

ソーシャルメディア特有の言語情報に基づく 機械学習による市場予測の補助

関岡 冨俊[†] 秦野 亮[†] 西山 裕之[†]

東京理科大学 理工学部[†]

1 背景・目的

近年、ソーシャルメディア上に投稿される不特定多数の文章が株や仮想通貨といったオンラインで取引される商材の価格変動に大きな影響を与えている。そのため、ソーシャルメディアは、現実世界の出来事を観察・予測するために注目されている。その顕著な例として、2021年5月14日に起業家イーロンマスクがTwitter上でツイートをした直後に仮想通貨の価格が大きく変動した。この原因としては、暗号通貨は需要と供給が投資家のセンチメントに影響されやすいため、ソーシャルメディアでのやりとりが暗号通貨の価格に大きな影響を与えるからだとと言える。そこで、ソーシャルメディアが現実世界の事象を動かすという仮説のもと、ソーシャルメディアの言語表現の情報が市場予測の有用な補助となることが期待出来ると考えた。

本研究では、Twitter上の情報やツイート内の自然言語表現が仮想通貨の価格変動といった市場情報の予測における特徴量になり得るかの調査目的として、ソーシャルメディアの言語表現から Sentence-BERT を使い作成した特徴量を使用した場合としなかった場合とで、価格の急変の予測を行う手法を比較する。

2 関連研究

Caron[1]らは、10-K という会計年次報告書の MD&A セクションのテキスト情報から BERT ベースの学習器を使い、株式リターンのボラティリティの回帰予測を行い平均絶対誤差(MAE)を使用して評価したというものである。この研究は明確に定義された意味を持つ数値データであるハード情報と企業報告書、プレスリリース、ニュース記事などの文章で表現されたデータなどであるソフト情報を用いて予測をおこなっている。

Market prediction aids using machine learning based on social media specific linguistic information

[†] Tokyo University of Science, Faculty of Science and Engineering

結果としては、BERT ベースの学習器を使用した予測(ソフト情報)とヒストリカルボラティリティ(ハード情報)の平均による学習器のアンサンブルによる予測の性能が一番高かった。また、学習で使用するデータを近年のものに限った場合の性能が高かった。

Ohana ら[2]は、150 以上のテクニカル、ファンダメンタル、マクロ経済のセットから LightGBM によるアプローチによって、S&P500 の大きな値下がり予測した。その結果、SHAP (SHapley Additive exPlanation) 値によって、株式市場の危機を予測する最も重要な変数をロバストに特定することができ、これにより、各日付に対する危機確率を説明可能にすることがわかった。また、2020年3月のS&P500金融暴落を20年2月3日には、急暴落の確率を27%に上昇させ、20年3月2日のCovidによる急暴落の開始時には、急暴落の確率を61%と急上昇させ予測することができた。本研究では、ソーシャルメディアから情報を収集し、特徴量を作成して、仮想通貨の終値の急変の予測を行う。

3 提案手法

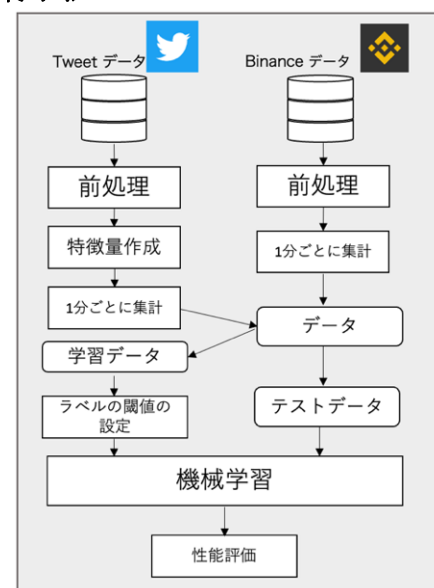


図1. 提案手法の概要

提案手法の概要を図1に示した。手順は以下の通りである。まずTwitter APIとBinance APIをそれぞれ使用して、BTC/USDTにまつわるTweetデータと1分刻みのタイムバーの情報を取得する。その後、タイムバーのデータには終値の差分をつけ、学習で使うデータのみから一定の割合を急変のラベルを振るための閾値を得て全体にラベル付けを行う(終値の差分のうち75%以上のものを急上昇, 25%以下のものを急降下, それ以外を普通とする)。また、Tweetデータを整形し、Sentence-BERTを使用し768次元の分散表現を獲得する。これらの自然言語データをタイムバーの時間の形にあるように1分間ごとに集計する。その後、LightGBMにより、終値の急上昇, 普通, 急降下の三クラス分類を行いソーシャルメディアに基づいた特徴量の有無で結果を比較する。

