

ドメイン特化疑似データを用いた Xの 感情分析による日経平均株価騰落予測の精度比較

今本 竜樹¹¹ 明治大学大学院 先端数理科学研究科
cs2733233@meiji.ac.jp櫻井 義尚²² 明治大学 総合数理学部
sakuraiy@meiji.ac.jp

概要

近年、株価予測の分野では、SNS 文書の感情分析を通じて市場の感情を推定し、予測精度を向上させる研究が注目されている。しかし、SNS 特有の主観的な感情表現と金融特有の専門用語が混在する金融 SNS ドメインに適応したアノテーション済みデータの不足が課題となっている。本稿では、汎用、SNS、金融、金融 SNS といった4つのドメインにおける感情分析の精度比較とターゲットドメイン特化型学習の評価を行った。また、疑似データ生成を活用し、ボジネガ分析を用いた株価予測実験において、予測精度の向上を実現した。本稿は、金融市場において有効性を示し、他の複雑なドメインにおいても適応可能性を持つものである。

1 はじめに

株価予測は、経済活動や市場の動向を予測し、その精度向上に向けた研究が長年にわたり進められてきた。近年、投資家心理が株価の動きに与える影響が認識されており、ニュース記事や SNS データを感情分析を通じて株価予測の精度を向上させる研究が行われている。

感情分析はテキストデータから投資家や市場参加者の感情や意見を抽出する上で重要な役割を担っている。感情分析手法は、大きく辞書型手法と言語モデル手法に分類される。辞書型手法は、あらかじめ定義された辞書に基づき単語の感情的な意味を判断する。この手法はシンプルで解釈が容易である一方、文脈を考慮しないため、Twitter のように文脈依存性が強い顕著な短文データに対しては分析精度が低下する。一方、BERT や GPT のような事前学習言語モデルは、文脈を考慮した単語のベクトル化が可能であり、柔軟かつ高精度な感情分析を実現する。

しかしながら、これらの高度なモデルにも課題が存在する。GPT のような大規模言語モデルは高い解析性能を持つが、その推論プロセスにおける計算コストが高く、何十万以上ある膨大な SNS データを全て処理することは現実的ではない。また、BERT のような比較的軽量なモデルは、特定のドメインに対応するために、対象ドメインのデータを用いたファインチューニングが必要であり、学習データの特性が性能に大きな影響を与えるため、精度の高いモデルの構築には対象ドメインの教師データの整備が不可欠である。

SNS を用いた、株価予測に必要な教師データは、金融ドメインと SNS ドメインの特性を持つ独自の領域であり、「金融 SNS ドメイン」と呼ばれる。このドメインでは、金融ドメインに見られる専門的な用語の理解が求められる一方で、SNS ドメイン特有の主観的で非構造的な言語的特徴を持つ。こうした複雑なドメインのデータであるにも関わらず、日本語の金融 SNS ドメインテキストに感情値がアノテーションされた公開言語コーパスは整備されていない。そのため、SNS を用いた株価予測における既存の研究では、金融 SNS ドメインではなく金融ドメインの言語コーパスを用いて感情分析が行われており、SNS の投稿内容の感情分析精度低下の一因となっている。

そこで、本研究では、関連するドメインのコーパスを活用したり、生成した疑似データを活用したりすることにより、金融 SNS ドメインの感情分析精度の向上を試みる。まず、関連する金融ドメイン、SNS ドメインのコーパスが学習データとしてどの程度寄与するのかを検証することで有用なコーパスを明らかにする。次に GPT を用いて金融 SNS ドメインの疑似データを生成し、これを学習データとした感情分析を用いて株価予測を行う。この予測精度を評価す

ることにより擬似データを用いた感情分析の有効性を明らかにする。

2 関連研究

SNS を用いた株価予測に関連する研究として、いくつかの先行研究が存在する。Bollen らは[1]、Twitter テキストデータを感情分析することで、2~5 日後のダウ平均株価の騰落を 86.7%の精度で予測した。日本国内においては、日経平均株価とツイートの研究が行われている。山野ら[2]は日経平均株価と「経済ワード」を含むツイート感情との相関を示し、ニュース記事よりもツイートの方が株価との因果関係が高い可能性を明らかにしている。さらに、片山・高橋[3]は小型株に関連する日本語ツイートを対象に感情分析を行い、その結果が株価や出来高に与える影響を定量的に評価した。しかし、同研究では Googl. Colab の Natural. Language. API を用いた感情分析モデルを使用しており、金融 SNS 特有の専門用語や文脈を十分に反映した解析が行われていない可能性が指摘される。

