インフルエンサーのツイートを用いた暗号資産の価格変化予測

Prediction of Crypto-Asset Price using Influencer Tweets

山本 寛史[1*]　坂地 泰紀[1]　松島 裕康[1]　山下 雄己[2]
大澤 恭平[2]　和泉 潔[1]　島田 尚[1]
Hirofumi Yamamoto[1]　Hiroki Sakaji[1]　Hiroyasu Matsushima[1]　Yuki Yamashita[2]
Kyohei Osawa[2]　Kiyoshi Izumi[1]　Takashi Shimada[1]

[1] 東京大学
[1] The University of Tokyo
[2] 株式会社電通国際情報サービス
[2] Information Services International-Dentsu, Ltd.

Abstract: 現在，暗号資産は金融分野で注目を集めており，代表的な暗号資産の一つであるビットコインの 1 日の取引量は 5 億を超える．本研究では，ソーシャルネットワークサービス上で，強い影響力を持ったインフルエンサーと呼ばれる人々の暗号資産に関するツイートの影響に着目する．我々はインフルエンサーのツイートが暗号資産価格に影響すると考え，インフルエンサーのツイートを用いて，ビットコイン価格の上昇/下降を予測する手法を提案する．インフルエンサーのツイートを収集し，それを言語処理の手法を用いて特徴抽出し，機械学習に用いる素性を生成した．実験の結果，我々はインフルエンサーツイートが暗号資産の価格に影響する可能性があることを示唆した．

1　はじめに

暗号資産[1]は現在，金融の分野において多くの投資家や研究者の関心を集めている．Bitcoin(BTC)[Nakamoto 08] は，そのような暗号資産の 1 つで，1 日の取引量は 50 億を超える．暗号資産のために考案されたブロックチェーン技術は，他の分野への応用が期待されているため，エンジニアからの注目も集めている．このように様々な分野からの関心があり，ブロックチェーンテクノロジーに関するトピックについて，多くのツイートがされている．一方で，Twitter のデータ活用についても機関投資家から注目され始めている．2011 年に，Bollen らはつぶやきから得られた気分状態がダウ工業平均株価の予測に役立つことを示しており [Bollen 11]，彼らの研究は多くの研究者や投資家に大きな影響を与えた．我々は，この結果を参考にして，ツイートを使用して株価ではなく BTC 価格を予測することを目的としている．具体的には，ソーシャルネットワーキングサービス（SNS）においてユーザーに大きな影響を与えるインフルエンサーと呼ばれるユーザーの暗号資産に関するつぶやきと暗号資産の価格に着目した．我々はインフルエンサーのつぶやきは暗号資産の価格に影響を与えている，つまり，インフルエンサーのつぶやきは暗号資産に影響を与えると仮定している．この仮説を検証するために，インフルエンサーのつぶやきを使って BTC 価格を予測する実験を行った．Twitter データと BTC 価格を使用する方法の予測可能性が，BTC 価格のみを使用する方法よりも優れている場合，私たちの仮説が証明されると考えられる．

我々は，以下に示す方法でインフルエンサーのつぶやきを使用して BTC 価格が増減するかどうかを予測した; まず，インフルエンサーのつぶやきと BTC の価格のデータを収集する．次に，自然言語処理技術を使用してツイートから特徴を抽出し，機械学習に与える入力データとする．与えられた入力データに対して，BTC 価格の上昇/下降を分類する出力として学習させる．最後に，学習したモデルを用いてインフルエンサーのツイートから BTC 価格を予測する．

我々はこの実験においてニューラルネットワーク，サポートベクターマシン（SVM）[Hearst 98]，ランダムフォレスト [Breiman 01] など，いくつかの機械学習方法を採用し，仮説を確認するための実験を行った．

*連絡先： 東京大学
E-mail: yamamoto-hirofumi277@g.ecc.u-tokyo.ac.jp

[1]暗号資産はしばしば「暗号通貨」と呼ばれる．しかし，日本の金融庁は暗号通貨を「暗号資産」と呼ぶべきであると論じているので，本稿ではそれを暗号資産と呼ぶ．https://www.fsa.go.jp/news/30/singi/20181214.html

