Title	機械学習手法を用いた金融市場分析 : 深層学習及び分散表現学習によるテキストマイニング
Sub Title	
Author	片倉, 賢治(Katakura, Kenji) 高橋, 大志(Takahashi, Hiroshi)
Publisher	慶應義塾大学大学院経営管理研究科
Publication year	2014
Jtitle	
JaLC DOI	
Abstract	
Notes	修士学位論文. 2014年度経営学 第2933号
Genre	Thesis or Dissertation
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO40003001-00002014-2933

慶應義塾大学大学院経営管理研究科修士課程

学位論文（　**2014** 年度　）

論文題名

機械学習手法を用いた金融市場分析

― 深層学習及び分散表現学習によるテキストマイニング ―

主　査	高橋 大志
副　査	林 高樹
副　査	岡田 正大
副　査	―

学籍番号	81330416	氏　名	片倉 賢治

所属ゼミ	高橋大志 研究会	学籍番号	81330416	氏名	片倉 賢治

（論文題名）

機械学習手法を用いた金融市場分析

― 深層学習及び分散表現学習によるテキストマイニング ―

（内容の要旨）

近年, 金融市場に対する関心が高まっている. 従来から金融市場の分析には様々な手法が提案されており, これまで市場データを用いた数多くの分析が報告されている. 株式価格を対象とした分析においては, 例えば, 合理的な投資家等を想定したモデルである資本資産価格評価モデル（CAPM : Capital Asset Pricing Model）や, Fama-French の 3 ファクターモデルといった手法が広く用いられている. 一方, 必ずしも合理的でない投資家を考慮した行動ファイナンス（Behavioral Finance）に関する議論も関心を集めており, その議論に広がりをみせている.

更に, 情報処理技術の向上によって市場参加者が利用可能な情報は飛躍的に増大すると共に分析の重要性が増している. 昨今, インターネット上の情報はビッグデータと呼ばれる程に, 年々大規模化, 多種多様化しており, 例えば, 構造化データ（数値データ, テキストデータ）のみならず, 非構造化データ（音声, 動画）といったデータの分析を行う手法も関心を集めている. また, 情報処理技術に求められる役割も従来の効率化や迅速化のみならず, 意思決定支援, 顧客の嗜好やニーズの把握が期待される等, 益々重要性が増している.

以上を背景として, 本稿では, 金融市場を取り巻く大規模情報に対して二つの最先端の機械学習手法を用いたアプローチで分析を試みる. 第一に, 深層学習を用いた評判分析手法によって分析するアプローチ. 第二に, 分散表現学習によって新たなファイナンス辞書を作成し分析するアプローチである.

分析対象のテキストデータに関しては, 世界で最も広く知られたニュース提供会社の一つである Thomson Reuters 社により提供されているニュースを採用し, 市場関連データについては, 日経 NEEDS および Thomson Reuters Datastream を用いた. また, 本分析では、市場データよりファクターリターンの算出を行い、それらリターンを用いた分析を行った.

分析の結果, 次に示すいくつかの興味深い結果を得ることができた. はじめに, (1) 分散表現学習を通じて金融ニュース特有の表現極性を有する単語群を抽出することができた. 次いで, (2) 株式市場は、ネガティブな経済ニュースに継続的に反応しやすいとの結果を見出すことができた. 更に, (3) 新辞書によって得られたポジティブスコアと企業規模との間に相対的に強い関連性があることを見出すことができた.

これらの結果を通じ, 今回採用した手法について一定の有用性を示すことができたが, 一方で, 今後の課題も残った. 頑健性の確認をはじめとして, 新たな辞書を抽出する際のいくつかのパラメータ（学習パラメータ, コサイン距離）調整や, 新辞書の極性付与の方法, 深層学習を行う分析環境整備等,については今後の課題である.

目次

1　序論

1.1　研究の背景と目的

近年，金融市場に対する関心が高まっている．世界的な金融危機以降，先進諸国のみならず多くの国々において経済金融政策運営は大きな転換点を迎えており，個々の金融機関の健全性を確保するだけでは，金融システム全体としての安定を必ずしも実現できるわけではないとの見方が強まっている．このような中，例えば，金融安定のためのマクロプルーデンス政策，インフレ目標の設定，物価安定，ゼロ金利政策，量的緩和政策等，先進諸国において様々な協調政策がなされており金融市場への関心は益々増している．

金融市場には，さまざまなステークホルダーが混在しているが，主要なステークホルダーの一つとして企業を挙げることができる(Fig.1-1)．企業は，金融市場に関する多くの意思決定を行っており，その中でも，企業価値評価は重要な要素の一つとなっている．なぜなら，企業価値の最大化は，株式会社（Corporate）として組織された企業経営者の意思決定において拠り所とすべき基準であり**[1]**，投資や資金調達といったコーポレートファイナンス領域における企業の財務意思決定も，企業価値の最大化，株主価値の最大化が主要な目的とされているためである[2]．

Fig. 1-1: 金融市場を取り巻く様々なステークホルダー．

企業価値評価を行うためにはいくつかの要素が必要となるが, とりわけ資本コスト（Cost of Capital）は重要な役割を果たす. この資本コストは, (1) 投資家や株主にとっての期待リターン, (2) 資金調達コスト, (3) 投資判断基準収益率（キャッシュフロー割引率）, (4) 業績評価基準（ハードルレート）, といった意味を含み, 企業価値評価にとって必要不可欠な要素である. 2000年代に入り, 日本では金融ビックバンによる直接金融の規制緩和をはじめとした金融市場の変化し, 日本的経営の代名詞であったメインバンク制が崩壊する等, 企業統治（コーポレートガバナンス）システムが変化すると共に企業価値や株主重視の声が高まり, この資本コストを意識した経営指標であるEVA®[1]（Economic Value Added）やキャッシュフローを導入する企業が相次いで増加している[3]. その意味においても金融市場の分析, そして, 資本コスト推計の重要性は大きい.

更に, 近年の情報処理技術の向上等を背景とし, 市場参加者が利用可能な情報は飛躍的に増大している. 昨今, インターネット上の情報はビッグデータと呼ばれる程に, 年々大規模化, 多種多様化しており, 例えば, 構造化データ（数値データ, テキストデータ）のみならず, 非構造化データ（音声, 動画）といったデータの分析を行う手法も関心を集めている. また, 情報処理技術に求められる役割も従来の効率化や迅速化のみならず, 意思決定支援, 顧客の嗜好やニーズの把握が期待される等, 益々重要性が増している. そして, これらのことが従来からの分析手法に加え, 新たな分析手法に対する要望を一層高めている.

これらを背景として, 本稿では, 市場参加者が利用可能な大規模情報に対して, 昨今注目を集めている新しい分析手法を用いて金融市場の実証分析を行う. また, これによって, 従来研究から一歩進んだ議論及び新たに採用した分析手法の有用性を示すことを目的とする.

[1] EVA は, 米スターン・スチュワート社の登録商標.

1.2　本論文の構成

本稿では，自然言語処理及び機械学習手法を用いた二つのアプローチによって金融市場の分析を試みる．第一に，深層学習を用いた評判分析手法によって分析するアプローチ（分析手法Ⅰ）．第二に，分散表現学習によって新たなファイナンス辞書を作成し分析するアプローチ（分析手法Ⅱ）である（Fig.1-2）．

そこで本稿の構成は，全体に関かわる関連研究を第二章で述べ，分析に用いたデータの共通部分を第三章で説明する．続いて，分析手法Ⅰを第三章で，分析手法Ⅱを第四章で説明する．それぞれの章では，はじめに採用した手法の概略について説明を行った後，分析に用いるデータの個別的な部分及び分析方法を説明する．次いで，分析結果，考察，まとめを示す．最後に，むすびとして本稿の結論および今後の課題を示す．

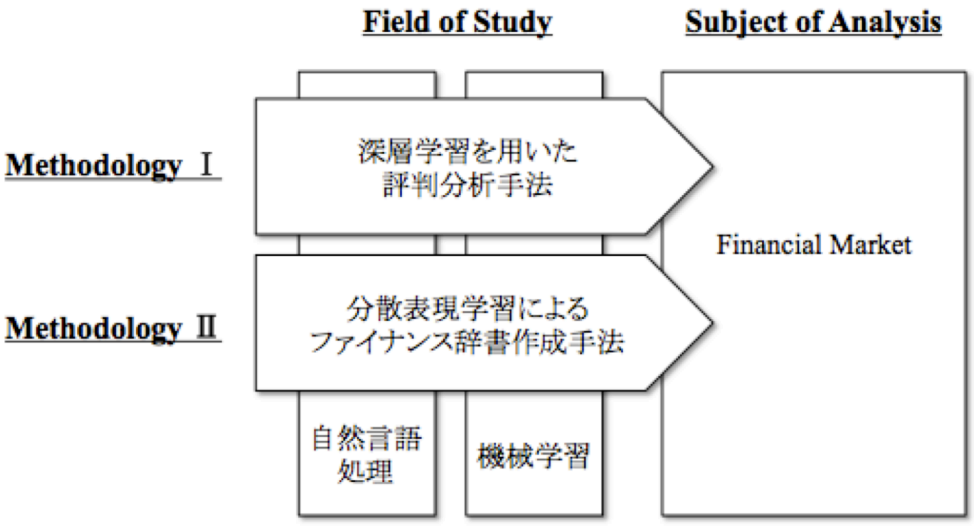

Fig. 1-2: 本稿における金融市場分析のアプローチ．

2 関連研究

本章では, 資本資産価格理論, 機械学習手法の一つである深層学習と呼ばれる多階層ニューラルネットワーク, 自然言語処理にニューラルネットワークを適用した分散表現学習の三つの領域における関連研究について説明する.

