金融指標予測のためのソーシャルメディアに適した分散表現獲得手法の検討

上田 健太郎[1]　諏訪 博彦[1]　小川 祐樹[2]　梅原 英一[3]　山下 達雄[4]　坪内 孝太[4]　安本 慶一[1]

概要：効率的市場仮説により金融時系列データのみを用いた金融指標の予測では，チャンスレートを大きく越えることは困難であることが知られている．そのため，これまで金融時系列以外のデータを活用し，市場の予測可能性を示す研究が多く行われてきた．その中で我々は日本最大の株式に関するソーシャルメディアである「Yahoo!ファイナンス掲示板」に着目した．ソーシャルメディア投稿文書には投稿者の意見や態度が反映されているため，金融指標を予測するための重要な情報源となると考えられるが，文書のノイズのために予測に有効な特徴表現を抽出することが難しい．本研究では先行研究の金融指標予測手法を拡張し，新たな特徴表現抽出法を行うことで，ソーシャルメディア文書においても金融指標予測に有効な特徴を抽出する．獲得した特徴表現を用いて予測を行い，先行研究と比較した結果，提案法が従来法の精度を上回ることを示す．

A Distributed Representation Acquisition Method Suitable for Social Media for Forecasting Financial Indicators

Kentaro Ueda[1]　Hirohiko Suwa[1]　Yuki Ogawa[2]　Eiichi Umehara[3]　Tatsuo Yamashita[4]　Kota Tsubouchi[4]
Keiichi Yasumoto[1]

1. はじめに

金融指標の一つであるボラティリティ・インデックス (VI) の上下動を予測することは投資におけるリスク低減につながるため，非常に重要なテーマである．しかし，金融指標の予測は市場の不確実性のため依然として困難な課題である．金融指標の予測は効率的市場仮説 [1] により金融時系列データのみを用いた予測では精度がチャンスレートと同程度になることが報告されている．そのため，多くの研究者は金融時系列データ以外のデータを用いて機械学習を行うことで金融指標の予測を行う手法を開発してきた．例えば，ニュースを用いた手法 [2], [3], [4] や企業の公式発表を用いた手法 [5], [6] などがある．我々はその中でもソーシャルメディアに着目する．ソーシャルメディアには投稿者の集合知が反映されている．ソーシャルメディアを解析することで人が理解している市場に関する本質的な特性を抽出することが可能であり，金融指標の効果的な予測を可能にすると考える．

本研究は，ソーシャルメディアの投稿文書から日経平均 VI（以下，VI）の予測に有効な特徴表現を抽出する手法の開発を目的とする．ソーシャルメディアの投稿文書はニュースや企業のプレスリリース，有価証券報告書のようなものと異なり，ほとんどの文章で文法誤りが存在する．このような文章では BERT のようなモデルは性能が低下することが知られており [7]，予測に有効な分散表現を獲得することが難しい．上田ら [8] は Latent Dirichlet Allocation(LDA)，Doc2Vec，BERT の三手法を用いて Yahoo!ファイナンス掲示板の投稿文書から分散表現を獲得し，VI 予測に用いて精度の比較を行なっている．実験の結果，言語モデルの違いで明確に精度の違いが表れていない．文書の文法誤りが原因で，日々の話題を反映した特徴ベクトルが獲得できていないと考えられる．我々は新たなアプローチでこの課題を克服し，VI 予測に有効な分散表現を獲得を目指す．第一に Sparse Composite Document Vector (SCDV)[9] を用いる

1　奈良先端科学技術大学院大学
2　立命館大学
3　新潟国際情報大学
4　ヤフー株式会社 Yahoo!JAPAN 研究所

ことで文書分散表現を獲得する．第二に獲得した分散表現を autoencoder[10] を用いて次元削減を行う．また，本研究ではベースラインとして Simple Word-Embedding-based Model(SWEM) での特徴量抽出の検証も行う．

