# 金融テキストマイニング研究の紹介

応専

和泉 潔　東京大学　　松井 藤五郎　中部大学

## 金融市場とスカート丈そしてテキストデータ

1926 年に経済学者の George Taylor 氏がヘムライン指数と呼ばれる経済理論を提唱した[1]．ヘムライン（hemline）とはスカート丈のことである．この理論は，スカート丈が短くなると株式市場が上げ相場になり，長くなると下げ相場になると主張している．金融市場は世の中の経済活動の活発さを反映しているはずである．だからもし，みんなが持っている平均的な景況感を早く正確に知ることができたら，株価が予測できるはずだ．これが，この理論の根底にある暗黙の仮定である．つまりヘムライン指数では，スカート丈が平均的な景況感の優れた指標になると主張している．

現代は，みんなの景況感を知るために，スカート丈よりもずっと良い情報がある．Web 上の大量のテキスト情報である．専門家でないごく普通の人たちが，経済と直接は関係ないような事柄について書いたものから，金融の専門家が発信する市場にかかわるさまざまなニュースや経済レポートまで，大量のテキスト情報が常に溢れている．

そこで近年，機械学習を用いたテキストマイニング手法によって，テキスト情報と市場変動の関係性を発見し市場分析に応用する研究が増えてきた．経済指標やマーケットのテクニカル指標等の数値情報には指標化されていないような情報を，テキスト情報から素早く自動的に抽出することが期待されている．本稿で具体的に研究事例を紹介する．

## 金融テキストマイニング研究の概観

金融テキストマイニング研究は，入力するテキストの性質・分析手法・予測対象となる市場の種類によって分類できる（表 -1）．

まず，分析対象となるテキストは，匿名の書き手に口語体で書かれたものから，特定の機関が文体で書いたものまで何種類かある．これらのテキストは，

| | | | | | | |
|---|---|---|---|---|---|---|
| 入力テキスト | テキスト | ツイッター[2] | 掲示板[4] | ニュース[6] | ニュース[7] | リポート[8), 9)] |
| | 量 | 1GB 以上 / 日 | 数百 KB/ 日 | 数百 KB/ 日 | 数百 KB/ 日 | 数十 KB/ 月 |
| | 書き手 | 1 億人以上<br>不特定多数 | 数百人<br>投資家 | 数百人<br>記者 | 数百人<br>記者 | 数十人<br>専門家 |
| | 内容 | 多様 | 少し限定 | 少し限定 | 少し限定 | 経済専門 |
| 分析手法 | 分析するテキストの期間 | 直近 24 時間 | 直近 24 時間 | 最新記事 | 直近 10 日間 | 直近 1 カ月間 |
| | 特徴の定義 | 手動 | 自動 | 手動 | 自動 | 自動 |
| | 処理 | Bag-of-words | 極性分析 | Bag-of-words | 構文解析 | Bag-of-words |
| 予測対象となる市場 | 価格の更新頻度 | 日次 | 日次 | 分次 | 日次 | 月次 |
| | 予測対象 | 市場平均 | 個別銘柄 | 個別銘柄 | 個別銘柄 | 市場平均・国債 |
| | 予測時間 | 1 日先 | 1 日先 | 20 分先 | 1 日～2 カ月先 | 約 2 週間先 |

表 -1　本稿で紹介する主な金融テキストマイニング研究の概要

量や内容そして書き手の多様性／専門性の軸によって整理できる．たとえば，ツイッターやブログなどは多様な内容と書き手を持つ膨大な量のテキスト情報であるが，書かれている内容は日常的な事柄も含む非常に雑多で統一性のないものである．オンラインのニュース記事や株式に関する掲示板などは，もう少し専門的なテキスト情報である．金融機関の発行する経済リポートは，少量だが一番専門的なテキスト情報である．テキストの様式や言葉使いも，ある程度の統一性を持っている．

金融テキストマイニングの分析手法は，テキストを単語の集合と見なして，単語の出現頻度情報を利用する bag-of-words が多く用いられている．テキストが大量にある場合は，比較的最新の短期間で得られたテキストを用いて，あらかじめ手動で列挙した単語リストを用いて特徴量を計算することが多い．テキストの量が少ない場合は，過去のより長期間のテキストから，自動的に抽出した単語リストを用いて特徴量を計算する傾向がある．