2　暗号資産価格予測手法

ここでは，本手法と機械学習方法のために選択した特徴量について説明する．本手法の新規性は，将来の暗号資産価格を予測するためにツイートデータと暗号資産価格の両方を使用することである．以下の節では，特徴量抽出とモデルについて説明する．

2.1　特徴量

最初に，英語と日本語それぞれのつぶやきの 0.5%以上に含まれる名詞，動詞，形容詞を選択し，それらの重要性を TFIDF によって評価した．$TFIDF(w,t)$ は次の式で計算される．

$$TFIDF(w,t) = tf(w,t)\log_2 \frac{N}{df(w)}, \qquad (1)$$

ここで，$tf(w,t)$ は，ツイート t 内の単語 w の頻度，N は，学習データセット内のツイートの数，$df(w)$ は，単語 w のツイート頻度をそれぞれ示している．

この実験には 2 種類の入力が選択された;(1) ツイートテキストの TFIDF データのみの入力，(2)TFIDF データにいいねの数とリツイート数，ツイートしたユーザのフォロワー数，および前日からの BTC の価格の変化率の 4 つをあわせた入力である．ただし，これらの入力は biLSTM には使用されず，代わりにツイートデータの word2vec[Mikolov 13] [2]（200 次元）が入力として使用される．

2.2　機械学習手法

この節では，実験で使用した機械学習方法を述べる．我々はツイートから BTC の価格変動を予測するために，ロジスティック回帰（LR），ランダムフォレスト（RF），および多層パーセプトロン（MLP）を使用した．LR と RF には Scikit-learn [3]で用意されているものを採用した．MLP と biLSTM は Pytorch [4]を使って実装した．MLP は，1 つの入力層，1 つの隠れ層，および 1 つの出力層で構成されており，活性化関数として双曲線正接関数（tanh）を採用した．各隠れ層は 100 単位で構成されている．biLSTM も，1 つの入力層，1 つの隠れ層，および 1 つの出力層で構成されている．

2.3　本手法

BTC の価格とツイートデータの関係をさらに調べるために，以下のモデルを提案する (Figure 1)．まず，前日からの BTC の価格変動が 1 %以上の場合，LR を用いてモデルの係数と切片を計算する．次に，切片と係数を使用して，結果が 1 になる確率を次の式を使用して計算する．

$$probability = \frac{1}{1 + e^{-\beta - \alpha \cdot v}} \qquad (2)$$

ここで，α は係数，β は切片，v は TFIDF と，LR の入力として使用されるその他のツイートデータをつなげたベクトルである．式 2 は，図 1 に示すように **Vectorize** と呼ばれる．RF 入力データは以下のように生成される．

Step 1: ツイートのテキストから TFIDF を計算する．

Step 2: ツイートの “いいね” の数とリツイート数，およびそのユーザーのフォロワー数を，計算した TFIDF につなげる．これは「ツイートデータ」と呼ばれる．

Step 3: Step 2 のデータによって LR を訓練し，その係数と切片を計算する．この訓練された LR は，Figure 1 における「Learned LR」である．

Step 4: ツイートごとに，LR の係数と切片，ツイートデータ（Formula 1）で BTC の価格が値上がりする確率を計算する．この確率を「LR スコア」とする．このスコアは，図 1 の入力ベクトルの「Vector B」部分のそれぞれの数である．

Step 5: 毎分，前日からの BTC 価格の変化率を計算する（これを「変化率」と呼ぶ）．これは Figure 1 の入力ベクトルの「Vector A」部分である．

Step 6: 1 分間にツイートされた 12 ツイートの LR スコアを，BTC の値段が翌日に 1%以上変動した時点から 1 時間前までの変動率につなげる．ツイートが足りない分は 0.5 に置き換える．これが RF の入力データであり，Figure 1 の入力ベクトルである．