SCDV は単語埋め込み表現にソフトクラスタリングを行い，各クラスターへの所属確率を考慮して分散表現の計算を行う手法である．したがって，LDA では不可能であった，文脈を考慮した文書分散表現の獲得が可能である．オートエンコーダは非線形な関係を処理できるため，ソーシャルメディアのセマンティック情報を考慮した次元削減手法として有効であると考える．本提案手法はソーシャルメディアに対して Doc2Vec や BERT を適用する際の課題（提案手法 3.2.2）も克服することができるため，ソーシャルメディア文書からの新たな特徴量抽出手法としてその有効性が期待できる．本研究の貢献は以下である．

- ソーシャルメディアの文書から新たなアプローチで特徴ベクトルを獲得し，機械学習による VI 上昇予測を行なった．先行研究との比較の結果，予測精度の向上が確認された．
- 金融時系列データのみを用いた手法と比較した結果，ソーシャルメディア文書を用いることで予測精度が改善することを示した．

本研究で用いたソーシャルメディアは日本最大の株式取引サイト「Yahoo!ファイナンス掲示板」であり，掲示板内の日経平均株価スレッドの投稿を分析した．先行研究 [8] と同様に，獲得した特徴量と金融時系列データを機械学習の入力とし，VI の大幅な上昇予測のタスクを行なった．VI の大幅な上昇を予測することで投資におけるリスク低減が可能である [11]．実験の結果，先行研究に比べ精度が向上し，F 値が 0.28 の VI 予測モデルを開発した．金融時系列データのみを用いて構築した予測モデルでは F 値が 0.25 の予測精度であったことから，ソーシャルメディアを用いることで VI 予測の精度が改善することが明らかになった．

2. 関連研究

2.1 機械学習による金融指標の予測

金融指標の予測は収益やリスク回避につながるため，非常に有益であり，これまで多くの先行研究が行われてきた．例えば，Chen ら [12] は LSTM を用いて中国株式のリターンをモデル化することで中国株式市場の予測を行なった．その結果，ランダムな手法と比較して，彼らのモデルは 14.3%から 27.2%へと精度を向上させることを達成している．さらに Long ら [13] は，金融時系列の特徴抽出と値動き予測タスクに特化したマルチフィルター・ニューラルネットワーク（MFNN）という新しい end-to-end なモデルを提案している．彼らは異なる特徴空間を捉えるために CNN と LSTM を合わせたシステムを提案し，結果として精度や収益性，安定性が従来の機械学習モデルや単一構造の深層学習モデルよりも優れていると報告している．また，株価の予測に敵対的学習を使用した研究もある．feng ら [14] は敵対的学習を用いることで，株価の確率を考慮した株価予測モデルを考案し，Xu ら [15] よりも平均的な w.r.t の精度で 3.11%の相対的な改善を報告している．金融時系列データのみを用いた研究では効率市場仮説 [1] により，チャンスレート程度の精度しかでないことが知られている．そこで，金融時系列データだけでなく，ニュース，ソーシャルメディアなどのテキスト情報を利用して市場の予測可能性を示す研究も多く行われている．

2.2 テキストを用いた金融指標の予測

Li ら [16] は記事の要約に注目し，ニュース記事の要約を用いた時の精度とと全文記事を使用した時の精度の比較を行なった．Finet のニュースアーカイブから取得した企業別と市場全体の両方のニュース記事を使用している．香港証券取引所の 5 年間のデータで実験を行なった結果，ニュース記事の要約を予測に用いることで，記事本文を用いるよりもパフォーマンスが向上することが示された．Li ら [17] は夜間ニュースを用いることで，前日の終値と始値の間の夜間の株価の動きを予測する方法を提案している．銘柄間の関係を利用するため，LSTM-RGCN モデルを提案しており，グラフを導入することで従来の手法では予測できなかった，ニュースとは直接関係のない銘柄の動きや市場全体の動きを予測することができることを示した．Chen ら [18] はミューチュアル・ファンドのポートフォリオ・データに含まれる投資行動のリポジトリから、潜在的な株式特性の表現を抽出することを行なった．実世界の株式市場のデータを用いた実験により，抽出した株式特性を用いて株式予測を行うことの有効性を示した．