2.1 資本資産価格理論

金融市場の分析, 特に, 資本コスト推計は一般にマーケットデータを用いた推計が行われており, 資本資産価格評価モデル（CAPM：Capital Asset Pricing Model）, Fama-Frenchの3ファクターモデルといったモデルが広く用いられている. CAPMは合理的な投資家を想定した資産価格モデル, Fama-Frenchの3ファクターモデルは企業の超過収益をマーケットファクターや割安株効果, 小型株効果といった要素で説明しようとする実証モデルである[4][5] . 更に, 昨今, 必ずしも合理的でない投資家を考慮した行動ファイナンス（Behavioral Finance）[6] に関する議論も近年関心を集めており, その議論に広がりをみせている[7][8][9] .

近年, 情報処理技術の進展を背景として従来の推計モデルに加え, 金融市場を取り巻く大規模情報を分析しようとする新たな試みが盛んに行われている. 具体的には, 企業の開示情報やニュース, マイクロブログといった広く利用可能なテキスト情報を分析対象とし, その大規模情報に対して機械学習手法を用いて分析する研究である. これらは, 金融市場を説明する資本資産価格理論, テキスト情報を分析するため単語や接続詞等をはじめとして, 文章の特性を考慮して処理を行う自然言語処理（NLP：Natural Language Processing）, さらにはデータから一定の特徴やパターンを学習し, 分類する機械学習（Machine Learning）といった研究領域との交点であり, 学際的な領域である.

先行研究で言えば, ニュース記事を機械学習手法の一つ SVM（Support Vector Machines）によって分類し, 株価動向に関して分析を行った研究[10] や, Twitter フィードから二種類の方法, (1) Opinion Finder による Positive, Negative の 2 値分類, (2) Google-Profile of Mood States （GPOMS）による Calm, Alert, Sure, Vital, Kind, Happy の 6 値分類, で個人の感情を分類し, それらの情報を用いて投資意思決定に及ぼす影響について分析した研究[11] , 深層学習（Deep Learning）と呼ばれる多階層ニューラルネットワークモデルの一つでもある RNN-RBM（Recurrent Neural Networks Restricted Boltzmann Machine）を用いて, ニュース記事から時間的に変動する株価の上昇, 下落を予測した研究[12] 等, 国内外問わず数多くの研究報告がなされている. 深層学習について詳しくは, 次章にて説明する. また, SEC（米証券取引委員会）に提出された 10-Ks（年次報告書）を対象とし, ファイナンスに特化した辞書(ファイナンス用辞書)を用いた分析も行われ

ている. ファイナンス用辞書には, Positive や Negative 等の極性を持つ単語群が定義されているが, これら辞書を用いて分析することにより, 心理社会学辞書の H4N（Harvard-IV-4 TagNeg）を用いて分析をしたものと比較して, 誤分類による影響が緩和され, 説明力が向上するとの報告が行われている[13]. 更に, これらファイナンス用辞書を金融市場の分析に活用した報告なども行われている[14].

このように, 新たな手法によって金融市場を説明しようとする試みは, 国内外において数多く提案されており研究の意義は大きい.

2.2　深層学習

近年，深層学習（Deep Learning）と呼ばれる手法が関心を集めている．深層学習とは，機械学習手法の一つであるニューラルネットワーク（Neural Network）を応用し，多層に重ねることによって抽象度の高い複雑なデータの表現を学習する手法である．そして，その特徴のため，表現学習（Representation Learning）とも呼ばれている[15]．深層学習は，従来のニューラルネットワークと比べ，より深い（Deep）階層構造を持つモデルであるとされ（Fig.2-1），複雑なデータの特徴を学習する精度が高まり，画像認識分野[16]，音声認識分野[17]，さらに化合物反応予測等において，従来手法と比較して圧倒的な性能を達成する等，様々な分野への応用が試みられている．ニューラルネットワークは，原型と言えるパーセプトロン（Perceptron）が 1958 年に Rosenblatt, F.によって提案[18]されて以降，長い間研究されてきた領域であり，その名の通り人間の脳神経回路網を模した情報処理モデルである．

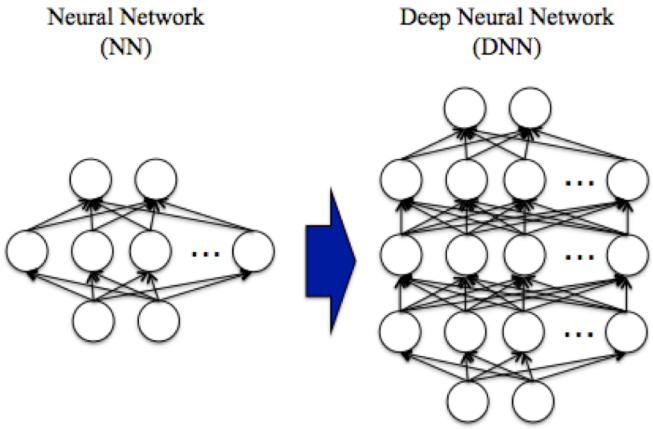

Fig. 2-1: ニューラルネットワークと深層学習の代表的モデル（DNN）．

近年になってこの多階層ニューラルネットワークによる機械学習技術が急速に進展した．その理由は大きく三点ある．第一に，背景的な進展．第二に，技術的な進展．第三に，新たな手法の提案である．過去，多層ニューラルネットワークの様々なモデルが提案されてきたが，学習の際に訓練データのみを学習し未知データに対しては適合できないといった過学習（Overfitting）が起こる学習の困難さや，学習に膨大な計算が必要なため非常に長時間かかるといった計算処理能力が依然として解決されない課題であった．しかしながら，昨今，(1) HW コストの顕著な低下からくる大規模データ獲得の容易さ等といった背景的な進展，(2) 大幅なチップ処理能力の増加といった技術的な進展，(3) 膨大なデータを多層ニューラルネットワークに入力し，層ごとに事前学習（pre-training）という教師なし学習を行うことによる過学習の回避と，事前学習後にネットワーク全体

の微調整を行う教師あり学習（fine-tuning）という新たな手法が提案され，これらの課題に対する解決策の糸口が提案された[19][20][21]．そして，現在では数多くの海外および国内の研究者が，これらの手法を応用する等，意欲的に研究に取り組んでいる．

深層学習は，自然言語処理分野（NLP, Natural Language Processing）の様々なタスクにも応用がなされており，従来手法と比較し性能は向上している．例えば，単語を固定長ベクトルで表現し，類似する単語が類似したベクトルを持つよう意味を埋め込み（word embedding）表現する分散表現（Distributed Representation）[22][23]，言語における単語の連なりの条件付き確率を学習する NNLM（Neural Net Language Model）と呼ばれる統計的言語モデル[24]，そして，品詞タグ付け，構文解析(Parse)を行うチャンキング，固有名詞等を抽出する固有表現抽出，動作主や相手，対象等に単語を分類するための意味役割付与，構造解析といった NLP タスクを統一的に扱う畳み込みアーキテクチャー（Convolutional Neural Network）の適用[25]，といった様々なタスクにおいて，深層学習の圧倒的に優れた精度が報告されている．とりわけ，構成的意味論[26]分野において，Socher らが提案した RNTN（Recursive Neural Tensor Network）の応用である句，文の評判極性分類手法[27] は，Stanford Sentiment Analysis とも呼ばれており，従来の RNN（Recursive Neural Network）[28][29] や MV-RNN（Matrix-Vector Recursive Neural Network）[30]，ナイーブベイズ（Naive Bayes），SVM 等の分類性能を超え，長文に対する評判極性の分類タスクにおいて最先端の精度 80%～85.4%を達成している．

評判極性分類手法（RNTN）は，文章を単語群から成る二分木構造によって表現し，ボトムアップ方式で句や文の評判極性を合成するモデルである．RNTN の入力として一様分布からランダムサンプリングによって生成した単語ベクトルを使用し，中間ノードでテンソル（Tensor）と呼ばれる極性の重み付け演算を行い，評判極性を出力する．このテンソル演算がパラメータの指数的増加を防ぐことに成功している．従来の RNN には，テンソルという概念はなく，否定文では高い精度を達成することが出来なかった．また，MV-RNN は，RNN に比べて高い性能を持つものの扱う文書の語彙数によってパラメータ数が膨大になるという課題があった．

Socher らの提案した評判極性手法（RNTN）の学習データは，きめ細かいセンチメントラベルが付与された，215,154 句，11,855 文のレビューデータを用いており，Pang and Lee(2005)[31] のデータを利用したものとなっている．また，デモモジュール，訓練・テストコード，Stanford Sentiment Treebank のデータセットが，Web ページ[2] に公開されている．

[2] 本稿においては，評判極性分類において，Web に公開されている RNTN（Stanford Sentiment Analysis）モジュールを採用した．
http://nlp.stanford.edu/sentiment/

2.3　分散表現

分散表現とは，単語をK次元で一意に表現するという1-of-K符号化によって得られるベクトルをより低次元で表現したものであり，意味が近しい単語同士はそのベクトル距離が近くなるような性質を有する表現である．そして，分散表現を獲得する方法は，様々あるがCBOW（Continuous Bag-of-Words）は，近年注目を集めている新たに提案されたニューラルネットワークモデルの一つであり，単語の分散表現を高精度で獲得できるという特徴を有している[33][34][35].