3　評価実験

本節では，本手法を評価する．BTC の価格変動とインフルエンサーのつぶやきの関係を調べるために，次の実験を行った．

- 本手法による予測とツイートデータなしの予測の精度の違いを確かめる実験．
- 入力として word2vec を使った biLSTM による実験．
- 本手法と同じ条件の，他のアプローチによる実験．

[2] https://radimrehurek.com/gensim/
[3] http://scikit-learn.org/stable/
[4] https://pytorch.org

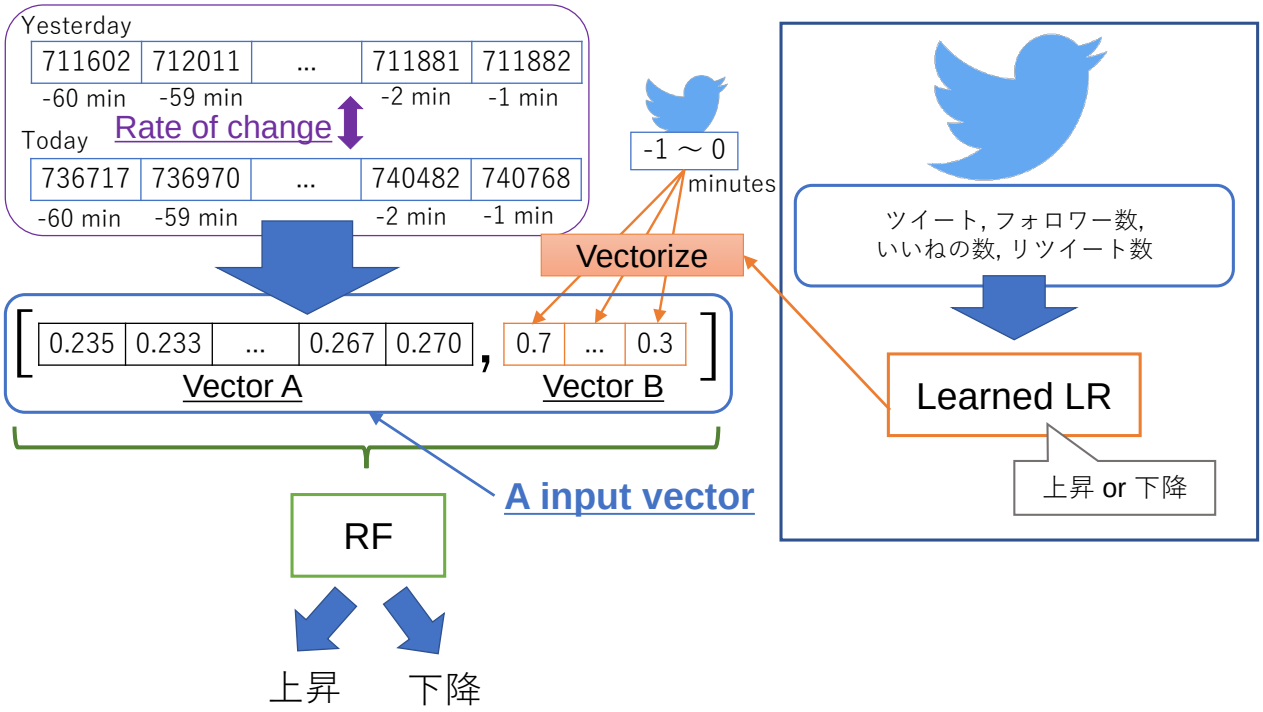

図 1: 本手法

表 1: ツイートの例

No.	Tweet
1	ＮＹ株大幅続落，市場関係者の見方
2	米インテル，振るわぬ株価と狭まる包囲網
3	南海電鉄，台風で浮かんだ訪日客銘柄の課題
4	ＧＭＯ，仮想通貨で再出発　ホリエモンの助言
5	株も先物も　総合取引所実現へ，政府が年内答申
6	ヤマトＨＤが「空飛ぶトラック」，米社と開発　人手不足解消へ無人輸送機
7	放置預金に注意，10 年で国が召し上げ　来年から