また，テキストの量が多いときには，直近のテキストでも大量のデータがあるので，その中から特定の銘柄に対する比較的短期間先の市場への影響を予測しようとする研究が多い．量が少ないが専門的なテキストの場合は，過去の長い期間でのテキストの特徴の時間変化を調べて，より長期間で広範囲な市場への影響を見ることが多い．

次章以降で個別の研究事例を紹介する．

## ツイッターに現れる意見と株価平均

Bollen らは，2008 年 2 月 28 日から 11 月 28 日の 9,853,498 個のツイッターデータを分析し，米国のダウ・ジョーンズ工業株価平均との関係性を調べた[2]．ユーザ数は約 2.7 百万人にもなり，1 日平均で 3.2 万個のツイートが投稿された．これだけ膨大なテキスト情報があれば，経済に対する世の中の平均的な見方のトレンドが抽出できるのではないかと考えたのである．

このテキスト情報のうち，彼らは書き手が自分の心的状態を明言していると思われるツイートだけを分析対象とした．そのために，“i feel” や “i am feeling”，“i'm feeling”，“i don't feel”，“I'm”，“I am”，“makes me” を含むツイートを抽出した．次に，各日の抽出されたツイート集合から，どのような心理状態に関連する表現が多いかを指標化した．心理学で使われる気分プロフィール検査（POMS）をベースとした，Google-Profile of Mood States（GPOMS）指数を新たに提唱している．もとになった POMS は，被験者に対して現在の自分の心的状態を，「友好的な」「不機嫌な」「活発な」「限界ギリギリの」「パニック状態の」等の 72 種類の表現への 7 段階程度の当てはまりを聞く質問紙調査を行い，この回答データから被験者の心的状態を表す，平穏・警戒・確信・活気・善意・幸福の 6 次元の尺度を計算する心理検査法である．GPOMS は，Google の 4,5-gram 共起語（25 億語）から，POMS の 72 表現と共起しやすい 964 語を抽出し，これらの単語の出現頻度も用いて，各日のツイートから先ほどの 6 次元の尺度のスコアを計算する．

テキスト情報を取得した，2008 年 2 月 28 日から 11 月 28 日について，6 次元の GPOMS 指数とダウ平均株価指数を用いて，Granger 因果性検定を行った．その結果，「平穏」の尺度が 2 ～ 5 日後の平均株価との因果性があった．さらに，1 日前から 3 日前までの「平穏」のスコアと平均株価を入力として，翌日の平均株価を予測するモデルを，Self-organizing Fuzzy Neural Network（SOFNN）手法を用いて構築した． 訓練用に用いたツイッターデータの期間は 2008 年 2 月 28 日から 11 月 28 日であり，テストに用いた期間は 2008 年 12 月 1 日から 19 日である．その結果，翌日の平均株価の騰落の方向性を，86.7% の精度で予測することができた．しかし，テキスト情報を用いずに，過去 3 日間の平均株価だけから予測した場合でも，73.3% の予測精度があった．

ツイッター情報を用いたほかの研究として，グラフ構造に着目した研究もある[3]．分析対象の企業名をハッシュタグなどに用いているツイートを抽出する．次に，ツイート・ユーザ・ハッシュタグ・

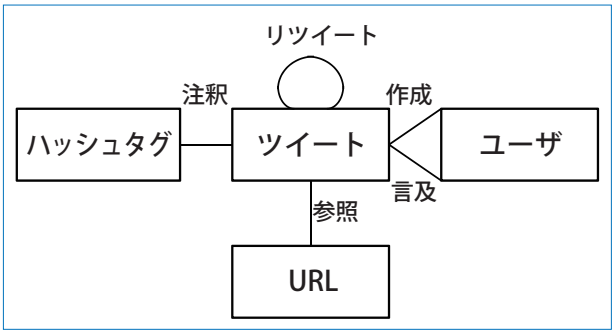


図 -1　ツイッター情報から抽出するグラフ構造の枠組み
文献 3）図 -2 より一部改変.