従来から言語情報の分析には，文章中に現れる単語を個別に扱うBag-of-words（BOW）によって表現する手法が一般的であった．しかしながら，単語の順序性の欠如，扱う単語数によって扱う次元数が膨大となる等の欠点があった．これに対して，CBOWによって学習した分散表現はBOWの課題を克服し，更に精度，学習速度も向上するとの報告が行なわれている．精度及び学習速度の向上には，ロジスティック回帰を階層的なグループに対して用いることでソフトマックスを近似する階層的ソフトマックスや，ランダムに偽の入力を選び，その偽の入力で正解の出力が出る確率が下がるように学習するネガティブサンプリングといった手法が採用されていることによって実現されている．

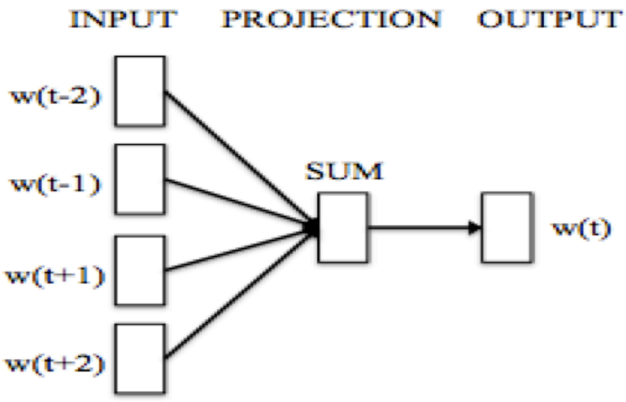

Fig. 2-2: CBOWの概略.

Fig.2-2は，CBOWによる分散表現の獲得方法の概略を示したものである．図の左から，Input, Projection, Outputとなっている．CBOWにおいては，注目する単語w(t)の前後の単語群w(t-2)，w(t-1)，w(t+1)，w(t+2)から構成されるBOWを入力とし，注目する単語

w(t)を出力するニューラルネットワークの学習により，分散表現が獲得される[3]．また，分散表現（単語ベクトル）を学習した結果を基に，各単語と距離が近い単語群を出力することが可能である[4]．なお，本分析では，コサイン距離を採用した．

[3] 本稿においては，分散表現学習において，word2vec と呼ばれるモジュールを採用した．
https://code.google.com/p/word2vec/
[4] 本稿においては，距離算出において，distance モジュールを採用した．

3 データ

3.1 金融市場データおよび市場関連データ

本稿では, 金融市場を取り巻くテキストデータおよび市場関連データを用いて分析を行う．テキストデータに関しては，世界で最も広く知られたニュース提供会社の一つである Thomson Reuters 社により提供されているニュースを採用した. 具体的には, 世界のマーケット動向に関するニュース News Feed Direct （NFD）を用いた. NFD は, News Scope Direct としても知られており, ニュースのヘッドラインや経済イベントを極小の遅延で配信し, 発表時刻もミリ秒単位で保持している等, 分析に適した特徴を有している. また, NFD は，世界各国の市場を対象としたニュースが含まれており，ニュースの言語も英語，フランス語，ドイツ語，日本語などをはじめ多岐にわたる.

市場関連データについては, 日経 NEEDS および Thomson Reuters Datastream より入手した．また，本分析ではファクターリターンを対象とした分析を行う．ファクターモデルに関するデータ（以下, FF ファクター）は, 久保田, 竹原(2007) [32] に従い、日本における東証 1 部, 東証 2 部から構成される銘柄から算出した 3 ファクター（マーケットファクターを表す Rm - Rf , 小型株効果を表す Small minus Big : SMB, 割安株効果を表す High minus Low : HML）データを用いた.

3.2　サンプル期間

分析に用いたデータのサンプル期間は，2003 年 1 月 1 日～2012 年 7 月 31 日とした．また，このような膨大な量のニュースの中から本分析では，とりわけ，日本市場及び日本企業に関する英語ニュース記事 411,531 件を対象として分析を行った．日次の NFD のニュースの件数は，最もニュースが多い日で 947 件．最もニュースが少ない日で 0 件となり，平均すると 117.6 件であった（Fig.3-1）．また，ニュース記事は，1,349 万行，9,265 万単語を含み，対象記事の内，日本企業に関連する記事は，363,970 件，該当する企業数は 308 件であった．

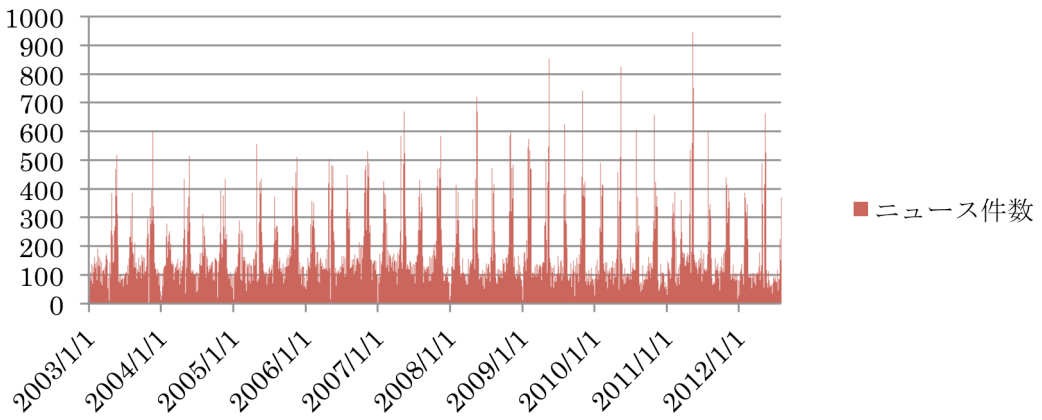

Fig. 3-1: 日次 NFD 件数.

4 深層学習を用いた評判分析手法

本章では, 第一の分析アプローチとして, 深層学習による評判分析手法を用いた分析について説明する.

本章の構成は, はじめに本分析手法の概略について説明し, 次に, 分析に用いるデータ及び分析方法, 次いで, 分析結果, 考察, まとめを説明する. 尚，本分析手法が，既存の研究と比較して新たな点は，三点ある．第一に分析するデータに即時性の高い経済金融ニュースである NFD を用いている点，第二に深層学習という近年急速に発展した機械学習手法を採用している点，第三に機械学習手法によって得られた結果と市場のファクターリターンとの関係性を分析している点が挙げられる.

4.1 はじめに

本分析は, Socher らによって提案された深層学習による評判分析手法（RNTN）を用い金融市場を対象とした分析を行う. 本手法を採用した目的は, 従来文章の極性判断の際に一般的であった単語極性による分類ではなく, 最先端の可変長文章の評判極性で極性分類することによって, 従来手法と比較して分析精度の向上, そして, より進んだ分析を試みるためである．具体的には, RNTN（Stanford Sentiment Analysis）モジュールを用いて文章の評判極性を分類し, その分類結果と資産価格変動の分析を行う. なお, 本分析では, 日本の株式市場を分析対象とし,ファイナンス分野において広く用いられているファクターリターンとの関係性について分析を行う.

尚，本章では，次の三点の仮説が検証を試みる．第一に, 金融市場は経済ニュースをはじめとした市場の情報量と関係性を有するという仮説. 第二に, NFD 記事の評価極性の分類が可能となり, その評価スコアが市場に対して影響力を有するという仮説. そして, 第三に, ネガティブな評判情報は数日間市場に対して影響力を有するという仮説である.

