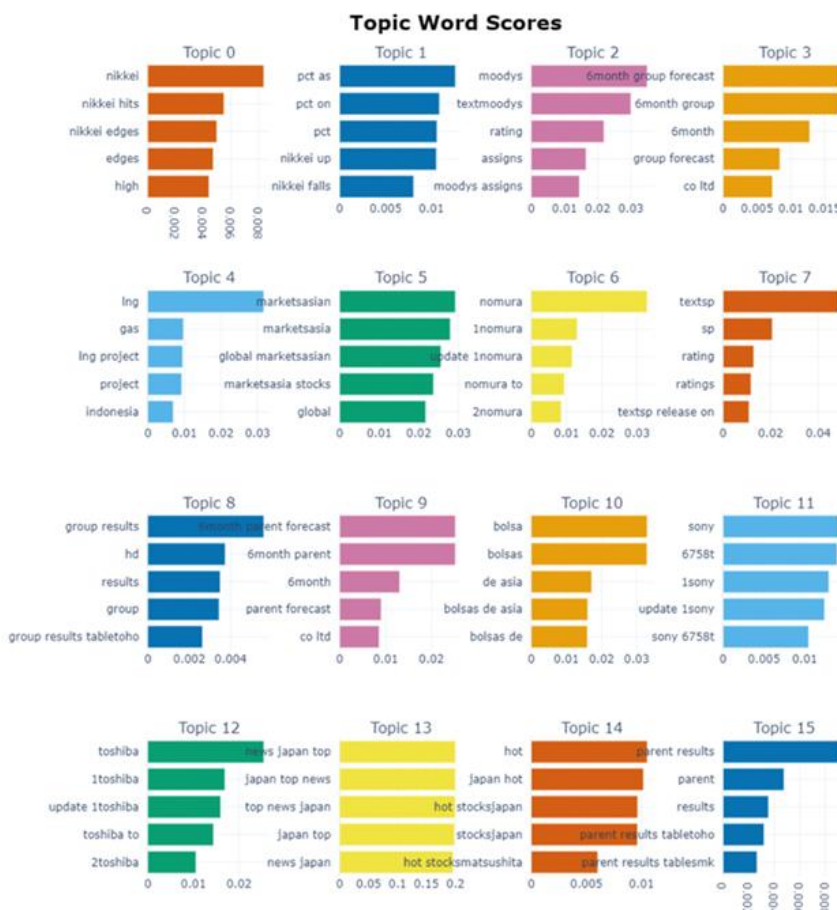
3-2 分析結果

分析対象となる文章データは大規模でありそれらの適切な分類は分析における重要な要素の一つである。本研究では、はじめにニュースの分類を行った。分析においては、ロイターニュースの分析を通じ、企業動向に関する有効な情報の獲得を試みた。ロイターニュースにおいては、各国の企業、経済に関する数多くのニュースが、日本語、英語を含め様々な言語を通じ配信されているが、本分析では、最も記事数の多い英語記事を対象として分析を行った。英語のテキストデータは、最先端の自然言語処理技術を通じ効率的に分析が可能である。

図4は、1996年から2019年のロイターニュースを対象に文章分類を行った結果である。本分析では、トピックモデルの一つであるBERTopicを使用して分析を行った。トピックモデルとは、文書中の出現単語の出現確率やその単語の意味によって文書からトピックを推定するモデルであるが、BERTopicは、BERT(Bidirectional Encoder Representations from Transformers)を用いることで、文脈を意識したトピック分類を行うことが可能なモデルである。図より、当モデルを通じ、ニュースをいくつかのタイプに分類することが可能となっていることを確認できる。

一般に、研究開発をおこなっている企業などをはじめ一つの企業内に複数の事業を有している場合が多く存在するが、Okazaki/Takahashi [2023, 2024]は、研究開発型企業に焦点をあてた分析を実施し、分析の