金融ニュースや企業の公式発表は信頼性が高く，ノイズが少ないため金融指標予測に広く利用されてきた．近年，Twitter などのソーシャルメディアは急激な発展を遂げ，膨大なデータを提供するようになった．そのため，ソーシャルメディアを活用した金融指標予測に用いるアプローチの開発も発展してきた．

2.3 ソーシャルメディアを用いた金融指標の予測

Twitter を分析した事例として Bollen ら [19] は，ツイートを 2 値の感情レベル（ポジネガ）及び 6 種類の感情レベルに分類し，日別のツイートの感情と，ダウ平均株価との相関を調査している．その結果，ポジネガではダウ平均株価との相関関係は見られなかったとした．一方で“Calm”な感情は 2 日後から 6 日後の，“Happy”な感情は 6 日後のダウ平均株価と正の相関があることを示している．さらに，ダウ平均株価の変動を 87.6%の精度で予測できたとしており．この報告はソーシャルメディアの応用可能性を示した．Liu ら [20] は Transformer encoder と capsule

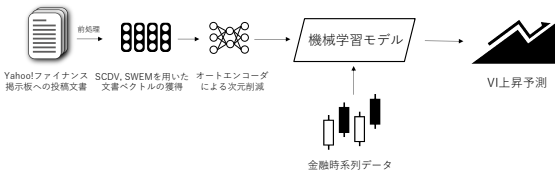

図 1 VI 予測モデルの概要

network の長所を組み合わせた，CapTE(Capsule network based on Transformer Encoder) モデルを提案し，ソーシャルメディアの投稿のみを用いて株価の上下動の予測を行なった．CapTE モデルにより，株式の動きの予測精度を向上させることができるツイート間の特定の関係を取得した．ランダムに選んだ 6 つの銘柄に対し Lavrenko の提案した戦略に従い実際の株式取引シミュレートを行うことで 20 取引日で最大リターン 17%越を得ている．

諏訪ら [21] は Yahoo!ファイナンス掲示板の投稿に対して LDA を用いることでトピックを獲得し，日経平均 VI の上昇予測を行なった．彼らは投稿を形態素解析で形態素に分割し，LDA を用いて各文書における 100 種のトピックに所属する確率を獲得した．獲得したベクトルを日別にまとめ，機械学習アルゴリズムへの入力としている．日経 VI の 2σ 以上の上昇とその他の 2 値分類問題を設定し，ロジスティック回帰および，ランダムフォレストでモデルを作成している．結果として，ロジスティック回帰を用いたモデルにおいて，2014 年 11 月 17 日から 2016 年 6 月 29 日の 395 営業日で評価したところ，precision, recall ともに 0.45 の精度を得ている．上田らは諏訪ら [21] の手法を拡張し，Yahoo!ファイナンス掲示板を用いた新たな VI 上昇予測モデルを開発した．諏訪らのモデルと同様の検証期間で実験した結果，recall 精度の僅かな向上を確認している．しかし，言語モデルの違いで精度の違いが明確に表れておらず，ソーシャルメディアから言語モデルを用いて金融指標の予測に有効な特徴ベクトルを獲得できていないという問題がある．そこで我々は投稿文書から特徴ベクトルを獲得する新たな手法を提案し．上田らの精度との比較を行うことで提案手法の有効性を評価する．

3. 提案手法

3.1 VI 予測モデルの構築プロセス

VI 予測モデルについての概要を**図 1** に示す．Yahoo!ファイナンス掲示板の日経平均株価スレッドへの投稿から文書の特徴表現を抽出するために，投稿された文書を日別にまとめ，形態素解析により形態素に分割する．言語モデルを用いることで形態素に分割した文書から分散表現を獲得し，オートエンコーダによる次元削減を行う．獲得した特徴ベクトルに金融時系列データを加えて特徴量を作成し，機械学習モデルへの入力とする．最後に機械学習アルゴリズムを用いて先行研究と同様に投資リスクとなる VI 指数の大幅な上昇の予測を行う．