4.2 データ

テキストデータに関しては, 世界で最も広く知られたニュース提供会社の一つである Thomson Reuters 社により提供されているニュースを採用した．また，市場関連データについては，日経 NEEDS および Thomson Reuters Datastream より入手した．本分析ではファクターリターンを対象とした分析を行うため，ファクターモデルに関するデータ（以下, FF ファクター）は, 久保田, 竹原(2007) [32] に従い、日本における東証 1 部, 東証 2 部から構成される銘柄から算出した 3 ファクター（マーケットファクターを表す

Rm - Rf，小型株効果を表す Small minus Big : SMB，割安株効果を表す High minus Low : HML）データを用いた.

本分析で用いる NFD データのサンプル期間は，　2003 年 1 月 1 日～2003 年 12 月 31 日とした.

4.3　分析方法

本分析では，金融市場において配信されているニュースとマーケットとの関係性を調査する．具体的には，日次 NFD を RNTN（Stanford Sentiment Analysis）モジュールによってセンチメント・インデックスを算出し，マーケットのファクターリターンとの関係性について分析を行う.

分析に先立ち Fig.4-1 に示すシステムを構築した．以下のシステムは主に 4 つのモジュールから構成されている．具体的には，　(1) NFD データベースから記事を抽出, (2)　抽出した記事群の整形, (3) Stanford Sentiment Analysis(RNTN)モジュールをバッチ形式で連続実行, (4)　結果を CSV ファイルで保存，等のモジュールである.

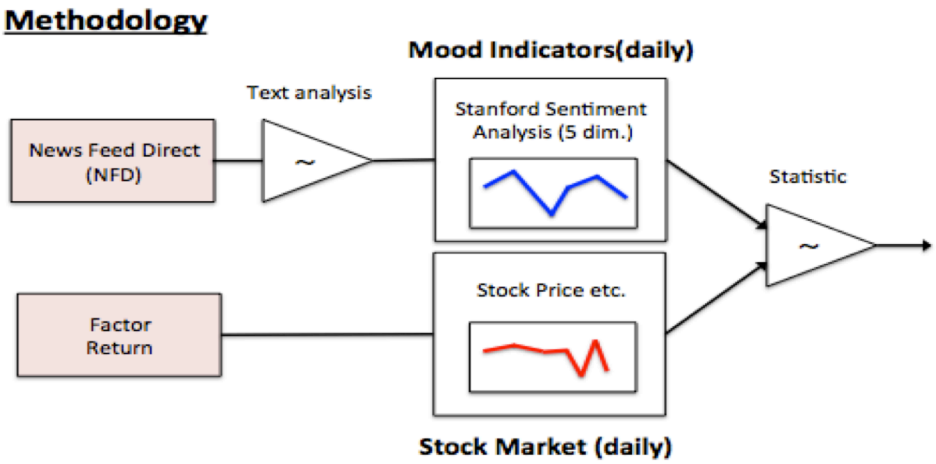

Fig. 4-1: システム概要図.

具体的な分析方法は次のとおりである.

(1) NFD を日次記事群として 1 日分を 1 ファイルとして抽出する．ただし, 1 日分の記事群は，東証の取引時間(9:00～15:00)を考慮して，ある t 日に発表された記事群は，日本標準時+0900(JST)　で　t -1 日(前日)の 15:00 以降, t 日(当日)の 15:00 迄の間に発表された記事を含めるものとした.

(2) 全ての日次記事群に対して, URL や E メールアドレス, 記事途中のピリオドの除外処理を行う.

(3) 全ての日次記事群を 1 ファイル毎に学習済みの Stanford Sentiment Analysis モジュールにインプットする.

(4) モジュールがアウトプットした評判極性分類を結果ファイル(1 日分 1 ファイル)として保存する.

(5) 結果ファイル中の評判極性毎に設定したセンチメントスコア（評判極性のスコアは, Very Positive , Positive を +1 , Neutral を 0 , Negative , Very Negative を -1 として設定）を合計し, 日次センチメント・インデックスを算出する.

(6) 算出した日次センチメント・インデックスとファクターリターンとの関係性を調べる. 日次センチメント・インデックスは, 他手法との比較可能性を考慮し, Bollen(2011)[11] に従い, Z-score に規格化した時系列データを算出した. Z-score の定義式を以下に示す.

$$Z_{Xt} = \frac{Xt - \bar{x}(X_{t\pm k})}{\sigma(X_{t\pm k})}$$

ここで，Xt はスコアの時系列データである.

$\bar{x}(X_{t\pm k})$, $\sigma(X_{t\pm k})$は, それぞれ，k 日間平均, k 日間標準偏差である.

上記プロセスを通じて，日次センチメントスコアを算出し, 金融市場との関係性を分析する.

4.4　分析結果

まず，日次 NFD 記事群の各文章を，学習済み Stanford Sentiment Analysis モジュールによって, Positive 及び Negative の評判極性を付与し，日次センチメントスコア及び Z-score(k=2)を算出した．算出した Positive, Negative それぞれの Z-score 値の基礎統計量を以下に示す．また，その時系列スコアをグラフに示す．（巻末 **A-1　深層学習を用いた評判分析によるセンチメントスコア** Fig.A-1, Fig.A-2 参照）

基礎統計量	Z(Positive)	Z(Negative)
平均	0.0	0.0
中央値	0.0	0.1
標本分散	0.8	0.7
標準偏差	0.9	0.9
尖度	-0.8	-0.7
歪度	0.1	-0.3
標本数	361	361

Fig. 4-2: センチメントスコア（Z-score）の基礎統計量.

次に，算出した日次のセンチメントスコア自身の特性について分析するため，時系列分析（ARMA: Auto Regressive Moving Average）を行った結果, Negative スコアに数日間に渡った自己相関がみられた．（巻末 **A-2　センチメントスコアの時系列分析** Fig.A-3, Fig.A-4 参照）

更に，算出したセンチメントスコアとファクターリターンの関係性を分析した結果, Negative スコアとマーケットファクターに強い関係性が見られた．具体的には, Negative スコアは，マーケットファクターを説明（有意水準 5%, T 値-2.243）し，特に，時価総額が小さい企業に対して強い関係性（SL～SH の指標について有意水準 1%, T 値-3.612～-3.409）を有することがわかった．

'***' P< .001 , '*' P< .05 , '.' P< .1

被説明変数	説明変数	回帰係数	T値	調整済み決定係数
Rm – Rf	Negative	- .00052 *	-2.243	.016
BL	Negative	- .00059 *	-2.214	.016
BM	Negative	- .00045 .	-1.950	.011
BH	Negative	- .00076 *	-2.591	.023
SL	Negative	- .00075 ***	-3.409	.042
SM	Negative	- .00067 ***	-3.612	.047
SH	Negative	- .00065 ***	-3.536	.045

Fig. 4-3: センチメントインデックスとマーケットファクターとの関係性.

4.5 考察

まず, 深層学習を用いた本手法による分析の結果, 次の三点の仮説が検証できたと考える. 第一に, 金融市場は経済ニュースをはじめとした市場の情報量と関係性を有するという仮説. 第二に, NFD 記事の評価極性の分類が可能となり, その評価スコアが市場に対して影響力を有するという仮説. そして, 第三に, ネガティブな評判情報は数日間市場に対して影響力を有するという仮説である.

次に, 本手法による分析結果の特筆すべき点は, 株式市場が経済ニュースのネガティブな記事に継続的に反応しやすく, 企業規模, 特に, 中小企業に対して大きな影響力を持っているという結果である. この点について, いくつかの視点で考えることにする. まず, 株式市場の投資家の行動についての視点であるが, 記事にネガティブな記述が見受けられると投資家は, 中小企業に対しては数日内の短期的な期間内に株式売却する短期的な投資行動をとる可能性が高い. 一方, 大企業に対しては即時売却に繋がらない比較的長期的な投資行動をとる可能性が高いということが考えられる. 次に, 企業規模と事業の安定性についての視点では, ネガティブな記事が発表された場合でも, 大企業の株式に対しては反応しにくいということから, 大企業は事業の安定性が高いということが言える. そしてこれは, 中小企業と比較して幅広い事業ドメインで事業を行っておりネガティブ要因をヘッジしているといった可能性も高い. 一方, ネガティブな記事が発表された場合, 中小企業の株式が反応するということは, 中小企業は事業の安定性が低く, 大企業と比較してより特化や領域が少ない事業ドメインで事業を行っているため, ネガティブ要因に対してより直接的に影響が出てしまう可能性が高いのではないかということが考えられる.

また, 本手法の課題は, 三点挙げられる.

第一に, 大規模なニュース記事の可変長文章を一文ずつ分類することによる RNTN モジュールの実行速度である. 著者が分析に用いた環境は通常個人で持っているデスクトップコンピュータでありサーバ等と比較して分析等負荷が大きい処理には向いていない. そしてこれによって, 1 ヶ月分のニュースの分析時間が数ヶ月かかる等膨大な時間がかかり, サンプリング期間が非常に短くなってしまった. 処理性能の高いコンピュータを用いる, 分析する文章を事前に選定する等, 解決すべきであると考える.