これらの研究は、SNS データの株価予測における有用性を示しているものの、多くの研究が学習データセット内での性能評価に終始しており、学習データのドメインについては十分に検討されていない点が課題である。また、日本語タスクにおいては、藤井らの研究[4]が、大規模言語モデルを用いた疑似データ生成で学習した小規模モデルが、特定の条件下で LLM 本体を超える精度を達成できることを示しており、特化型モデルの可能性を示唆している。

3 感情分析ドメイン依存性評価

本章ではアノテーション済みのデータを使用して、2つの実験を行い、ドメインによる感情分析について分析する。

3.1 アノテーション済み感情データ

以下で使用した各ドメインにおける感情値付きデータについて説明する。

金融 SNS ドメインに近いデータとして FinGPT Sentimen. Dataset[5]を使用した。このデータは金融関連のニュースやツイートのテキストに対してポジティブ、ネガティブ、ニュートラルが付与されたデータセットである。この中でツイートのみを使用する。なお、日本語のアノテーション済みの金融 SNS ドメインでの感情データが存在しないため、英語の

データセットを翻訳して使用する。金融ドメインの専門性を反映しつつ、SNS ドメインに特有の主観的な言語表現を含む点で、翻訳により一部のニュアンスが変化する可能性を考慮する必要はあるものの、金融 SNS ドメインの分析の目的に十分に適合する。

汎用ドメインとして Japanes. Realisti. Textua. Entailmen. Corpus(以下 JRTE コーパス)[6]を使用した。JRTE コーパスは株式会社リクルートが旅行情報サイト「じゃらん net」で公開されている宿泊施設への利用者によるクチコミが収録されており、ポジティブ、ネガティブ、ニュートラルのアノテーションがされているデータである。文がどの判定ラベルに対応するかはアノテーションを行った . 名による多数決によって決定している。

金融ドメインとしては chABSA-Dataset[8]を使用した。chABSA-dataset は TI. 株式会社によって公開されたデータセットであり、2016 年度の上場企業の有価証券報告書をベースに作成されている。各文に含まれるアスペクトに対してポジティブ、ネガティブニュートラルの感情ラベルが付与されている。ポジティブなものが含まれるものは+1、ネガティブは-1、ニュートラルは0の加算平均を文の感情値として使用した。

SNS ドメインとしては WRIME[7]を使用した。WRIM. は、SN. に投稿された文章に対して感情極性(強いネガティブ、ネガティブ、ニュートラル、ポジティブ、ジティブ)が付与されている。本稿では感情の強度は考慮せず、ポジティブ、ネガティブ、ニュートラルの3クラスに変換して使用した。

以下の表に、使用したデータセットとそれぞれの対応ドメインを示す。

表. ドメインと使用データ

| ドメイン | 使用データ |
|-------------|--------------------------|
| 汎用ドメイン | JRTE コーパス |
| SNS ドメイン | WRIME |
| 金融ドメイン | chABSA-dataset |
| 金融 SNS ドメイン | FinGP. Sentimen. Dataset |

3.2 実験設定

感情分析モデルには東北大学乾・鈴木研究室が公開している事前学習済み日本語 BER. (cl-tohoku/bert-base-japanese)を使用した。感情分類タスクは、各文書を「ポジティブ」「ネガティブ」「ニュートラル」の3つのカテゴリに分類する多クラス

分類問題として設定した。この際、学習が偏らないように、層化分割をおこなった。

3. 3 実験 1：同ドメインによるテスト検証

実験 1 では学習データとテストデータを同じドメインとした。汎用ドメイン、SNS ドメイン、金融ドメイン、金融 SNS ドメインの 4 つのデータセットで感情分析の精度を比較した。各ドメインで 1, 113 件ずつのデータを用いて 9 割を学習データ、1 割を評価データとして使用した。この際、学習が偏らないように、層化分割をおこなった。結果は表 2 の通りになった。

表 2 実験 1：同ドメインによるテスト検証結果

| | Accuracy | Precision | Recall | F1-score |
|--------|----------|-----------|--------|----------|
| 汎用 | 0. 850 | 0. 839 | 0. 851 | 0. 845 |
| 金融 | 0. 704 | 0. 718 | 0. 696 | 0. 707 |
| SNS | 0. 510 | 0. 486 | 0. 495 | 0. 490 |
| 金融 SNS | 0. 679 | 0. 666 | 0. 614 | 0. 639 |

3. 4 実験 2：SNS 金融テストデータ検証

実験 2 ではテストデータを SNS 金融ドメインとし、学習データを異なるドメインのデータとした時、どの程度有用なのかを検証した。各ドメインで 1, 002 件ずつのデータを用いて学習を行った。評価データは金融 SNS ドメインの 111 個のデータを使用した。金融 SNS ドメインの感情分析を行った。結果は表 3 のようになった。