さらに，我々は BTC の価格だけでなく，取引量とツイートも関連している可能性があると仮定した．そのため，ツイートから取引量の変化を予測する実験も行った．

3.1 実験設定

実験を行う上での設定について説明する．

· **被説明変数:** 本研究では，予測対象である BTC の価格について非説明変数である出力値について大まかに 3 ケース用意した．まず，(1)BTC の価格が一定期間後に上昇または下降するかの予測である．1 日後，1 週間後，2 週間後という 3 つの期間を考えた．次に，(2)BTC の取引金額が一定期間後に増減するかどうかを予測することである．これについても，価格の場合と同様に 3 つの期間が考慮された．最後に，(3)BTC の価格が 1 日後に 1 %以上上昇したとき，または減少したときに限定し，価格が上昇または下降するかどうかを予測した．

· **入力変数 (ツイート) の種類の組み合わせ:** ツイートは次の 3 つの基準に従って分類された : (1) 言語（日本語または英語），(2)BTC または GEN（一般）としてタグ付けされているかどうか，(3) ツイート後に BTC の価格が 1 %以上変動したかどうか．ここで，ツイートに BTC 関連のトピックが含まれている場合，そのツイートは BTC とタグ付けされていた．さらに，ツイートに暗号資産関連の一般的なトピックが含まれている場合，それらは GEN とタグ付けされていた．これらの基準の組み合わせに基づいて，我々は 8 種類の入力を得た．

3.2 評価データ

ここでは，私たちの方法を評価するためのデータを生成する方法を示す．MeCab[5] を日本語の，spaCy[6] を英語の形態素解析のために使用した．RT で始まるツイートは，リツイートであるため削除した．評価には

[5] http://taku910.github.io/mecab/
[6] https://spacy.io/

2018 年 7 月 16 日から 10 月 24 日までのツイートデータを使用した．我々は，以下のようにしてインフルエンサーのツイートを収集した．まず，暗号資産に関連するツイートを頻繁に投稿する人々を調査し，彼らのフォロワーをたどった．そして，私たちはこれらの人々の中からインフルエンサーを選び，彼らのつぶやきを集めた．7 月 16 日から 9 月 24 日までのツイートデータを学習データとし，10 月 4 日から 10 月 24 日までをテストデータとした．

表 2 に我々のデータの内容を示す．

表 2: 入力に使われたツイートの数

	train data	test data
英語	25,986	14,542
日本語	70,400	23,581
英語 (BTC または GEN のタグ)	2,765	1,693
日本語 (BTC または GEN のタグ)	7,915	2,431
英語 (変化率 $>$ 0.01)	16,864	2,552
日本語 (変化率 $>$ 0.01)	47,864	4,305
英語 (BTC または GEN のタグ, 変化率 $>$ 0.01)	1,824	336
日本語 (BTC または GEN のタグ, 変化率 $>$ 0.01)	5,508	413

特徴のうち，TFIDF 値は，Scikit-learn によって正規化した．追加の特徴，すなわち，“いいね” およびリツイートの数，前日からの BTC 価格の変化率，およびそのユーザのフォロワーの数は，それぞれの絶対値の最大値で割り，最大値は 1 とした．

3.3 評価結果

表 3 は，本手法と，ツイートデータではなく BTC 価格のみを使った RF 予測の予測結果を示している．さらに，本手法と同じデータを使用し，LR，RF，および MLP 法を使用して BTC の価格変動を予測した．これらの実験の入力には翌日までの変化率が 1%以上の入力のみを使用した．

表 3: 予測の結果

	本手法	BTC 価格のみ	LR	RF	MLP
ACC	**0.666**	0.646	0.435	0.608	0.499
PRE	**0.665**	0.639	0.344	0.517	0.403
REC	**0.605**	0.598	0.440	0.435	0.499
F1	**0.629**	0.613	0.386	0.472	0.446