第二に, より進んだ分析が必要な点である. 市場のセンチメントが企業規模の大小に対して異なる影響力を持つとした場合, 異なる業界に対する影響力はどうなるか等, より進んだ分析によって金融市場に対して影響を与える要因を明らかにしていくことが必要であると考えられる.

第三に, 本手法で用いた RNTN モジュールによって分類したポジティブな評判情報が補足出来, 市場との関係性が明らかになっていない点である. 考えられる理由の一

つは, 本手法学習データはファイナンス分野のデータではない点である. 評判と言っても分野が違えば別の表現になるため, いくつかの記事が誤分類されていることも想定でき, 精度に関しては今後改善の余地があると言える. 本手法の有用性についてもこの点を改善した後, 他手法との比較することが好ましいと考える.

4.6 まとめ

深層学習による評判分析を用いて算出されたセンチメントスコアは, スコア値がとりわけ大きい日があること等の結果を確認した. また, 算出した Negative スコアは, 数日間に渡った自己相関を有する特性があり, マーケットファクター, 特に, 時価総額が小さい企業に対して強い関係性を有することがわかった. ただし, 今回採用した手法の中で, サンプル期間, 分析環境, 学習データ, 精度, に関して今後改善の余地があると言えることがわかった.

5 分散表現学習によるファイナンス辞書作成手法

本章では, 第二の分析アプローチとして, 分散表現学習による辞書作成手法を用いた分析について説明する.

本章の構成は, はじめに分散表現の学習手法である CBOW（Continuous Bag-of-Words）の概略について説明し, 次に, 分析に用いるデータ及び分析方法, 次いで, 分析結果, 考察, まとめを説明する．尚，本分析手法が，既存の研究と比較して新たな点は，四点ある．第一に分析するデータに即時性の高い経済金融ニュースである NFD を用いている点，第二に分散表現学習という近年急速に発展した機械学習手法を採用している点，第三に機械学習手法によって得られた結果とファイナンス領域で広く用いられている辞書を活用して新たな辞書の作成を試みた点，第四に新たな辞書を用いて市場のファクターリターンとの関係性を分析している点が挙げられる.

5.1 はじめに

本分析は，金融市場ニュースのテキスト情報に対して，CBOW（Continuous Bag-of-Words）を用いて学習した分散表現を活用し, 新たなファイナンス辞書の作成を試みる. これは, 当手法を通じて金融市場ニュース特有の表現やファイナンス分野における評判極性を持つ単語群を新たに抽出し, その単語群と金融市場との関係性の分析を試みるものである. なお，本分析においても，日本の株式市場を対象としたニュースを分析対象とし,ファイナンス分野において広く用いられているファクターリターンとの関係性について分析を行う.

尚，本章では次の三点の仮説の検証を試みる．第一に, ファイナンス辞書に定義されておらず経済金融ニュース特有かつ極性を有する単語（表現）が存在しているであろうという仮説. 第二に, 分散表現によって新たに獲得した単語群を加えた新辞書によって算出したスコアが金融市場に対する説明力を向上させるという仮説. そして, 第三に, 算出したスコア（NFD の時系列極性スコア）はマーケットファクター及び企業規模ファクター（小型株効果）等に対しても影響力を有するという仮説である.

5.2 データ

本分析では, 金融市場ニュースおよび既存の辞書[13]をベースにし，新たにファイナ

ンス用辞書の作成を行う[5].

辞書作成に用いるテキストデータに関しては，世界で最も広く知られたニュース提供会社の一つである Thomson Reuters 社により提供されているニュースを採用した．また，市場関連データについては，日経 NEEDS および Thomson Reuters Datastream より入手した．本分析ではファクターリターンを対象とした分析を行うため，ファクターモデルに関するデータ（以下, FF ファクター）は，久保田，竹原(2007) [32] に従い、日本における東証 1 部，東証 2 部から構成される銘柄から算出した 3 ファクター（マーケットファクターを表す Rm - Rf，小型株効果を表す Small minus Big の SMB ファクター，割安株効果を表す High minus Low の HML ファクター）データを用いた.

金融市場ニュースのサンプル期間は，2003 年 1 月 1 日～2012 年 7 月 31 日とした.

[5] 既存の辞書については，前述のファイナンス用辞書データ[13]を用いた.
http://www3.nd.edu/~mcdonald/Word_Lists.html

5.3　分析方法

本分析では，金融市場において配信されているニュースから分散表現を学習し，新たなファイナンス辞書の作成を試みる．具体的には，全期間の NFD を分散表現学習モジュールによって，単語の分散表現を学習し，既存辞書の単語リストとの関係性に注目して分析を行う．

分析に先立ち Fig.5-1 に示すシステムを構築した．以下のシステムは主に 4 つのモジュールから構成されている．具体的には，(1) NFD データベースから記事を抽出, (2) 抽出した記事群の整形, (3) ファイナンス辞書から単語を読み込み, distance モジュール（インプットした単語とコサイン距離が近い単語を出力）をバッチ形式で連続実行, (4) 出力された単語リストを読み込み，日次 NFD での出現頻度を日次カウントし結果を CSV ファイルで保存，等のモジュールである．

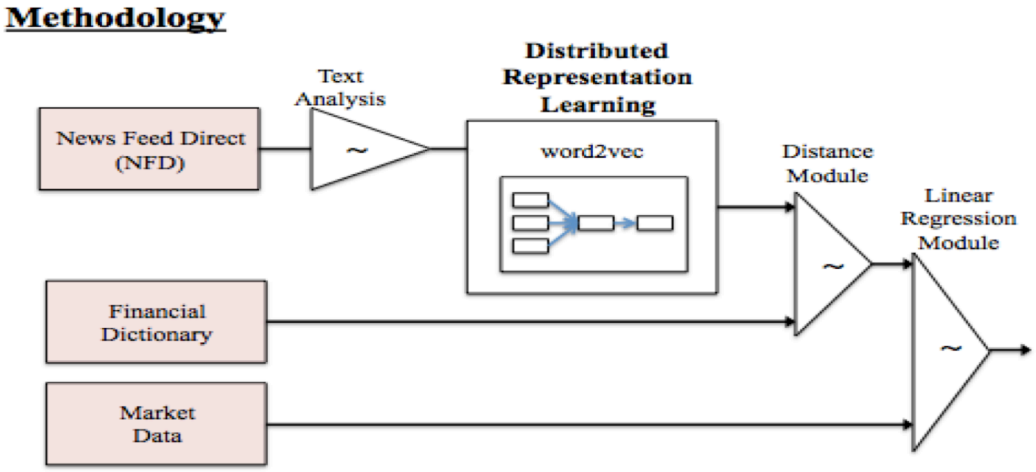

Fig. 5-1: システム概要図.

Fig.3-2 は，分析方法の概略を示したものである．図中の左部分は，ニュース情報(NFD)および既存のファイナンス用辞書(Financial Dictionary)を示している．本分析では，これら情報を基に単語群の作成を行う．

具体的な分析方法は，次のとおりである．

(1) 全期間のニュースデータ(NFD)を 1 つのファイルに抽出する．

(2) 分析対象のニュースデータ(NFD)に対して，必要な前処理を行う．今回採用するモジュールによって文章を学習するためURL及びEメールアドレスといった文章以外の除外処理を行う．

(3) ニュースデータ(NFD)ファイルを分散表現学習モジュールに入力し，分散表現を学習する．本分析では，学習手法は階層化ソフトマックスとし，考慮する文脈サイズは 5（w(t-5)〜w(t+5)）とした．

(4) 分散表現学習モジュールの学習結果を，距離算出モジュールに入力し実行する．

(5) ファイナンス辞書の Positive, Negative 単語を距離算出モジュールに順次入力し，分散表現のコサイン距離が近い単語群の一覧を取得する．尚，今回新たな工夫した点は本工程である．ファイナンス辞書の単語群は数千単語存在しているため手動でモジュールに入力して結果を得ると膨大な時間と手間がかかるため，この処理スクリプトを作成し自動化した点である．

(6) ファイナンス辞書で定義済み単語 及び 獲得した新単語 に対して日次 NFD 出現頻度を計上し，センチメントスコアを算出する．

(7) センチメントスコアとマーケットデータの関係性を分析する．

上記プロセスを通じ，単語群の作成を行い金融市場との関係性を分析する．

5.4　分析結果

まず, ファイナンス辞書に Negative 単語として定義されている単語群に対して分析を説明する. 次に, 定義済み単語群と金融市場ニュースの分散表現学習結果の双方を用いた分析結果, Positive 単語についての分析結果を説明する. 更に, それらの結果から抽出した新たな単語群の分析結果, 辞書を用いた市場センチメントスコアの算出結果, センチメントスコアと金融市場との関係性分析結果を説明する.