3.2 文書特徴ベクトルの獲得

3.2.1 投稿文書の収集

Yahoo!ファイナンス掲示板から投稿文書を取得する．VI が大阪証券取引所に上場している日経 225 オプションに対応しているため，本研究では日経平均株価スレッドへの投稿を分析対象とし，収集する．投稿文書から抽出したデータは，日付，時刻，投稿文書である．日経平均株価および VI のデータは JPX データクラウド[*1]から収集する．

3.2.2 言語モデルを用いた文書特徴ベクトルの獲得

日経平均株価スレッドへの投稿から言語モデルを用いて，文書の分散表現を獲得する．先行研究では Latent Dirichlet Allocation(LDA) トピックモデルや Doc2Vec，BERT を用いることで，投稿文書から文書分散表現の獲得を行なっている．LDA では単語の出現頻度のみを参照しているため，単語の語順を考慮できないという問題がある．Doc2Vec のアルゴリズムでは，単語ベクトルはテキストの異なる文書間にまたがる意味性を捉えるのに対し、文書ベクトルは同じ段落から生成された文脈語に対して学習されるため，文書ベクトルが局所的な意味性しか捉えられないという問題や，短い文書に対してオーバーフィッティングしやすいという問題 [22] がある．ソーシャルメディアでは 1 単語で構成されるような短い文書も存在するため，予測に有効な文書ベクトルを獲得する上で特に問題となる可能性がある．Kumar ら [7] はスペルミスやタイプミスが混在する文書に対しては BERT の性能が著しく低下することを示している．ソーシャルメディア文書は多くの文法誤りが存在するため，分散表現を獲得する手法として不適切であると考えられる．

そこで我々は SCDV を使用し，分散表現の獲得を行うことでこれらの問題を克服する．SCDV アルゴリズムでは文脈中の単語の各クラスタへの所属確率の違いによって獲得される分散表現が異なる．そのため単語埋め込み表現ベースでありながら実質的に文脈を考慮でき，ソーシャルメディア文書からでも有効な特徴ベクトルを獲得できると考える．本研究では埋め込み表現を用いたより単純な文書分散表現獲得手法である Simple Word-Embedding-based Model(SWEM) をベースラインとして実装し，各モデルとの比較を行う．SWEM とは word2vec により得られた文書中の単語埋め込みの各配列を単純に加算や平均して文書分散表現を獲得する手法である．

SCDV や SWEM により得られた分散表現はオートエンコーダを用いて 16 次元に次元削減を行い，文書特徴ベクトルを獲得する．

[*1] http://db- ec.jpx. co.jp/item/C430509.html

4. 実験

4.1 データセット

2012 年 11 月 26 日から 2020 年 09 月 30 日までに Yahoo! ファイナンス掲示板の日経平均株価スレッドに投稿された文書を収集した結果，9,463,597 件であった．日経平均株価および VI のデータは JPX データクラウドから収集した．収集期間は 2012 年 11 月 26 日から 2020 年 10 月 07 日 (1920 営業日) である．

投稿データの収集期間と日経平均株価および VI のデータの収集期間に若干の誤差があるのは最後の 5 営業日を正解ラベルの計算に使うためであり，実験期間はさらに最初の 5 営業日を除いた 2012 年 12 月 3 日から 2020 年 9 月 30 日 (1910 営業日) である．そのうち正例とされた日は 288 日であった．検証期間は，2016 年 12 月 14 日から 2020 年 9 月 30 日 (925 営業日) である. そのうち正例とされた日は 133 日で、検証期間の 14.37%にあたる．

4.2 文書の前処理

前処理として，投稿文書の URL，HTML，改行コードを除去した．さらに，半角ひらがなとカタカナを全角に変換し，全角英数字は半角に変換した．その後形態素解析を行なった．形態素解析には Mecab[23] を使用し，辞書には新規語や固有表現に強い NEologd[24] を使用した．なお、NEologd は 2020 年 5 月 21 日更新分を使用した．その後の前処理の方法として以下の二種類（前処理 A，前処理 B）を行なった．

前処理 A　各文書から，名詞, 動詞, 形容詞の中で Subtype が数値，非自立，代名詞，接尾でないものを抽出した．さらに京都大学が公開しているストップワードのリストに,「ある、する、ちゃう、ない、なる、やる」を追加し，これらをストップワードとして除去した．