URL をノードとして，リツイートや引用，作成などをリンクとする図 -1 のようなグラフを日次で構築する．グラフ構造を表す複数の指標と翌日の対象銘柄のリターン（価格の変化率）や取引高との相関を分析した．その結果，グラフの連結要素の数が取引量と正の相関が見られたが，リターンとは相関が見られなかった．つまり，この手法では，ツイッターのネットワーク構造からある銘柄が話題になって取り引きされやすいかどうかは分かるが，それが株価上昇または下落のどちらの内容で話題になっているかまでは解析することは難しかったのである．

## 掲示板の投稿と株価変動の分析

一定期間に蓄積された一般の市場参加者が書いたテキスト情報を分析する研究もある [4]．個別銘柄をテーマとするインターネット上の掲示板に，その銘柄の株価に興味があるユーザから投稿された記事を取り扱う．掲示板の記事から，市場に対する集合的な意見を抽出しようとする試みである．

ここでの分析における入力は，主に次の 3 種類である．(a) 投稿数：株式の銘柄ごとの掲示板における数時間から数日の投稿数，(b) 強気比率：一定期間内の強気の内容を持つ投稿数と弱気内容の投稿数の差．強気と弱気は投稿者自身のタグ付けや好悪表現の頻度等から判断される，(c) 合意インデックス：一定期間の強気投稿数または弱気投稿数のどちらか一方への偏り度合い．主な予測対象は，各掲示板が取り扱う銘柄に関して，掲示板のテキストが収集された翌日の，株価リターン・出来高・ボラティリティ（価格変動の標準偏差）である．これらの市場データとさきほどの 3 種類の特徴量との相関関係を分析した．

その結果，出来高とボラティリティに関しては，どの入力変数ともある程度有意な相関関係が見られた．つまり，投資家の関心が高く投稿数が多くなると，その銘柄の取り引きが活発になり，出来高やボラティリティが大きくなることを示した．強気または弱気どちらか一方に意見が偏った合意インデックスが高い状態は出来高とボラティリティが増加する傾向があった．しかし，株価リターンに関しては，有意な相関が得られなかった．また強気比率のみに対して関係性が見られることがあった．以上の結果より，掲示板のテキスト情報では，どの銘柄が活発に取り引きされているか（されそうか）という判断には有効であるが，その方向性の抽出までは今のところ難しいという状況である．この結果は，先ほどのツイッターのグラフ構造による分析と同様に，実際の投資行動の方向は取引戦略的な要素が関連するので，掲示板やツイッターのテキストに含まれる意見とは直結していないことが原因かもしれない．

## オンラインニュースと短期市場変動

テキストマイニングを用いた市場分析研究で一番多いのが，直近のニュース記事テキストの特徴から，数時間程度の短期的な市場変動の方向性を予測するものである．ニュース記事を用いた先行研究での学習の多くは，最新ニュースまたは今から数時間以内に配信されたニュースのテキストを入力として，対象となる金融価格の今から数時間後のトレンド（上昇，下降，横ばい）またはボラティリティを予測対象として行われる．まず，ニューステキストから，重要そうな単語やカテゴリまたは単語の組合せの頻度（または TF-IDF 値）を計算し，テキストの特徴ベクトルとする．過去のニュース記事の特徴ベクトルとその記事が配信された翌日の市場変動データを用いて，機械学習により両者の関係性を学習する．よく使われている学習手法は，ナイーブベイス，サポートベクタマシン（SVM），分類子システ