結果、1) BERTopic による分析を通じ企業が注力している事業に関する情報獲得可能性のあること、2) 市場全体のニュースと個別企業のニュースの比較を通じ、個別企業の特徴に関する情報の獲得可能性のあることを見出している。当研究では、ニュース分析を通じた、タイムリーな企業動向や製品戦略に関する情報獲得可能性のあることを示すものであり、学術的および実務的にも重要性の高い研究である。よりパフォーマンスのよい自然言語処理技術の成果を取り込んだ詳細なニュースデータ分析や日中価格変動データを用いた分析などは継続して取り組み中である。

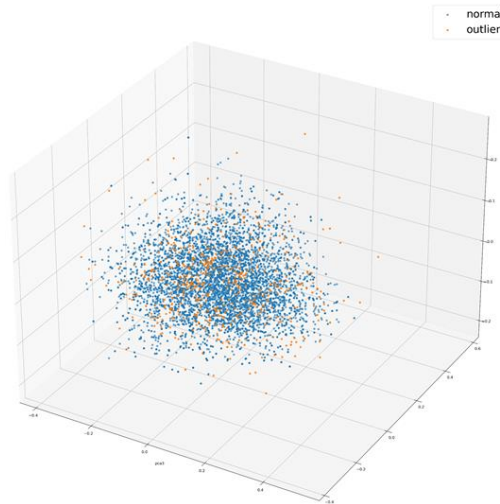


出所: Okazaki/Takahashi[2023]: BERTopic study on the acquisition of information on business activities through news analysis: a year-by-year analysis. International Workshop Artificial Intelligence of and for Business, 2023. より抜粋加工.

図 4 BERTopic による文章分類

自然言語処理技術を通じ、詳細な企業属性を取り込んだ分析も可能となる。本研究では、企業属性の詳細に焦点を当てた分析も取り組んだ。図5は、研究開発型の特定の企業の特許文章をベクトル化し可視化を行った分析結果である。図中に数多くの点が存在するが、各点は、一つの特許を示している。分析対象企業は、保有する特許の数だけ図中の点が存在している。図中の特許をベクトル化することにより特許間の位置関係に関する情報を獲得することが可能である。その結果、特許間の距離や特許のばらつき度合いなどについても定量的に評価が可能である。

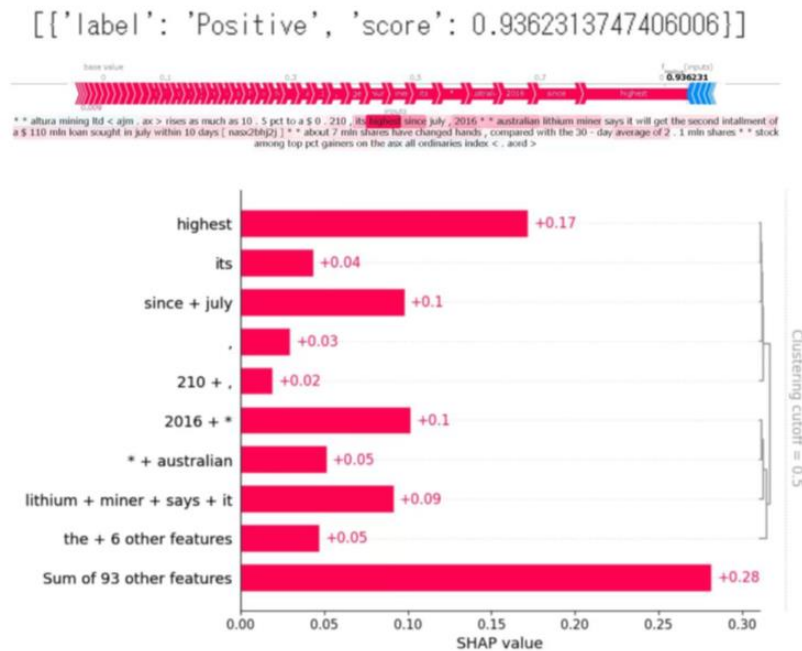
本研究では、特許のばらつき度合いを多角化の程度を示す指標と捉え、企業価値と多角化度合いの関連性について国内上場の製造業に属する企業を対象に分析を行った。分析の結果、多角化ディスカウントと整合的な結果を見出している (Matsumoto/Suge/Takahashi 2023)。また、企業のM&Aに焦点を当てた分析においても企業間の技術的距離と企業が生み出すイノベーションに一定の関係がある可能性を見出している (Yamamoto/Takahashi [2023])。