ここで，「ACC」は accuracy，「PRE」は precision，「REC」は recall，「F1」は f_measure をあらわす．これらの値は以下のように定義される．

	真の結果	
予測	TP	FP
結果	FN	TN

上記のように定義したとき，

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$
$$PRE = \frac{TP}{TP + FP}$$
$$REC = \frac{TP}{TP + FN}$$
$$F1 = \frac{2 \times PRE \times REC}{PRE + REC}.$$

また，word2vec を入力とし，biLSTM による学習結果は表 4 のとおりである．

表 4: biLSTM の結果

	英語 word2vec			日本語 word2vec		
	1 日	1 週間	2 週間	1 日	1 週間	2 週間
PRE	0.515	0.120	0.424	**0.521**	0.157	0.411
REC	**0.544**	0.398	0.480	0.516	0.531	0.499
F1	**0.529**	0.184	0.450	0.518	0.242	0.451

また，LR，RF，MLP を用いて 1%以上の変化率で学習した結果は，図 2 および図 3 のとおりである．入力データとして，ツイートのテキストに加えて，いいねとリツイートの数，前日からの BTC 価格の変化率，およびそのフォロワーのフォロワー数が使用された．英語と日本語のつぶやきでは，BTC または GEN とタグ付けされたつぶやきとすべてのつぶやきの 2 種類の入力が使用された．

最後に，ツイートデータから取引量を予測できるかどうかを検証した．入力データとして，英語と日本語のすべてのツイートの TFIDF データ，およびその他の 4 つの特徴量（いいねとリツイートの数，そのユーザーのフォロワー数，および前日からの BTC 価格の変化率） が使われた．

4 考察

表 3 は，本手法が他のメソッドよりも優れていることを示している．ここでは，本手法のパフォーマンスが「BTC の価格のデータのみ」を使うときのパフォーマンスより優れているという事実に焦点を当てる．