前処理 B　各文書から記号を抽出し，除去した．

4.3 VI 予測モデルの構築

4.3.1 VI の大幅な上昇の定義とラベリング

投資リスクとなる VI の大幅な上昇を先行研究と同様に以下のように定義し，ラベリングを行なった．

投資リスクの定義

「今後 5 日営業日以内に日経 VI が閾値以上上昇する」

上記投資リスクの定義に当てはまることを正例としてデータにラベリングを行なった．ここで，閾値とは日経 VI の 1 日差分の 2σ とし，σ は次のように表される．

$$\sigma = \sqrt{\frac{1}{T-1}\sum_{k=0}^{T-1}(x_k - \bar{x})^2}$$

(T：実験期間中の営業日の日数，x：日経 VI の 1 日差分)

4.3.2 前処理された文書と言語モデルの組み合わせ

前処理された文書と言語モデルの組み合わせを**表 1** に示す．

表 1　前処理と言語モデルの組み合わせ

前処理 A_SCDV
前処理 B_SCDV
前処理 A_SWEM

本研究では SCDV アルゴリズムや SWEM アルゴリズムで必要となる単語埋め込みは東北大学：乾・岡崎研究室が公開している学習済み word2vec モデル（300 次元）[*2]を用いて取得した．さらに SCDV アルゴリズムでのソフトクラスタリング時のクラスタ数は 60 とした．

4.3.3 特徴量の作成

ソーシャルメディア投稿文書から獲得した特徴ベクトルと金融時系列データから先行研究と同様に次の特徴量を作成した．

- 投稿数
- 投稿数の 1 日差分，5 日差分
- 投稿数の 1 日比，5 日比
- 投稿数の 2 日間移動平均，5 日間移動平均
- 投稿数と投稿数の 2 日間移動平均の差，5 日間移動平均の差
- 投稿数と投稿数の 2 日間移動平均の比，5 日間移動平均の比
- 特徴ベクトル
- 特徴ベクトルの 1 日差分，5 日差分
- 特徴ベクトルの 1 日比，5 日比
- 特徴ベクトルの 2 日間移動平均，5 日間移動平均
- 特徴ベクトルと特徴ベクトルの 2 日間移動平均の差，5 日間移動平均の差
- 特徴ベクトルと特徴ベクトルの 2 日間移動平均の比，5 日間移動平均の比
- 日経平均株価の始値
- 日経平均株価の始値の 1 日差分，5 日差分
- 日経平均株価の始値の 1 日比，5 日比
- 日経平均株価の始値の 2 日間移動平均，5 日間移動平均
- 日経平均株価の始値と日経平均株価の始値の 2 日間移動平均の差，5 日間移動平均の差
- 日経平均株価の始値と日経平均株価の始値の 2 日間移動平均の比，5 日間移動平均の比
- VI の始値
- VI の始値の 1 日差分，5 日差分
- VI の始値の 1 日比，5 日比
- VI の始値の 2 日間移動平均，5 日間移動平均

[*2] https://github.com/singletongue/WikiEntVec

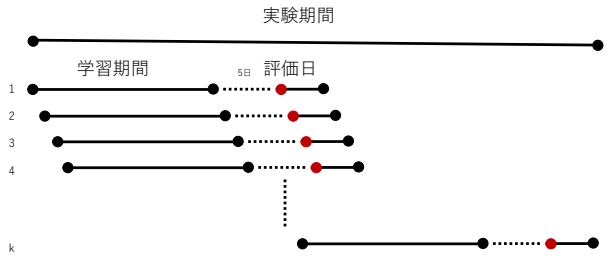

図 2　モデル構築の設計

- VI の始値と VI の始値の 2 日間移動平均の差，5 日間移動平均の差
- VI の始値と VI の始値の 2 日間移動平均の比，5 日間移動平均の比

4.3.4　機械学習手法

4.3.3 で作成した特徴量を機械学習アルゴリズムへの入力し，予測を行った．休日明けの営業日を予測する際は，前日のソーシャルメディアから作成した特徴量と前営業日の金融時系列データ特徴量を入力として予測を行なった．