出所：Yamamoto/Takahashi [2023]: The Relationship Between Technological Distance and Innovation Emergence in M&A Through Patent Data with Outlier Detection Methods, Agents and Multi-Agent Systems: Technologies and Applications 2023, Springer, Singapore, pp. 195–205, 2023.より抜粋加工。

図5 企業の特許文章の可視化の例

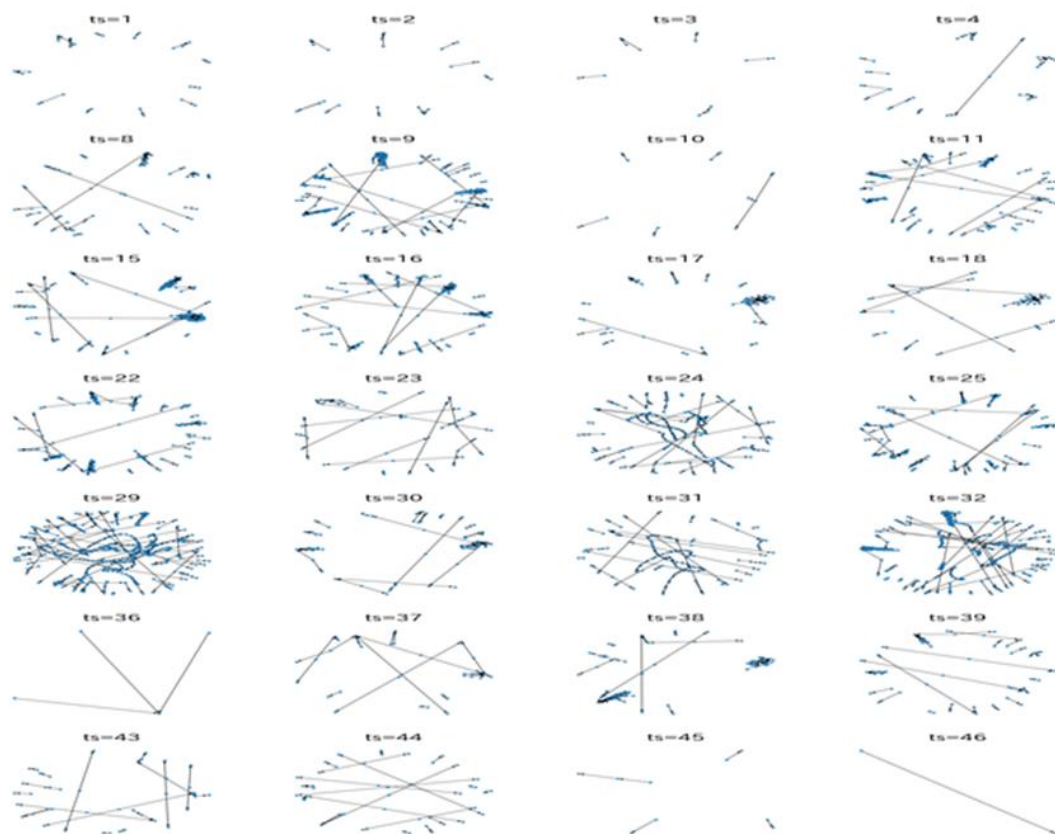
近年の情報技術の進展によりこれまで取り組みが困難であった課題に取り組めるようになってきた一方で、利用するモデルのブラックボックスとなる課題も指摘されている。本研究では、これらを背景としてモデルのブラックボックス化の課題を緩和するための取り組みも行った（図6）。