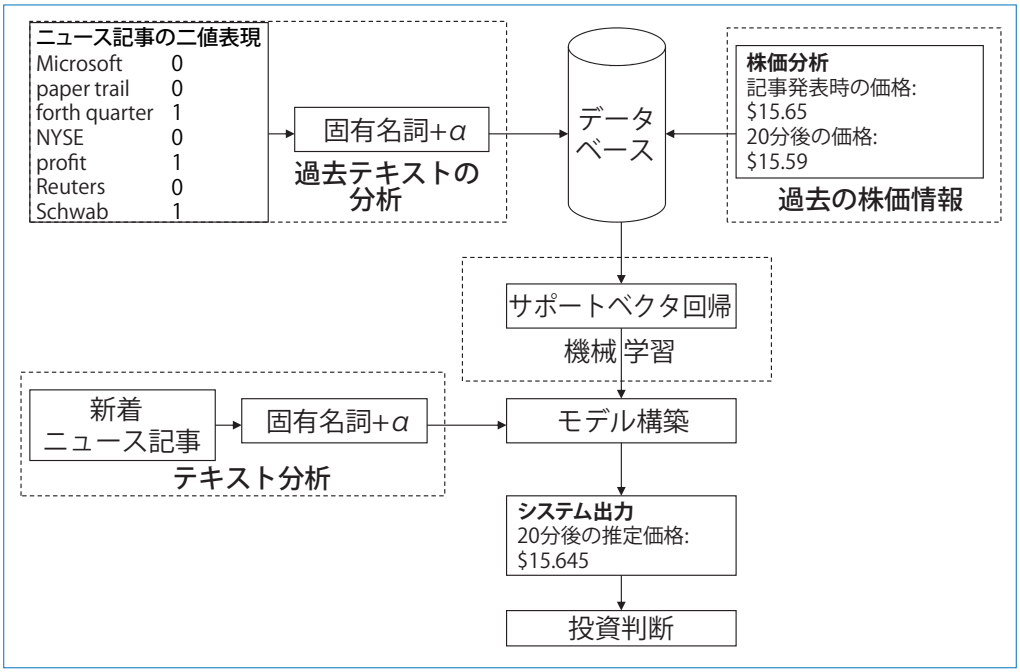


図 -2　AZFinText の概要．文献 6）図 -3 より一部改変．

ム（classifier system）である．過去データを用いた学習により獲得したルールに，最新のニューステキストを入力して，実際に数時間後の市場変動を予測する．訓練データと異なるデータで予測テストを行うと，先行研究では大体 40 ～ 50% の予測精度である [5]．上昇，下降，横ばいの 3 種類の状態への予測なので，ランダムに予測すれば精度は 33% である．ランダム予測よりは有意に精度は高いが，まだ精度向上が必要である．

最新の研究ではテキストマイニングによる予測をもとに実際の運用に活用しようと試みている．Schumaker らは，Yahoo! Finance の記事から米国の個別銘柄の 20 分後の株価動向を予測した [6]．2005 年 10 月 26 日から 11 月 28 日の 5 週間のデータを用いて，Yahoo! Finance から集めた 9,211 記事から取引時間（10:30am ～ 3:40pm）のニュースに限定した 2,809 記事を，図 -2 に示す AZFinText と呼ばれるシステムで分析した．各記事において会社名や要人などの固有名詞とあらかじめ決めておいた用語の出現を見る．その記事が配信されてから 20 分後の S&P500 の構成をする個別銘柄の株価の変化との関係を，サポートベクタ回帰を用いてモデル化する．このモデルを用いて，新たに配信されたニュース記事の単語出現パターンから，20 分後の特定の個別銘柄の価格変化を予測する．運用テストでは，20 分後に 1% 以上の株価変動が起きると予想された銘柄を売買した．同じ期間で，S&P500 の構成銘柄で運用しているクォンツ・ファンド（定量分析を基に運用を行うファンド）と比較すると，どのファンドよりもテキストマイニングの運用成績が良かった．

ほかにも特に有望だと思われるのは，数値データの時系列解析とテキストマイニングを組み合わせる手法である [7]．証券アナリストによる企業の格下げ変更の発表が，その企業の株価に与える影響を分析した．このときに，格下げ発表前のボラティリティに，オンラインニュースのテキストから抽出したポジティブ／ネガティブの市場心理（センチメント）を表す指数を組み合わせて分析することによって，より正確に発表後の株価下落を推定でき，安定した運用を可能にした．数値データだけで運用した場合に比べて，単位リスクあたりの収益率であるシャープ比が約 1.5 倍に増加した．