5.4.1 定義済み Negative 単語

はじめに, 定義済み Negative 単語を対象とした分析を行った. 分析の結果, 全 2,329 件中, 金融市場ニュースに出現した単語は 1,933 件, 出現しない単語は 396 件であることがわかった. 次いで, 距離算出モジュール(distance モジュール)に, この出現した単語のみ 1,933 件を入力し, コサイン距離の近い単語リストを獲得した. 結果について分析をしたところ, コサイン距離が 0.7 以上の単語数は, 67 単語であり, それらのうち, 既存辞書に含まれていな単語は 45 単語であった.

5.4.2 定義済み Positive 単語

次いで, 定義済み Positive 単語を対象とした分析を行った. 分析の結果, 全 354 件中, 金融市場ニュースに出現した単語は 335 件, 出現しない単語は 19 件であった. 次いで, Positive 単語についても同様にコサイン距離の近い単語群の抽出を試みた結果, コサイン距離が 0.7 以上の単語数は 35 単語であり, それらのうち, 既存辞書に含まれていな単語は 20 単語であった.

5.4.3 新たに抽出した単語群

ファイナンス辞書に定義されている単語群と距離が近いベクトルとして学習されていた単語群の内, 特に, コサイン距離 0.7 以上として新たに抽出された単語群は, 重複を除くと 54 件あった. この中には, drop, fell など Negative な極性を持つ単語が含まれていた. また, 一方では, rose, climbed などといった Positive な極性を持つ単語もあることがわかった. (巻末 **A-3　分散表現学習によって得られた新たな単語群** Fig.A-5 参照)

5.4.4 センチメントスコアの算出（ファイナンス辞書）

まず, ファイナンス辞書に定義されている Positive 及び Negative 単語が日次 NFD に出現する頻度を算出した. 算出した出現頻度（センチメントスコア）の時系列データの基礎統計量を示す.

	Positive	Negative
標本数	3500	3500
平均	124.78	290.47
標本分散	7394.84	66454.51
標準偏差	85.99	257.79
尖度	0.20	13.20
歪度	0.53	2.46
中央値	120.00	248.00
最頻値	0.00	0.00
最大	537.00	3130.00
最小	0.00	0.00

Fig. 5-2: センチメントスコアの基礎統計量.

5.4.5 センチメントスコアの算出（新辞書）

次に, 新たに抽出した単語が日次 NFD に出現する頻度を算出した. そして, 出現頻度が高い上位 10 単語を下表に示す. 最も出現していた単語は, rose であり, 次いで, fell, rise, fall, jumped, drop, recovery, climbed, believe, allow であった.

No.	word	出現回数	出現確率
1	rose	59,947	18.5%
2	fell	54,878	16.9%
3	rise	40,494	12.5%
4	fall	28,664	8.8%
5	jumped	15,430	4.8%
6	drop	14,728	4.5%
7	recovery	13,321	4.1%
8	climbed	8,252	2.5%
9	believe	7,966	2.5%
10	allow	6,262	1.9%
全単語出現回数		324,663	

Fig. 5-3: 新たに抽出した単語の出現頻度と出現確率.

とりわけ, 出現頻度が高い上位 6 単語について以下のとおり極性を付与し, 日次センチメントを算出した.

・Positive　　:　rose, rise, jumped
・Negative　　:　fell, fall, drop

5.4.6 マーケットとの関係性分析

算出したセンチメントスコア（従来辞書に新辞書を加えたもの）とマーケットファクターと関係性を分析した結果, 従来のファイナンス辞書に比べ新辞書によって算出したスコアはより強い関係性（p 値, T 値, 決定係数）を有していることがわかった. また, Negative スコアはマーケットファクターを説明(有意水準 1%, T 値-7.482～-4.109)し, Positive スコアは企業規模（SMB ファクター）を説明（有意水準 1～5%, T 値-3.307～-3.216）する等, 強い関係性を示していた（Fig.5-4）. また, Positive スコアは特に大企業に対して正の相関を有していた（Fig.5-5）. Fig.5-5 中の BL, BM, BH は, 東証 1 部および東証 2 部から構成される銘柄の中で, 時価総額が中央値以上かつ時価総額に対する自己資本比率が 30%以下, 30~70%, 70%以上の企業を分類している表記である.

'***' P< .001 , '*' P< .05 , '.' P< .1

被説明変数	説明変数	回帰係数	T値	調整済み決定係数
Rm – Rf	Negative(従来)	- .00047 ***	-4.109	.00672
Rm – Rf	Negative(新辞書)	- .00501 ***	-7.482	.02286
SMB	Positive(従来)	- .00059 **	-3.216	.00395
SMB	Positive(新辞書)	- .00047 ***	-3.307	.00421

Fig. 5-4: センチメントスコアとマーケットファクターとの関係性.

相関係数	Positive
SL	.00031
SM	.01317
SH	.01121
BL	.03038
BM	.04368
BH	.03821

Fig. 5-5: センチメントスコアとマーケットファクターとの関係性.

5.5 考察

まず, 分散表現学習を用いた本手法による分析の結果, 次の三点の仮説が検証できたと考える. 第一に, ファイナンス辞書に定義されておらず経済金融ニュース特有かつ極性を有する単語（表現）が存在しているであろうという仮説. 第二に, 分散表現によって新たに獲得した単語群を加えた新辞書によって算出したスコアが金融市場に対する説明力を向上させるという仮説. そして, 第三に, 算出したスコア（NFD の時系列極性スコア）はマーケットファクター及び企業規模ファクター（SMB）等に対しても影響力を有するという仮説である.

次に, 本手法による分析結果は特筆すべき点が二つある.

一つ目は, 経済ニュース特有の表現として, rose, fell, rise, fall, jumped, drop, recovery, climbed, believe, allow といった極性を有する可能性のある単語群が抽出できた点である. しかしながら, これらの単語群は極性を有する単語として常識的に考えられる. では, なぜ従来の辞書には含まれていなかったのか. 考えられる理由は, 従来の辞書が SEC（米証券取引委員会）に提出された 10-Ks（年次報告書）を分析対象としており, 企業が発表する情報から作成しているためである. つまり, 企業発表等の情報を受けてから記事にする経済ニュースに出てくる表現や単語は, 表現や単語が従来の辞書と異なり, 捉えることができなかったのではないかと想定できる.

二つ目の特筆すべき点は, NFD から算出した時系列ポジティブスコアが企業規模ファクターの特に時価総額が大きい企業群の収益に影響力を有している点である. この点についていくつかの仮説を挙げる. まず, そもそもニュースで発表されるような情報で考えられる内容は二つある. 政治や経済, 政府発表のようなマクロ記事, 大企業の発表のようなミクロ記事である. このため, どちらのニュースとも関係性が強い大企業が, ポジティブな情報が発表されたことによってダイレクトに反応したとする仮説. 次に, そもそも発表されたニュースが大企業に偏っていたという仮説. 更に, ポジティブな情報がニュースで発表された場合, 投資家行動として, 前述のとおり市場との関係性の大きい大企業の株式に投資するという仮説等が考えられる.

本手法の課題は, 三点挙げられる. 一点目は, 新たな単語群を抽出する手法の調整である. word2vec モジュールは, 様々なパラメータが存在しておりそれらを調整することによって学習結果はドラスティックに変わる. また, 分散学習後に抽出する単語群の閾値としたコサイン距離の範囲調整も検討すべき課題である. 本分析では, 比較的良い結果となるパラメータおよび閾値を選択したが, 詳細な比較検討が必要であると考える. 二点目は, 分析に対する剛健性の確認である. 本分析では, 分散表現によって文章特有の表現等を抽出できたが, 他の記事群を分析対象とした場合, また, より短い期間を分析対象とした場合等, 市場との関係性の分析結果が変わることも考えら

れる. これらの確認が必要である. 三点目は, 類義語, 多義語といった表現の扱いである. 今より踏み込んだ分析を行うためにはこれらをどう扱うかについても検討が必要である.

5.6 まとめ

本手法によって抽出された, 新たな単語群を用いた金融市場の分析は, 従来のファイナンス辞書を使用した分析と比較して分析精度を向上させることが確認された. これは, 従来のファイナンス辞書および分散表現学習手法の併用によって, 分析対象の文章特有の表現や単語の偏りを捉えることができ, より進んだ分析を実施できる可能性が高まった. また, 算出した Negative スコアは, マーケットファクターに対して強い関係性を有し, 一方, Positive スコアは, 企業規模, 特に大企業に対して強い関係性を有するということも確認され, 当手法について一定の有用性を示せたと考える. ただし, 今回採用した方法の中で, 新たな辞書を抽出する際のいくつかのパラメータ (学習パラメータ, コサイン距離) や, 新辞書の極性付与の方法, 剛健性の確認については今後の課題である.