出所：Nish/Takahashi [2023]: Building a Shapley FinBERTopic System to Interpret Topics and Articles Affecting Stock Prices, The 8th IEEE/ACIS International Conference on Big Data, Cloud Computing, and Data Science Engineering, 2023.より抜粋加工。

図6 モデルの説明可能性向上のためのシステム提案（SHAP 値に基づく指標）

株式市場分析にて用いられる機械学習手法は、異なる市場においても有効に機能する可能性がある。図6は、ビットコイン市場の不正取引分析に機械学習を分析した結果の一つである。同図においては、市場参加者間の取引を短い期間に区切り、時間ごとの違法取引の推移の可視化（抜粋）を行ったものである。本分析では、違法取引検出の難しさなどビットコインの課題についていくつか指摘を行っている（Ehara/Takahashi [2023]）。



出所：Ehara/Takahashi [2023]: Bitcoin Fraudulent Transaction Detection Vulnerability, Agents and Multi-Agent Systems: Technologies and Applications 2023, Springer, Singapore, pp. 183-193, 2023. より抜粋加工。

図6 違法取引の短期間ごとの推移の可視化（抜粋）

4 おわりに

本研究は、機械学習及びテキストデータの利活用を通じた分析により、情報が価格に反映されるプロセスについて分析を行った。分析においては、ニュース、有価証券報告書、特許文書を用いた分析を通じ、情報と資産価格の関連性について分析を行った。本研究は、近年急速に進展してきている情報技術を活用することで、企業が注力している事業に関する情報獲得可能性のあること、企業の多角化と資産価格の関連性など興味深い結果を見出すことができた。また、本研究では、モデルのブラックボックス化の課題を緩和するため取り組みやビットコインにおける違法取引に関する分析などについても興味深い結果を見出すことができた。

情報の果たす役割は、資産価格理論において重要な役割を果たしており、情報技術の利活用は今後ますます高まると期待される。本研究は、情報技術およびファイナンスの両者に亘る学際的でありかつ最先端の研究課題となっている。今後、本研究にて得られた研究を更に拡張させてゆくことを予定している。具体的には、分析対象の拡大、多様な手法を用いた分析などを計画している。