3 実験

データの収集期間は、2021年6月29日から2021年1月6日のうちでTweetデータがあるもののみを実験の対象とした(185511個)。学習データの割合を7割, 8割, 9割と変化させながら実験を行った結果が以下の表1である。

表1. 実験結果

	ベースラインの正答率	正答率	macro-F1
Binance データ + Tweet データ 学習データ: 2021/6/29 - 2021/11/9 テストデータ: 2021/11/9 - 2022/1/3	0.489	0.500	0.283
Binance データのみ 学習データ: 2021/6/29 - 2021/11/9 テストデータ: 2021/11/9 - 2022/1/3	0.489	0.500	0.279
Binance データ + Tweet データ 学習データ: 2021/6/29 - 2021/11/27 テストデータ: 2021/11/27 - 2022/1/3	0.492	0.500	0.293
Binance データのみ 学習データ: 2021/6/29 - 2021/11/27 テストデータ: 2021/11/27 - 2022/1/3	0.492	0.500	0.287
Binance データ + Tweet データ 学習データ: 2021/6/29 - 2021/12/15 テストデータ: 2021/12/15 - 2022/1/3	0.568	0.569	0.296
Binance データのみ 学習データ: 2021/6/29 - 2021/12/15 テストデータ: 2021/12/15 - 2022/1/3	0.598	0.569	0.254

表1のベースラインとは全て普通ラベルとした場合の正答率である。表1からわかることとして、Binance データのみで予測を行った時よりもTweet データに基づく特徴量を使用した場合の方が全体的に macro-F1 が大きくなっており、最大で 4.2%大きくなっている。また学習データの割合が大きくなると Binance データのみの場合よりも Tweet データを使用した場合の方が予測精度の向上が大きくなっている。次に Tweet データを使

用することで macro-F1 が 4.2%大きくなった期間の各ラベルに対する precision と recall を表2, 3に記す。

表2. Binance データと Tweet データによる予測結果

クラス	precision	recall	F1 スコア
急降下	0.354	0.050	0.087
普通	0.581	0.969	0.726
急上昇	0.383	0.040	0.073
正答率			0.569
macro 平均	0.439	0.353	0.296

表3. Binance データのみの予測結果

クラス	precision	recall	F1 スコア
急降下	0.300	0.001	0.003
普通	0.571	0.996	0.725
急上昇	0.462	0.018	0.035
正答率			0.569
macro 平均	0.444	0.338	0.254

表2, 3を見比べると Tweet データを使用した場合の方が急上昇と急降下のラベルの recall が大きいため、取りこぼしが少なくなっていることがわかる。しかし、precision を確認すると急上昇のラベルに対して Tweet データを使用した場合の方が小さくなっている。そのため、この場合 Tweet データを使用した場合、急上昇のラベルの誤検出が多くなっているといえる。

4 今後の課題

本研究ではTwitter APIの仕様上の制限のため、短期間のデータセットしか作成できなかった。そのため局所的なデータセットに対しての学習・予測による検証しか行えなかった。その他にもTwitterから収集する際の単語の選別を行う必要があると考えられる。また、実運用を考慮するとタイムバーの単位を大きくしたり、特徴量エンジニアリングを行う、より強力な学習器の使用が考えられる。

5 参考文献

- [1] Matthew Caron and Oliver Müller. Hardening soft information: A transformer-based approach to forecasting stock return volatility. In 2020 IEEE International Conference on Big Data (Big Data), pages 4383–4391. IEEE, 2020.
- [2] Jean Jacques Ohana, Steve Ohana, Eric Benhamou, David Saltiel, and Beatrice Guez. Explainable AI (XAI) models applied to the multi-agent environment of financial markets. In International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, pages 189–207. Springer, 2021.