表 3 実験 2：SNS 金融テストデータ検証結果

| | Accuracy | Precision | Recall | F1-score |
|--------|----------|-----------|--------|----------|
| 汎用 | 0.436 | 0.475 | 0.436 | 0.454 |
| 金融 | 0.471 | 0.510 | 0.473 | 0.491 |
| SNS | 0.437 | 0.499 | 0.437 | 0.467 |
| 金融 SNS | 0.679 | 0.666 | 0.614 | 0.639 |

3. 5 考察

実験 1 の結果から、以下のドメインの特性が分かる。まず、汎用ドメインと金融文書ドメインを比較すると、金融文書データには専門用語や形式的な文体が含まれるため、汎用ドメインよりも精度が低下している。一方、金融文書ドメインと金融 SNS ドメインの比較では、金融 SNS ドメインに口語的な表現や非公式な文体が含まれるため、精度がさらに低下

していることが確認された。SNS ドメインと金融 SNS ドメインを比較すると、金融 SNS ドメインの方が、専門的な内容を含みつつも口語的な表現が抑制されている傾向があるため、SNS ドメインよりも高い精度を示した。また、SNS ドメインと金融ドメインを比較した場合、SNS データは主観的かつ感情的な表現が多く、文脈が曖昧であるため、金融文書データのように形式的かつ一貫した内容を持つデータに比べて感情分析モデルの性能が大幅に低下することが明らかとなった。これらの結果から、金融文書ドメイン > 金融 SNS ドメイン > SNS ドメインの順で感情分析モデルにとって分析が容易である。

次に実験 2 の結果からターゲットドメインである金融 SNS データの感情分析精度は、学習に使用したデータのドメインとターゲットドメインの一致度に大きく依存することが示された。具体的には、金融 SNS データを学習ドメインとして用いた場合に最も高い精度を達成し、ターゲットドメインとの適合性が感情分析モデルの性能に大きく寄与することが確認された。一方、汎用ドメインや SNS データを学習ドメインとして用いた場合、ターゲットドメインとは文脈や語彙の特性が異なるため、精度が低下した。この結果は、事前段階で学習データが適切な精度を示していても、実際のターゲットドメインと一致しない場合には、性能が大幅に低下する可能性を示唆している。したがって、ターゲットドメインに適したデータを用いてモデルを学習させることが、感情分析タスクにおいて重要であることが明らかとなった。

4 提案手法

本章では、提案手法である疑似データを用いた感情分析を通じた株価予測手法について説明する。

4. 1 実験の流れ

以下に予測の流れについて説明する。

1. Twitte. API を用いてツイートを集集する。
2. GPT を用いて疑似データを作成する。
3. BERT を疑似データを用いて学習を行う。
4. ツイートに対してポジネガ分析を行う。
5. 日ごとにポジティブとネガティブをとり時系列データを作成する。
5. LST. を用いて、時系列予測のための深層学習モデルを構築する。

6. 株価の騰落データとツイート感情値の時系列データを予測モデルに組み込み、騰落予測を行う。

4. 2 疑似データ生成手法

疑似データを作成するための LL. には GPT-4 (Generativ. Pre-traine. Transformer-4) を用いる。疑似データ生成においては、感情ラベル (ポジティブ, ネガティブ, ニュートラル) と、日経シソーラスの証券[BI]中分野に含まれる上位単語の全ての組み合わせを網羅的に生成するよう設計した。具体的には、各感情ラベルごとに、word リスト内の単語を順に組み合わせたプロンプトで生成し、これにより全ての感情ラベルと全ての word について均等に条件を満たすデータを作成する。以下に、本稿で用いたプロンプトを示す

「以下の条件に基づいて、SNS 投稿の疑似データを 10 個生成してください。

条件:

1. 各投稿は日本語で記述してください。
2. 各投稿には「word」という単語を必ず含めてください。
3. 投稿は親しみやすく、現実の SNS 投稿らしい文面で記述してください。
4. 各投稿は 140 文字以内で、簡潔にしてください。
5. 投稿は[感情ラベル]の感情にしてください。
6. 金融ドメインに深い理解があるものとします。」

5 疑似データを用いた株価予測評価

本章では、各ドメインデータと疑似データを用いて学習を行ったモデルで感情分析を行い、日経平均株価予測の精度比較を行う。

5. 1 データ

株価データは日経平均株価の終値を使用する。Yahoo. Finance から 2017/9/1~2018/11/30 で取得し、終値を使用した。

ツイートデータは株価と同じ期間で「日経平均」または日経シソーラスの証券[BI]中分野に含まれる 37 単語のいずれかを含むツイートを収集した。

5. 2 モデル設定

感情分析モデルには、3 章で使用したモデルを採用し各ラベル (ポジティブ, ネガティブ, ニュートラル) が均等に分布するデータセット(計 1113 デー

タ)を用いてファインチューニングを行った。予測モデルの入力として感情スコアと株価終値の 2 つの特徴量を用い、5 日間の時系列データをウィンドウサイズとして設定した。