先行研究では機械学習アルゴリズム，学習期間，ダウンサンプリングの有無などを変更した精度比較実験を行なっている．その中で精度が最も良い組み合わせである手法を本研究でも用いて先行研究との比較を行う．そのため，本研究では機械学習アルゴリズムとしてランダムフォレスト，学習期間は 2 年に固定する手法（**図 2**），正例と負例は 1:1 にダウンサンプリングして学習を行う手法を用いる．

4.3.5　予測モデルの評価法

実験期間 T 日のうち，学習期間を n 日目から m 日目までの $m-n+1$ 日間として学習する．本研究では学習期間を 2 年（490 営業日）と固定するため，$m-n+1=490$ が常にみたされる．評価には $m+5$ 日目の 1 日を使用する．n と m を 1 ずつ増加させ，学習，評価することを $m+5=T$ になるまで k 回繰り返す．最後に評価日の合計 k 回で評価を行う．（図 2）

5.　実験結果

モデルの評価指標として，正例に対する Precision, Recall，F1-measure を用いてこれらの値を比較する．金融時系列データ特徴量のみで学習させた金融時系列モデルでの精度，評価の結果を先行研究のモデルの精度と合わせて**表 2** に示す．比較に使用した先行研究のモデルは 3 種類で，一つは LDA により 128 次元の分散表現を獲得し，使用したモデル．一つは，Doc2Vec により 64 次元の分散表現を獲得し，使用したモデル．一つは，BERT により 768 次元の分散表現を獲得し，使用したモデルである．いずれも機械学習アルゴリズム，ダウンサンプリング，評価法は提案手法と同じである．

実験の結果，SCDV を使用したモデルでは前処理 A を行なったものが精度が良く，precision が 0.18，recall が 0.53，F1 が 0.27 という精度となった．SWEM を用いた手法は，presicion，recall，F1 の全てにおいて最も精度が高く，precision が 0.19，recall が 0.56，F1 が 0.28 という精度となった．

表 2　各モデルの予測結果

モデル	Precision	Recall	F1
金融時系列モデル	0.18	0.4	0.25
LDA_128_5:5	0.18	0.46	0.26
Doc2Vec_64_5:5	0.14	0.28	0.18
BERT_768_5:5	0.18	0.45	0.26
SCDV_前処理 A	0.18	0.53	0.27
SCDV_前処理 B	0.17	0.47	0.25
SWEM_前処理 A	0.19	0.56	0.28

6.　おわりに

本研究では金融指標の予測に有効な特徴量をソーシャルメディアから抽出する新たな手法を提案した．獲得した特徴量を使用してモデルを構築し，先行研究と精度の比較を行うことで提案手法の評価を行なった．比較の結果，先行研究の精度を上回る F1 が 0.27 と 0.28 のモデルが構築された．この結果は SCDV と SWEM がノイズの多い文書に対して LDA，Doc2Vec，BERT よりもパフォーマンスが良いことを示唆しており，提案手法は金融指標予測におけるソーシャルメディア文書の特徴表現抽出手法として新たな可能性を示した．さらに，ソーシャルメディアを用いた場合，金融時系列データのみを用いた予測精度よりも精度が高くなることが確認された．

参考文献

[1] Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, Vol. 25, No. 2, pp. 383–417, 1970.

[2] Yangtuo Peng and Hui Jiang. Leverage financial news to predict stock price movements using word embeddings and deep neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 374–379, San Diego, California, June 2016. Association for Computational Linguistics.

[3] Takashi MATSUBARA, Ryo AKITA, and Kuniaki UEHARA. Stock price prediction by deep neural generative model of news articles. *IEICE Transactions on Information and Systems*, Vol. E101.D, No. 4, pp. 901–908, 2018.

[4] Manuel R. Vargas, Beatriz S. L. P. de Lima, and Alexandre G. Evsukoff. Deep learning for stock market prediction from financial news articles. In *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, pp. 60–65, 2017.