6 結論

本稿では, 二つの最先端の機械学習技術を用いて金融市場の分析を行い, いくつかの興味深い結果を得ることができた. 一つ目は, 株式市場は深層学習を用いた分析によって得られたネガティブな経済ニュースに継続的に反応しやすく, 企業規模, 特に, 中小企業に対して大きな影響力を有しているという結果. 二つ目は, 金融ニュースの分散表現を用いて極性を有する単語群を抽出できたという結果. そして, 三つ目は, 新辞書によって得られたポジティブスコアが大企業に影響力を有しているという結果である.

これらの結果から, 手法によらず共通的に得られる結果があること, 手法によって得られる結果が異なるということが言える. 共通的に得られる結果は, 市場はネガティブな情報と関係性を有するという点である. そして, 手法によって得られる結果が異なる点は, 一つ目はポジティブスコアと企業規模, 特に大企業との関係性について, 二つ目はネガティブスコアと企業規模, 特に中小企業との関係性である. 総じて, これら異なる結果が出たことに対するより進んだ分析を行う必要性があるということが明らかになった.

7 今後の課題

今後の課題は，大きく三つある.

第一に剛健性の確認である. これは，今回採用した二つの最先端の機械学習技術を応用した手法が未だ確立されておらず発展途上であることに起因する. これらの手法を用いた手順,及び検証環境も更なる整備が必要と考える.

第二に，より詳細なグループに対するファイナンス辞書の作成である．本研究により，ファイナンスの文脈でポジティブ，ネガティブを定義した辞書が，マーケット全体や企業規模に応じたグループの評価において一定の有用性を示せた．しかしながら，業界別グループや個別企業（例えば，輸出企業と輸入企業等）に対しては本研究の結果とは全く違う関係性を有する事が考えられる．これらの検証についても今後の課題である.

第三に，資産価格の予測である．本研究では，総じてセンチメントスコアと株式資産価格の同時点の関係性を分析している．しかしながら，本研究の延長線としては資産価格の予測および投資戦略の有用性に関する分析がある.

謝辞

本研究を進めるにあたり，丁寧かつ熱心なご指導を頂いた指導教員の 高橋大志 教授に感謝致します．また，日常の議論を通じて多くの知識や示唆を頂いた副査の方々および研究室の皆様に心より御礼を申し上げ，謝辞とさせて頂きます．

参考文献

[1] 久保田敬一 : 決定版コーポレートファイナンス. 東洋経済新報社, (2006).

[2] Brealey, Richard A., Myers, S., and Allen, F.: Principles of corporate finance., Tata McGraw-Hill Education, (2012).

[3] 加護野忠男, 砂川伸幸,吉村典久. コーポレート・ガバナンスの経営学: 会社統治の新しいパラダイム. 有斐閣, (2010).

[4] Jensen, Michael C.: Some anomalous evidence regarding market efficiency. , Journal of financial economics 6.2, 95/101. , (1978).

[5] Fama, Eugene F., and Kenneth R. French. : Common risk factors in the returns on stocks and bonds. , Journal of financial economics 33.1, 3/56. (1993).

[6] Shleifer, Andrei. Inefficient markets: An introduction to behavioral finance. , Oxford university press, (2000).

[7] 砂川伸幸,山崎尚志. : マーケットの非効率性と企業の投資・財務戦略., 國民經濟雜誌 186.3 , 65/77, (2002).

[8] Ikenberry, David, Josef Lakonishok, and Theo Vermaelen. : Market underreaction to open market share repurchases. , Journal of financial economics 39.2 , 181/208, (1995).

[9] DeLong, J. Bradford. : Noise trader risk in financial markets. , (1989).

[10] Schumaker, Robert P., and Hsinchun Chen. : Textual analysis of stock market prediction using breaking financial news: The AZFin text system. , ACM Transactions on Information Systems (TOIS) 27.2, 12, (2009).

[11] Bollen, Johan, Huina Mao, and Xiaojun Zeng. : Twitter mood predicts the stock market. , Journal of Computational Science 2.1, 1/8, (2011).

[12] 吉原輝, 藤川和樹,関和広. : 深層学習による経済指標動向推定., 人工知能学会全国大会論文集 28 , 1/4, (2014).

[13] Loughran, Tim, and Bill McDonald.: When is a liability not a liability? Textual analysis, dictionaries, and 10 - Ks, The Journal of Finance 66.1 , pp. 35-65 (2011).

[14] Yamashita, Y., Jotaki, H., and Takahashi, H.: Analyzing the Influence of Head-Line News on the Stock Market in Japan, International Journal of Intelligent Systems Technologies and Applications,12, pp.328-341, (2013).

[15] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. : Representation learning: A review and new perspectives. , Pattern Analysis and Machine Intelligence, IEEE Transactions on 35.8: 1798/1828. (2013).

[16] Le, Quoc V.: Building high-level features using large scale unsupervised learning. , Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, (2013).

[17] Deng, Li, et al.: Recent advances in deep learning for speech research at Microsoft. , Acoustics,

Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, (2013).

[18] Rosenblatt, Frank. : The perceptron: a probabilistic model for information storage and organization in the brain. , *Psychological review* 65.6, 386 (1958).

[19] G.E. Hinton, S. Osindero, and Y. Teh,: A fast learning algorithm for deep belief nets, Neural Computation, vol 18, (2006).

[20] Bengio, Yoshua, et al. : Greedy layer-wise training of deep networks. , Advances in neural information processing systems 19: 153. (2007).

[21] Poultney, Christopher, Sumit Chopra, and Yann L. Cun.: Efficient learning of sparse representations with an energy-based model. , Advances in neural information processing systems. (2006).

[22] Hinton, Geoffrey E.: Learning distributed representations of concepts, Proceedings of the eighth annual conference of the cognitive science society. Vol. 1. (1986).

[23] Mikolov, Tomas, et al.: Distributed representations of words and phrases and their compositionality. , Advances in Neural Information Processing Systems. (2013).

[24] Bengio, Yoshua, et al.: Neural probabilistic language models. , Innovations in Machine Learning. Springer Berlin Heidelberg, 137-186, (2006).

[25] Collobert, Ronan, and Jason Weston. : A unified architecture for natural language processing: Deep neural networks with multitask learning. , Proceedings of the 25th international conference on Machine learning. ACM, (2008).

[26] Socher, Richard, et al.: Semi-supervised recursive autoencoders for predicting sentiment distributions. , Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, (2011).

[27] Socher, Richard, et al.: Recursive deep models for semantic compositionality over a sentiment treebank. , Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). (2013).

[28] Goller, Christoph, and Andreas Kuchler.: Learning task-dependent distributed representations by backpropagation through structure. , Neural Networks, 1996., IEEE International Conference on. Vol. 1. IEEE, (1996).

[29] Socher, Richard, et al.: Parsing natural scenes and natural language with recursive neural networks. , Proceedings of the 28th International Conference on Machine Learning (ICML-11). (2011).

[30] Socher, Richard, et al.: Semantic compositionality through recursive matrix-vector spaces. , Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, (2012).

[31] Pang, Bo, and Lillian Lee. : Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. , Proceedings of the 43rd Annual Meeting on

Association for Computational Linguistics. Association for Computational Linguistics, (2005).
[32] 久保田敬一,竹原均. : Fama-french ファクターモデルの有効性の再検証. , 現代ファイナンス 22, 3/23 , (2007).
[33] Mikolov, Tomas, et al.: Efficient estimation of word representations in vector space., arXiv preprint arXiv:1301.3781 (2013).
[34] Mikolov, Tomas, et al.: Distributed representations of words and phrases and their compositionality., Advances in Neural Information Processing Systems. (2013).
[35] 西尾泰和.: word2vec による自然言語処理, オライリージャパン, (2014).

Appendix

A-1　深層学習を用いた評判分析によるセンチメントスコア

本章では，第 4 章の**深層学習を用いた評判分析手法**で算出したセンチメントスコアについて説明する．

Fig.A-1, Fig.A-2 は, 2003 年 1 月 1 日～2003 年 12 月 31 日までのセンチメントスコアの時系列変動を表している. Fig.A-1 はポジティブスコア, Fig.A-2 はネガティブスコアを表し，図の横軸はニュース記事の発表日，縦軸はセンチメントスコアの Z-score 値を表す．分析対象のデータは，東証取引日以外の休業日(日曜日，国民の祝日，前日及び翌日が国民の祝日である日，土曜日，年始 3 日間及び 12 月 31 日)を除いた取引時間内の記事群のみを対象としている．スコアは，全体的に-1.0～+1.0 の間に収まっていることを確認でき, Z-score の値がとりわけ大きい日があることを確認できるが, これらの日においては，何らかのイベントが生じている可能性がある．

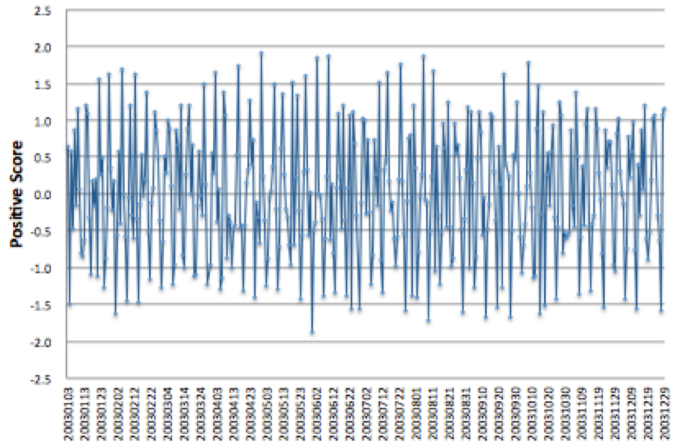

Fig. A-1: 日次センチメント（Positive Z-Score）.