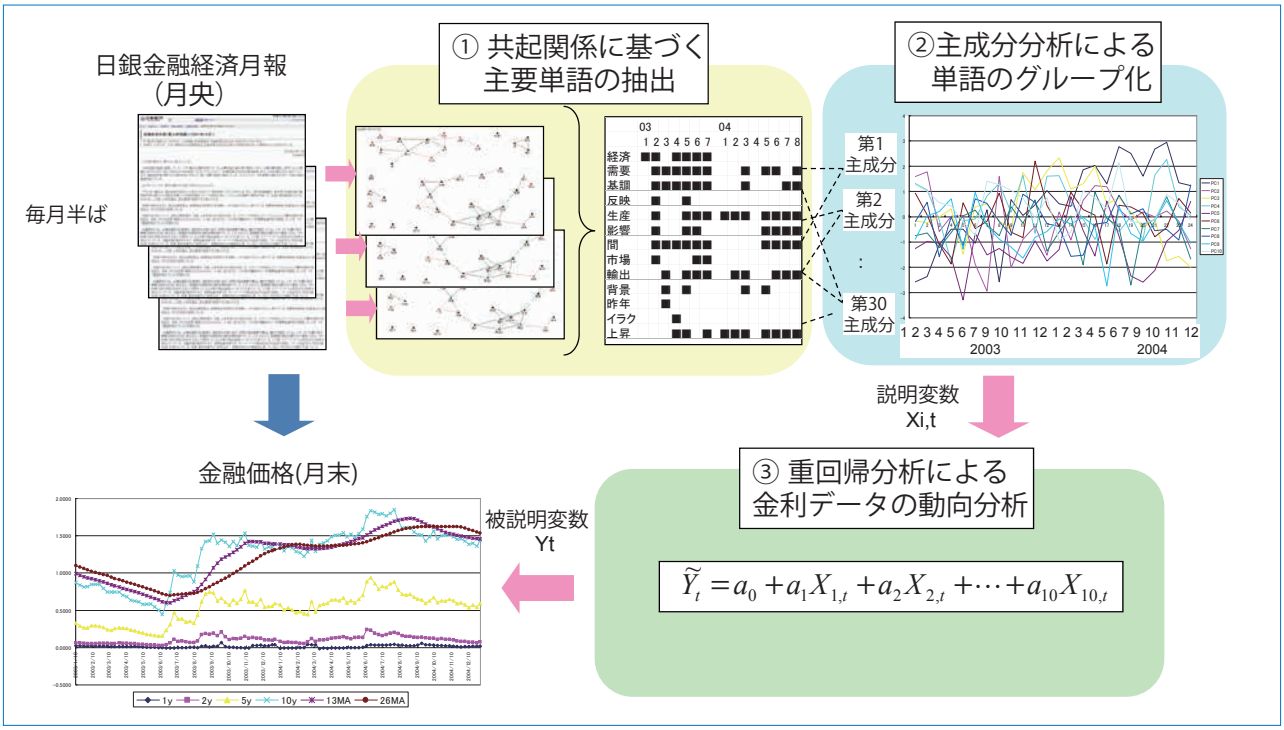


図 -3　経済リポート分析手法の概要　月央に発表されるテキスト情報から①から③の分析ステップを経て，月末の金融価格を推定する．

## 経済リポートと長期市場変動

定期的に発行され形式も定まった経済リポートから，テキストの特徴の時間変化を抽出し，月次以上の長期的な価格時系列データの変動との関係性を発見する研究事例を紹介する[8), 9)]．

入力は，金融経済月報と呼ばれる経済リポートで，日本銀行が日本全体の金融・経済情勢を分析した資料であり，毎月半ばに A4 で 15 ～ 20 ページの分量で公開されている．解説内容の順番や段落構成等がほぼ定式化されていて，月ごとのテキスト内容の変化が比較しやすい．

分析の枠組みを図 -3 に示す．最初に，1998 ～ 2007 年の過去 10 年間のテキスト情報での単語の共起頻度をもとに主要な単語を抽出し，さらに出現頻度の時間変化パターンの主成分分析により人間にも理解しやすい 30 個の特徴量を抽出した．次に，これらの特徴量の時系列データを用いて，過去 10 年間の国債市場の価格データに関して回帰分析を行った．得られた回帰式に 2008 年の各月のテキストデータを入力し，各市場の外挿予測を行った．さすがに 10 月の 3,000 円近くの歴史的な暴落は推定できなかったが，4 月の 1,500 円の高騰や 9 月の 2,000 円の下落などを推定することができ，変動が激しい時期であったにもかかわらず全体的に市場の方向感をよく捉えることができた[8)]．