予測モデルの構造は、2 層の LSTM (各層 128 ユニット) および 1 層の Dense 層 (出力 1 ユニット) からなる深層学習モデルとし、最適化アルゴリズムには Adam を、損失関数には平均二乗誤差 (MSE) を採用した。学習は 20 エポックで行い、過学習を防ぐためにアーリーストッピングを導入した。

5. 3 株価予測実験結果

2017/9/1~2018/3/30 で学習を行い、2018/4/2~2018/11/30 でテストを行った。各データを用いた株価予測の評価結果は表 4 の通りになった。

表 4 日経平均株価騰落予測結果

| | Accuracy | Precision | Recall | F1-score |
|--------|----------|-----------|--------|----------|
| 汎用 | 0. 556 | 0. 584 | 0. 625 | 0. 604 |
| 金融 | 0. 541 | 0. 567 | 0. 638 | 0. 601 |
| SNS | 0. 541 | 0. 569 | 0. 625 | 0. 596 |
| 金融 SNS | 0. 560 | 0. 583 | 0. 562 | 0. 572 |
| 疑似データ | 0. 571 | 0. 600 | 0. 65. | 0. 625 |

5. 4 考察

疑似データを用いた場合、全ての評価指標で株価予測の精度が高くなった。これは、疑似データ生成がドメイン間のギャップを埋め、ターゲットドメイン (金融 SNS ドメイン) への適応力を高めたと考えられる。英語を翻訳した金融 SNS ドメインでは、Accuracy と Precision は他のドメインと比較して良好な結果を示した一方、Recall が 0. 526 と低い値にとどまった。この結果は、翻訳過程で文脈情報が失われたり、言語間の感情表現の微妙な違いがモデルの感情分析に影響を与えた可能性を示唆している。

6 おわりに

本稿では、感情分析を用いた株価予測において、疑似データ生成を活用することで予測精度が向上することを示した。本研究の成果は、膨大な SNS データを効率的に活用しつつ、計算コストを抑えながら高精度な予測を実現する手法の有効性を示している。

また、株価予測の精度向上に寄与するだけでなく、他の複雑なドメインにおいても適応可能な分析フレームワークの基盤を提供することが期待される。

謝辞

本研究は JSPS 科研費 20K11960 の助成を受けたものです。

参考文献

- [1] Joha. Bollen, Huin. Mao, Xiao-Ju. Zeng : Twitte. moo. predict. th. stoc. market. arXiv:1010. 3003v. [cs. CE] (1. Oc. 2010)
- [2] SNS 及び, RSS と株価変動の相関. 山野剛志, 山下海, 横川隼也, 広瀬啓雄 第 84 回全国大会講演論文集, 2021 年
- [3] 株式市場におけるツイート情報と小型株の関連性. 片山慎也, 高橋大志, 人工知能学会第二種研究会資料, 第 BI-01. 巻, 16, 2020 年
- [4] 日本語タスクにおける LL. を用いた疑似学習データ生成の検討. 藤井巧朗, 勝又智. 言語処理学会 第 30 回年次大会 発表論文集, 2024 年 3 月.
- [5] AI4Financ. Foundation. (n. d.). *FinGP. Sentimen. Trai. Dataset*. Retrieve. Octobe. 1, 2024, , fro . <https://huggingface.co/datasets/FinGPT/fingpt-sentiment-train>
- [6] Hayashibe, Yuta. "Japanes. Realisti. Textua. Entailmen. Corpus. " *Proceeding. o. th. Europea. Languag. Resource. Associatio. (ELRA)*, Marseille, 11–1. Ma. 2020, pp. 6827–6834.
- [7] Takahir. Kubo, Hirok. Nakayama. chABSA: Aspec. Base. Sentimen. Analysi. datase. i. Japanese. UR. <https://github.com/chakki-works/chABSA-dataset>
- [8] Tomoyuk. Kajiwara, Chenu. Chu, Norik. Takemura, Yut. Nakashima, an. Hajim. Nagahara. WRIM. : . ne. datase. fo. emotiona. intensit. estimatio. wit. subjectiv. an. objectiv. annotations. I. Proceeding. o. th. 202. Conferenc. o. th. Nort. America. Chapte. o. th. Associatio. fo. Computatio. Linguistics: Huma. Languag. Technologies, pp. 2095–2104, 2021. [4] Jaco. Devlin, Ming
- [9] Masahir. Suzuki, Hirok. Sakaji, Masanor. Hirano, Kiyosh. Izumi. "Constructin. an. analyzin. domain-specifi. languag. mode. fo. financia. tex. mining, " *Informatio. Processin. & Management*, Volum. 60, Issu. 2, 2023