この結果は，我々の実験設定においてはインフルエンサーのつぶやきが暗号資産価格に影響を与えることを示している．

我々は他の機械学習法によっても価格を予測しようとしたが，これらの方法では図 3 に示すように，予測が

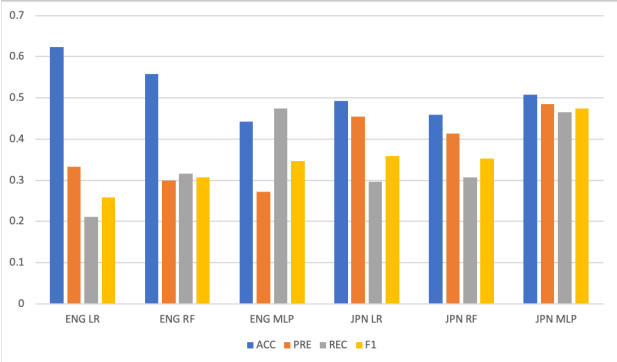

図 2: 価格の変化，変化率 > 0.01

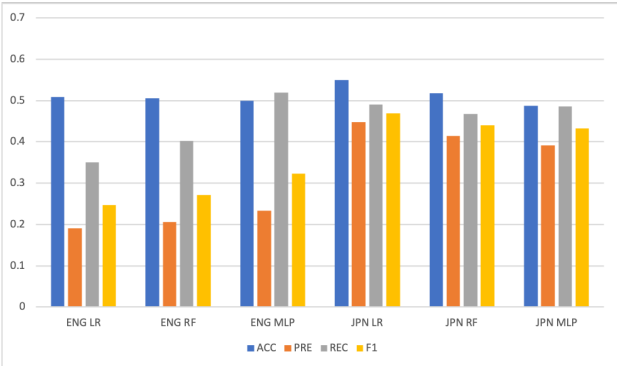

図 3: 価格の変化，変化率 > 0.01, BTC または GEN とタグづけされているツイートのみ

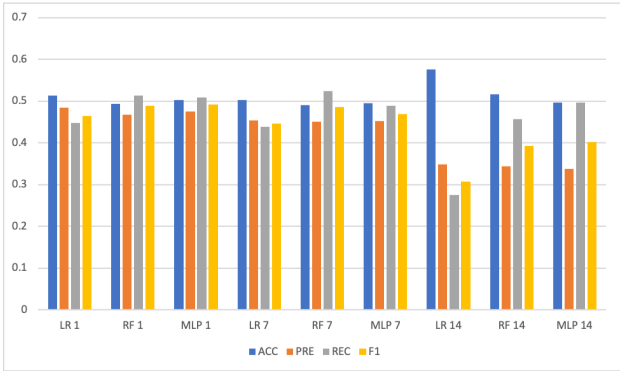

図 4: 取引量の変化, 入力: 英語

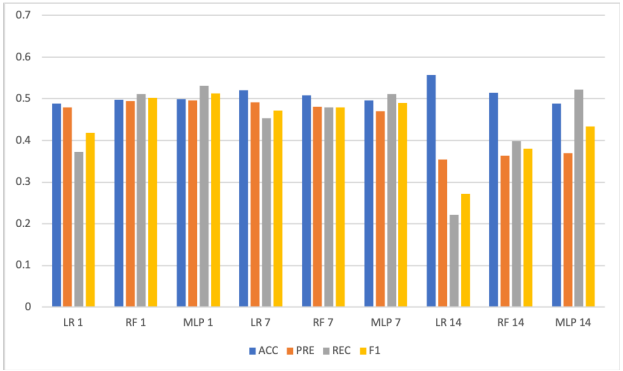

図 5: 取引量の変化, 入力: 日本語

可能であるという結果は得られなかった．結果として，この結果は，私たちの方法がツイートデータと暗号資産価格を効果的に使用できることを示している．すなわち，我々の方法は，暗号資産価格を予測するためにツイートデータを効果的に利用することができる．

図 2 と図 3 から，予測にタグ付きツイートを使用するとより良い結果が得られることがわかる．さらに，これらの結果では，日本語のつぶやきによる LR は他の方法よりも優れている．そのため，我々は LR を使用して手法を開発することにした．

図 4 と図 5 は，数量予測が言語に依存しないことを確認している．この結果から，ツイートを使って量を予測することは困難であることがわかる．そこで，ツイートを使って BTC の価格を予測する方法を開発した．

5 関連研究

Bollen らは，ツイートの気分がダウ工業平均株価の予測に役立つことを示した [Bollen 11]．彼らの研究では，予測に自己組織化ファジィニューラルネットワークを使用し，その結果，80%以上の精度で上昇と下降を予測することができた．Schumaker らは，金融ニュース記事分析を用いて株価を予測するための機械学習アプローチを提案した [Schumaker 09]．彼らの研究は指標と株価を予測しているが，暗号資産価格は予測されていない．

金融テキストマイニングについては，Koppel らが会社の株価のパフォーマンスへの形に見える影響に従って会社のニュース記事を分類するための方法を提案した [Koppel 06]．Low らは移動概念を表す用語を抽出するためのシソーラスとして WordNet[Fellbaum 98] を使用する，semantic expectation-based knowledge extraction methodology（SEKE） [Low 01] を提案した．Ito らは金融テキストデータを視覚化するためのニューラルネットワークモデルを提案した [Ito 18]．さらに，彼らのニューラルネットワークモデルは単語感情とそのカテゴリを獲得することができる．Milea らは欧州中央銀行が発行した報告書から抽出したファジーな grammar fragments に基づいて MSCI ユーロ指数（上方，下方，または一定）を予測した [Milea 10]．

日本語の金融テキストマイニングに関しては，Sakai らが日本の業績に関する金融記事から業績要因を抽出する方法を提案した [Sakai 07]．彼らの方法は業績要因を抽出するための手がかりを使用しており，ブートストラップ法を使用して手がかりを自動的に集めること