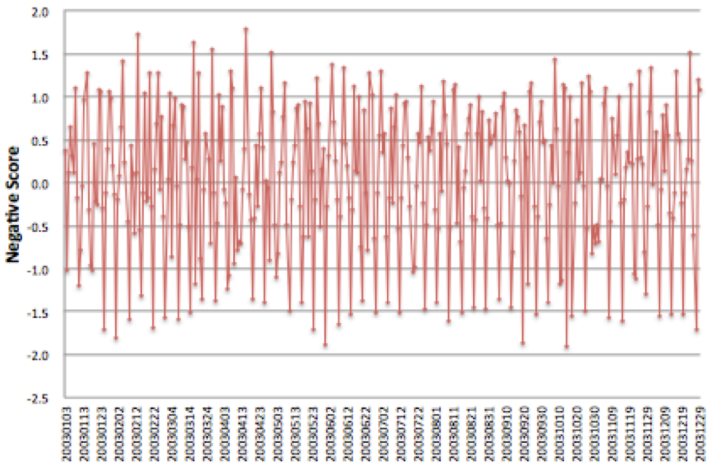

Fig. A-2: 日次センチメント（Negative Z-Score）.

A-2　センチメントスコアの時系列分析

本章では，第４章の**深層学習を用いた評判分析手法**で算出したセンチメントスコアの時系列分析結果について説明する．

Fig.A-3, A-4 は, 2003 年 1 月 1 日～2003 年 12 月 31 日までのセンチメントスコアの時系列データ自身の自己相関を表している．図の横軸は Lag を表し，縦軸は自己相関係数を表し，図の青破線は 95%信頼区間を示している．Fig.A-4 からは，Negative スコアが 1～5 日に渡って自己相関係数の値が大きくなっており連続的に自己相関があることがわかる．一方，Positive スコアは，自己相関係数が比較的ランダムな値をとっており連続性，周期性は確認できなかった．

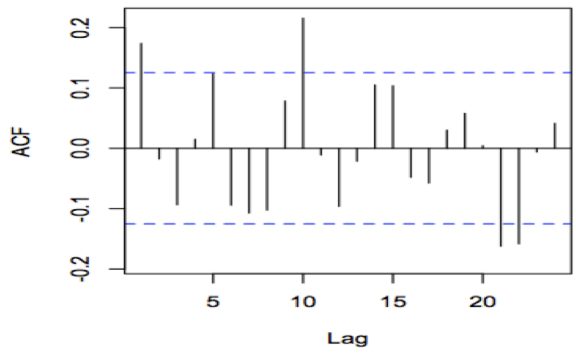

Fig. A-3:　自己相関グラフ（Positive スコア）．

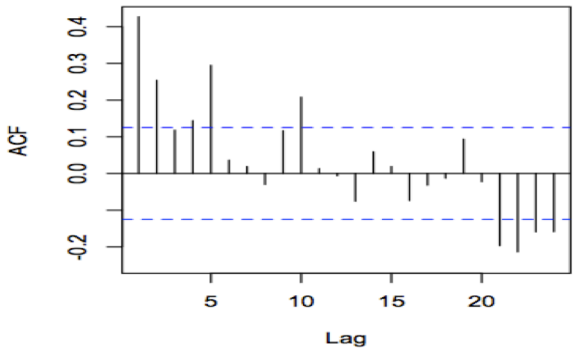

Fig. A-4:　自己相関グラフ（Negative スコア）．

A-3　分散表現学習によって得られた新たな単語群

本章では，第５章の**分散表現学習によるファイナンス辞書作成手法**で新たに獲得した極性を有する単語群について説明する．

Fig.A-5 は, 2003 年 1 月 1 日～2012 年 7 月 31 日までの NFD から CBOW によって学習した分散表現を元に，ファイナンス辞書の単語群[13]とコサイン距離が近い（閾値を 0.7 以上した）単語を示したものである．表の cosine distance はコサイン距離を表す．獲得した単語群の内，いくつかの単語は極性を有するものの, is , click のような極性を有さないものも含まれていること，また，類義語，多義語といった単語に対する扱いについても今後さらなる改善が必要であると言える．

No.	word	cosine distance	No.	word	cosine distance
1	accede	0.7239	28	midst	0.7314
2	acknowledged	0.7396	29	non-performing	0.7373
3	adhere	0.7204	30	noted	0.7371
4	allow	0.7575	31	probe	0.7252
5	bailing	0.7093	32	pull	0.7044
6	believe	0.7079	33	recovery	0.7119
7	bolster	0.7371	34	reeling	0.7204
8	buoyed	0.7446	35	refering	0.7291
9	capitalise	0.7313	36	relate	0.7003
10	capitalising	0.7070	37	reveal	0.7147
11	capitalize	0.7468	38	rise	0.7536
12	click	0.7934	39	rose	0.8305
13	climbed	0.8464	40	sensitivity	0.7035
14	cutting	0.7061	41	shed	0.7850
15	dipped	0.7091	42	slid	0.7696
16	drop	0.8082	43	slipped	0.7875
17	eke	0.7723	44	slump	0.7136
18	fall	0.7156	45	softer	0.7091
19	father	0.7047	46	supported	0.7397
20	fell	0.8147	47	surged	0.7803
21	fragility	0.7136	48	tumbled	0.7573
22	hereby	0.7535	49	upgraded	0.7643
23	influx	0.7154	50	upward	0.7012
24	is	0.7329	51	wake	0.7704
25	jumped	0.7696	52	weaker	0.7520
26	lead-up	0.7052	53	wipe	0.7026
27	lift	0.7879	54	wondered	0.7083

Fig. A-5:　分散学習の結果，新たに抽出された単語群（コサイン距離 0.7 以上）．

A-4　分析テキストの処理詳細

本章では，第５章の**分散表現学習によるファイナンス辞書作成手法**で分析対象テキストに対して機械学習手法を行う前の整形処理，コサイン距離算出モジュールの自動実行処理等，いくつかを取り上げ説明する．

１．　URL 及び E メール除外処理

テキストデータに含まれる URL 及び E メールアドレスを除外処理は以下のとおり実施した．厳密には URI（広義の URL）は RFC3986 で定義されている正規表現を用いることも考えられたが今回のテキストデータ中にはロイターの Web ページのトップ URL 等，より簡易的な表現で問題ないと判断した．また，E メールアドレスも RFC5321 及び 5322 に定義されているが，URL と同様の理由により簡易的な表現で除外処理を行った. スクリプト中の out.txt は変換元テキストデータを表す.

RemoveURI.sh

```
#!/bin/bash
perl -p -i.bak1 -e 's/https?:¥/¥/[-_.!~*a-zA-Z0-9;¥/?:¥@&=+$,%#]+//g' out.txt
perl -p -i.bak2 -e 's/^[a-zA-Z0-9._-]+¥@[a-zA-Z0-9.-]+//g' out.txt
```

２．　大文字から小文字への変換処理

ファイナンス辞書で極性が定義されている単語リストは全て大文字で定義されているため，それらを分析対象のテキストデータに対して検索してもヒットしない．このため，ファイナンス辞書の単語群及び分析対象データの双方を全て小文字に変換して検証した. 大文字から小文字への具体的な変換処理は以下のとおり実施した. スクリプト中の input.txt は変換したいデータ，out.txt は変換後出力されるデータを表す.

Upper2Lower.sh

```
#!/bin/bash
file=input.txt
cat $file | tr "A-Z" "a-z" > out.txt
```

３．　コサイン距離算出モジュール自動実行処理

ファイナンス辞書で定義されている単語群と NFD テキストから分散表現を学習した結果ファイルの二つを用いて，NFD 中で新たに極性を有する可能性のある単語群（コサイン距離と共に）リストを抽出する処理を次に示す．具体的には，コサイン距離を算出する distance モジュールにファイナンス辞書のデータをヒアドキュメントとして連続的にインプットする処理とした．スクリプト中の word_dic.txt はネガティブ単語が一行に一単語ずつ定義されているテキストファイル，result.bin は分散表現学習結果のバイナリファイル，result.txt は distance モジュールの出力結果を表す．

Get_Negative_Word_List.sh

```
#!/bin/bash
file=word_dic.txt
word=`cat $file`
/Word2vec/distance result.bin << EOF >> result.txt
$word
EXIT
EOF
```