さらに，毎月の逐次的な外挿予測値を用いて国債市場・株式市場・外為市場で運用テストを行った．その結果，日経平均株価・日本国債 5 年利回り＞日本国債 2 年利回り・日本国債 10 年利回り＞円ドルレートの順で，運用成績が良かった．これは，日銀の動向が各市場に対してどれほど影響力を持ち得るのかということを表した結果だと思われる．運用テスト期間での価格変動の正答率を見ると，前月に比べて大きく下降また上昇した月は，提案手法による変動予測の精度が高かった．つまり提案手法は，市場が大きく動くときに，テキスト情報から市場動向の予兆を抽出することができたのである．さらに，日本国債の市場で運用テストを行った結果，数値データを使った計量経済モデルや同じテキストを使用したサポートベクタ回帰と比べて，どの市場でも安定してほぼ最高水準の運用益をあげることができ

た[9]．変動が大きい時期の騰落予測の精度が高い方が運用益を増加できるので，上述の運用テスト結果もこの手法が市場の大きな変動の予兆を抽出できたことを表している．

## まとめ

金融テキストマイニング研究はまだ新しい研究分野である．分析対象も手法も手探りの状態である．こうすればうまくいくという定石はまだない．現状では，どの手法も一長一短がある．ただし，単一の分析対象だけでなく複数種類のテキスト情報に分析範囲を拡大し，特徴量の工夫や背景情報の考慮などの共通する問題を克服できれば，今後この手法で金融市場に関する集合知を獲得できる可能性がある．

**参考文献**

1) Lewin, T. : The Hemline Index, updated, International Herald Tribune (Oct. 19, 2008).
2) Bollen, J., Mao, H. and Zeng, -J. X. : Twitter Mood Predicts the Stock Market, Journal of Computational Science, Vol.2, No.1, pp.1-8 (2011).
3) Ruiz, J. E., Hristidis, V., Castillo, C., Gionis, A. and Jaimes, A. : Correlating Financial Time Series with Micro-Blogging Activity, Proceedings of the fifth ACM International Conference on Web Search and Data Mining, pp.513-522 (2012).
4) 丸山　健，梅原英一，諏訪博彦，太田敏澄：インターネット株式掲示板の投稿内容と株式市場の関係，証券アナリストジャーナル，Vol.46, No.11-12, pp.110-127 (2008).
5) Mittermayer, -A. M., and Knolmayer, F. G. : Text Mining Systems for Market Response to News : A Survey, Technical Report, University of Bern (2006).
6) Schumaker, P. R. and Chen, H. : A Discrete Stock Price Prediction Engine based on Financial News, IEEE Computer, Vol.43, No.1, pp.51–56 (2010).
7) 岡田克彦，中元政一，東高　宏，羽室行信：負け犬は誰だ？　証券アナリストの格下げにより価値を失う企業，第7回ファイナンスにおける人工知能応用研究会資料，SIG-FIN-007-07 (2011).
8) 和泉　潔，後藤　卓，松井藤五郎：テキスト情報による金融市場変動の要因分析，人工知能学会論文誌，Vol.25, No.3, pp.383-387 (2010).
9) 和泉　潔，後藤　卓，松井藤五郎：経済テキスト情報を用いた長期的な市場動向推定，情報処理学会論文誌，Vol.52, No.12, pp.3309-3315 (2011).

(2012年6月1日受付)

**和泉 潔（正会員）** ‖ kiyoshi@ni.mints.ne.jp

1998年東京大学大学院博士課程修了．博士（学術）．同年より2010年まで，電子技術総合研究所（現 産業技術総合研究所）勤務．2010年より現職．金融情報学に関する研究に従事．人工知能学会，電子情報通信学会，電気学会各会員．

**松井 藤五郎（正会員）** ‖ TohgorohMatsui@tohgoroh.jp

2003年名古屋工業大学大学院工学研究科博士課程修了．博士（工学）．2003～2009年東京理科大学理工学部経営工学科助教．2009年とうごろう機械学習研究所設立．2010年より現職．機械学習およびデータ・マイニングに関する研究に従事．人工知能学会，ACM，AAAI各会員．