[5] Stefan Feuerriegel, Antal Ratku, and Dirk Neumann. Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation. In *2016*

49th Hawaii International Conference on System Sciences (HICSS), pp. 1072–1081, 2016.

[6] Stefan Feuerriegel and Julius Gordon. Long-term stock index forecasting based on text mining of regulatory disclosures. *Decision Support Systems*, Vol. 112, pp. 88–97, 2018.

[7] Ankit Kumar, Piyush Makhija, and Anuj Gupta. Noisy text data: Achilles' heel of bert. *arXiv preprint arXiv:2003.12932*, 2020.

[8] Kentaro Ueda, Kodai Sasaki, Hirohiko Suwa, Yuki Ogawa, Eiichi Umehara, Tatsuo Yamashita, Kota Tsubouchi, and Keiichi Yasumoto. Prediction of nikkei vi increase for reducing investment risk using yahoo! japan stock bbs. *The 6th International Workshop on Application of Big Data for Computational Social Science (ABCSS2021)*, 2021.

[9] Dheeraj Mekala, Vivek Gupta, Bhargavi Paranjape, and Harish Karnick. SCDV : Sparse composite document vectors using soft clustering over distributional representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 659–669, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[10] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, p. 153–160, Cambridge, MA, USA, 2006. MIT Press.

[11] Kodai Sasaki, Yui Hirose, Eiichi Umehara, Hirohiko Suwa, Yuki Ogawa, Tatsuo Yamashita, and Kota Tsubouchi. Simulation of volatility trading using nikkei stock index option based on stock bulletin board. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 4367–4374, 2018.

[12] Kai Chen, Yi Zhou, and Fangyan Dai. A lstm-based method for stock returns prediction: A case study of china stock market. In *2015 IEEE International Conference on Big Data (Big Data)*, pp. 2823–2824, 2015.

[13] Wen Long, Zhichen Lu, and Lingxiao Cui. Deep learning-based feature engineering for stock price movement prediction. *Knowledge-Based Systems*, Vol. 164, pp. 163–173, 2019.

[14] Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. Enhancing stock movement prediction with adversarial training, 2019.

[15] Yumo Xu and Shay B. Cohen. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1970–1979, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[16] X. Li, H. Xie, Y. Song, S. Zhu, Q. Li, and F. Wang. Does summarization help stock prediction? a news impact analysis. *IEEE Intelligent Systems*, Vol. 30, No. 03, pp. 26–34, may 2015.

[17] Wei Li, Ruihan Bao, Keiko Harimoto, Deli Chen, Jingjing Xu, and Qi Su. Modeling the stock relation with graph network for overnight stock movement prediction. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 4541–4547. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Special Track on AI in FinTech.

[18] Chi Chen, Li Zhao, Jiang Bian, Chunxiao Xing, and Tie-Yan Liu. Investment behaviors can tell what inside: Exploring stock intrinsic properties for stock trend prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, p. 2376–2384, New York, NY, USA, 2019. Association for Computing Machinery.

[19] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, Vol. 2, No. 1, pp. 1–8, 2011.

[20] Jintao Liu, Hongfei Lin, Xikai Liu, Bo Xu, Yuqi Ren, Yufeng Diao, and Liang Yang. Transformer-based capsule network for stock movement prediction. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pp. 66–73, Macao, China, August 2019.

[21] Hirohiko Suwa, Yuki Ogawa, Eiichi Umehara, Kento Kakigi, Keiichi Yasumoto, Tatsuo Yamashita, and Kota Tsubouchi. Develop method to predict the increase in the nikkei vi index. In *2017 IEEE International Conference on Big Data (Big Data)*, pp. 3133–3138, 2017.

[22] Qingyao Ai, Liu Yang, Jiafeng Guo, and W. Bruce Croft. Analysis of the paragraph vector model for information retrieval. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, p. 133–142, New York, NY, USA, 2016. Association for Computing Machinery.

[23] T. KUDO. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.jp*, 2006.

[24] Sato Toshinori. Neologism dictionary based on the language resources on the web for mecab, 2015.