ができる．Sakaji らは統計的手法を用いて新聞記事から経済動向を示す根拠表現を自動的に抽出する方法を提案した [Sakaji 08]．Kitamori らは決算短信から業績予想と経済見通しを示す文を抽出して分類する方法を提案した [Kitamori 17]．分類方法は，半教師付きアプローチを使用することによるニューラルネットワークに基づいている．

これらの金融テキストマイニング研究は 1 つの言語のみを対象としている．対照的に，私たちの方法は複数の言語のつぶやきを使いる．

6 まとめ

本稿では，BTC 価格を予測する方法を提案した．この方法は，機械学習方法として RF および LR を使用し，特徴としてツイートおよび暗号資産価格を使用する．特に，本手法の新規性は，RF の入力のために，LR の切片と係数を使用することである．この方法により，インフルエンサーのつぶやきが暗号資産に影響を与えることが示された．さらに，我々は 2 つの言語でツイートを使用して実験し，暗号資産の量を予測することも試みた．

私たちの方法では，1%以上変化した場合のみを考えている． したがって，将来の研究として，この制限なしに暗号資産価格を予測する方法を見つけることが挙げられる．さらに，私たちの方法を使って仮想取引をシミュレートする事も考えられる．

参考文献

[Bollen 11] Bollen, J., Mao, H., and Zeng, X.: Twitter mood predicts the stock market, *Journal of computational science*, Vol. 2, No. 1, pp. 1–8 (2011)

[Breiman 01] Breiman, L.: Random Forest, *Machine Learning*, Vol. 45, No. 1, pp. 5–32 (2001)

[Fellbaum 98] Fellbaum, C.: *WordNet: An Electronic Lexical Database*, The MIT Press (1998)

[Hearst 98] Hearst, M., Dumais, S., Osuna, E., Platt, J., and Scholkopf, B.: Support Vector Machines, *IEEE Intelligent Systems and their applications*, Vol. 13, No. 4, pp. 18–28 (1998)

[Ito 18] Ito, T., Sakaji, H., Izumi, K., Tsubouchi, K., and Yamashita, T.: GINN: gradient interpretable neural networks for visualizing financial texts, *International Journal of Data Science and Analytics* (2018)

[Kitamori 17] Kitamori, S., Sakai, H., and Sakaji, H.: Extraction of sentences concerning business performance forecast and economic forecast from summaries of financial statements by deep learning, in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–7 (2017)

[Koppel 06] Koppel, M. and Shtrimberg, I.: *Good News or Bad News? Let the Market Decide*, pp. 297–301, Springer Netherlands, Dordrecht (2006)

[Low 01] Low, B.-T., Chan, K., Choi, L.-L., Chin, M.-Y., and Lay, S.-L.: Semantic expectation-based causation knowledge extraction: A study on Hong Kong stock movement analysis, in *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pp. 114–123 (2001)

[Mikolov 13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed representations of words and phrases and their compositionality, in *Advances in neural information processing systems*, pp. 3111–3119 (2013)

[Milea 10] Milea, V., Sharef, N. M., Almeida, R. J., Kaymak, U., and Frasincar, F.: Prediction of the MSCI EURO index based on fuzzy grammar fragments extracted from European Central Bank statements, in *2010 International Conference of Soft Computing and Pattern Recognition*, pp. 231–236 (2010)

[Nakamoto 08] Nakamoto, S.: Bitcoin: A peer-to-peer electronic cash system (2008)

[Sakai 07] Sakai, H. and Masuyama, S.: Extraction of Cause Information from Newspaper Articles Concerning Business Performance, in *Proc. of the 4th IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI2007)*, pp. 205–212 (2007)

[Sakaji 08] Sakaji, H., Sakai, H., and Masuyama, S.: Automatic Extraction of Basis Expressions That Indicate Economic Trends, in *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pp. 977–984 (2008)

[Schumaker 09] Schumaker, R. P. and Chen, H.: Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System, *ACM Trans. Inf. Syst.*, Vol. 27, No. 2, pp. 12:1–12:19 (2009)