【参考文献】

- [1] 高橋大志: 人工知能・概観 ―機械学習、自然言語処理、計算機シミュレーションおよび 関連トピック―, 証券アナリストジャーナル, 9, 61, 6-15. (2023)
- [2] Antweiler, W., and Frank, M.Z.: Is All That Talk Just Noise? The Information Content of Internet Stock Message Board. *Journal of Finance* 59(3), 1259--1294 (2004)
- [3] Bishop, C. M. : *Pattern Recognition and Machine Learning*. Springer (2006)
- [4] Black, F. and Scholes, M. : The Pricing of Options and Corporate Liabilities. *Journal of Political Economy* 81(3), 637--654 (1973)
- [5] Brealey, R.A., Myers, S.C. and Allen, F.: *Principles of Corporate Finance*, McGraw-Hill (2010)
- [6] Campbell, J.Y., Grossman, S.J., and Wang, J.: Trading Volume and Serial Correlation in Stock Returns, *Quarterly Journal of Economics* 108(4), 905--939 (1993)
- [7] Campbell, J. Y., Lo, A. W. and MacKinlay, A. C. : *The Econometrics of Financial Markets*, Princeton University Press (1997)
- [8] Dougal, C., Engelberg, J., Garcia, D., and Parsons, C. A.: Journalists and The Stock Market, *Review of Financial Studies* 25(3), 639--679 (2012)
- [9] Engelberg, J., Reed, A. V., and Ringgenber, M. C. : How are Shorts Informed? Short Sellers, News, and Information Processing, *Journal of Financial Economics* 105(2), 260--278 (2012)
- [10] Fama, E. : Efficient Capital Markets: A Review of Theory and Empirical Work, *Journal of Finance* 25(2), 383--417 (1970)
- [11] Fama, E. F. and French, K. R. : Common Risk Factors in The Returns on Stock and Bonds, *Journal of Financial Economics* 33(1), 3--56 (1993)
- [12] Garcia, D. : Sentiment during Recessions. *Journal of Finance* 68(3), 1267--1300 (2013)
- [13] Goshima, K., Takahashi, H., Terano, T.: Estimating financial words' negative-positive from stock prices. *The 21st International Conference Computing in Economics and Finance* (2015)
- [14] Henry, A. : Are Investors Influenced by How Earnings Press Releases Are Written? , *Journal of Business Communication* 45(4), 363--407 (2008)
- [15] Hinton, G.E., Osindero, S., and Teh, Y.: A Fast Learning Algorithm for Deep Belief Nets, *Neural Computation* 18(7), 1527--1554 (2006)
- [16] Ingersoll, J. E. : *Theory of Financial Decision Making*. Rowman & Littlefield (1987)
- [17] Kearney, C. and Liu, S. : Textual Sentiment in Finance : A Survey of Methods and Models, *International Review of Financial Analysis* 33, 171--185 (2014)
- [18] Healy, A.D. and Lo, A.W. : Managing Real-Time Risks and Returns: The Thomson Reuters NewsScope Event Indices. In: Mitra, G. and Mitra L. (eds.) *The Handbook of New Analytics in Finance*. John Wiley & Sons, West Sussex (2011)
- [19] Luenberger, D.G. : *Investment Science*. Oxford University Press, Oxford (2000)
- [20] Sharpe, W.F. : Capital Asset Prices:A Theory of Market Equilibrium under Condition of Risk. *Journal of Finance* 19(3), 425--442 (1964)
- [21] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., and Potts, C.: Recursive Deep Models for Semantic Compositionality Over Sentiment Treebnk. *Conference on Empirical Methods in Natural Language Processing* (2013)
- [22] Tetlock, P.C. : Giving Content to Investor Sentiment:The Role of Media in the Stock Market. *Journal of Finance* 62(3), 1139--1168 (2007)
- [23] Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S. : More Than Words: Quantifying Language to Measure Firms's Fundamentals, *Journal of Finance* 63(3), 1437--1467 (2008)
- [24] Loughran, T. and McDonald, B. : When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance* 66(1), 35--65 (2011)

〈発 表 資 料〉

題 名	掲載誌・学会名等	発表年月
BERTopic study on the acquisition of information on business activities through news analysis: a year-by-year analysis	International Workshop Artificial Intelligence of and for Business (AI-Biz2023)	2023 年 6 月
Bitcoin Fraudulent Transaction Detection Vulnerability	Agents and Multi-Agent Systems: Technologies and Applications 2023, Springer	2023 年 6 月
The Relationship Between Technological Distance and Innovation Emergence in M&A Through Patent Data with Outlier Detection Methods	Agents and Multi-Agent Systems: Technologies and Applications 2023, Springer	2023 年 6 月
Analysis of the Relationship between Technological Diversification and Enterprise Value using Patent Data	Information Technology and Management	2023 年 10 月
Building aShapley FinBERTopic System to Interpret Topics and Articles Affecting Stock Prices	The 8th IEEE/ACIS International Conference on Big Data, Cloud Computing, and Data Science Engineering	2023 年 10 月
Nowcasting Corporate Product Development Activities Through News Article by BERTopic: The Case of the Japanese Chemical Company	JSAI-isAI 2023, Lecture Notes in Computer Science, vol. 14644, Springer	2024 年 (to appear)