

2017年度 工学院大学卒業論文

Twitter に表れる感情を利用した株価予測

Predicting the stock market using twitter mood



情報学部 コンピュータ科学科
応用経済分析研究室

学籍番号：J1-14070

氏名：高波 龍太郎

指導教員：矢崎 敬人准教授

目次

1. はじめに.....	1
2. 方法.....	2
2.1 実験環境.....	2
2.2 実験方法.....	3
2.2.1 感情辞書作成	3
2.2.2 感情値の時系列データ作成	8
2.2.3 1年間の感情値と株価の因果関係の推定	9
2.2.4 感情値を利用した株価の予測.....	10
3. 結果.....	11
3.1 グレンジャー因果分析の分析結果	11
3.2 XGBoost による予測精度算出結果	15
4. 考察.....	18
5. 結論.....	19

1. はじめに

近年ソーシャルネットワーキングサービス（SNS）が広く普及し、総務省情報通信政策研究所によると 2016 年においては日本人全体の約 70%以上の人々が SNS を利用している（図 1）. 数ある SNS の中でも本研究は Twitter に注目した. Twitter は日常で感じたことなどを不特定多数の人々に「つぶやく」ように自由に発信（ツイート）するマイクロブログサービスであり、その特徴からその時点のユーザの感情が表れやすいと一般的に考えられている. Twitter のツイート内容からツイートしたときの感情を分析する研究も進んでおり、山本他[1]では感情を表す単語を収録した『感情表現辞典』[2]をもとに熊本他[3]の感情語辞書構築法を応用し、ツイート内容から適切に発信者の感情を推測する研究を行った. 山本他[1]の研究では個々のツイート内容から感情を分析し、人間がそのツイートを見て感じた感情との比較を行っているが、本研究では一年間のツイートを時系列データとして取得し、個人の感情ではなく、大衆の感情を推測しているという点で相違がある.

ツイート内容を分析して得た感情データを株価の分析に役立てようという研究も進んでおり、佐々木他[4]は対象とする企業に関するツイートがポジティブな内容かネガティブな内容かを調べ、それが株価に影響しているかを調べた. 本研究は感情を 6 つに分類して分析している点、特定の企業の株価ではなく、日経平均株価を分析対象としている点で相違する. Bollen et al. [5]では英語のツイートに対して 6 次元の感情（Calm, Alert, Sure, Vital, Kind, Happy）と Positive, Negative の度合いに関してそれぞれ分析を行い、感情とダウ平均株価に因果関係があることを示した. Bollen et al. [5]が分析対象としたツイートは 2008 年の英語のもののみであった. 本研究の目的は、より SNS が発展した 2016 年の日本語のツイートにおいても人間の感情と日経平均株価との因果関係の有無を検証し、感情が株価の予測指標となりうるかを検討することである.

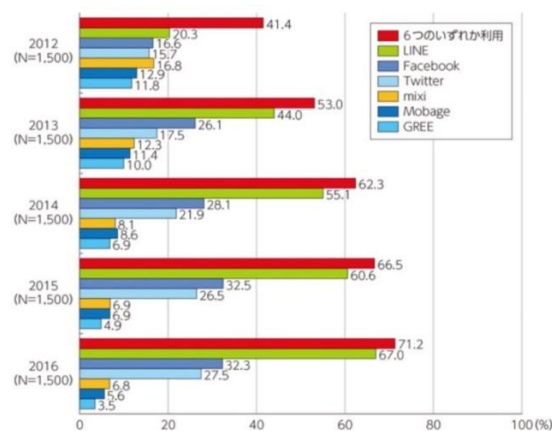


図 1：代表的 SNS の利用率の推移（出典）総務省情報通信政策研究所「情報通信メディアの利用時間と情報行動に関する調査」

2. 方法

本章では実験環境，実験方法について記述する．

2.1 実験環境

- OS X El Capitan (ver. 10.11.6)
 - Anaconda (ver. 4.3.30)
 - ✧ Python (ver. 3.5.4)
 - ✧ Selenium (ver. 3.6.0)
 - ✧ ChromeDriver (ver. 2.34)
 - ✧ MongoDB (ver. 3.4.10)
 - R Studio (1.1.383)
 - ✧ R (ver.3.3.3)
- Virtual Box (ver. 5.2.4)(OS: Ubuntu (ver. 16.04))
 - Anaconda (ver. 4.3.30)
 - ✧ Python (ver. 3.6.3)
 - ✧ JUMAN++ (ver. 1.02)
 - ✧ JUMAN (ver. 7.01)
 - ✧ KNP (ver. 4.18)
 - ✧ PyKNP (ver. 0.3)

本研究ではデータの収集，分析を行う際に主に Python を使用した．しかし，後に述べるグレンジャー因果分析においては Python よりも容易に行うことができる R を使ってプログラミングを行った．Selenium, ChromeDriver はデフォルトのものから拡張した Python のパッケージであり，後述する映画レビューデータをウェブスクレイピングする際に使用した．

英文とは異なり，日本語のツイートは単語ごとの区切りが明確にスペースなどで区切られておらず，機械的に単語を処理する場合，形態素解析をし，分かち書きという単語ごとにスペースで区切る処理を事前に行う必要があった．分かち書きを行うツールとして日本語形態素解析システム JUMAN++パッケージを使用した．Python にて処理を行う場合は JUMAN++の他に JUMAN, KNP, PyKNP を同時にインストールする必要がある．また JUMAN++は Linux 上で動作するように設計されているので Virtual Box の仮想環境上に Linux のディストリビューションの1つである Ubuntu をインストールし，分かち書きの処理のみ Ubuntu 上の Python で行った．

2.2 実験方法

本研究では 2016 年 1 月 1 日から 2016 年 12 月 31 日までのツイートをそれぞれ 1000 件取得する．その後熊本他[3]と山本他[1]をもとに作成した 6 次元の感情（喜，哀，昂，驚，怒，怖）の感情語辞書をもとに，取得したツイートを 1 日ごとに推定し，時系列データを作成した．その後，作成した各時系列データと日経平均株価に対してグレンジャー因果分析を行い，感情の各時系列データと株価との間の因果関係を推定した．最後に因果関係が認められた時系列データに関して実際の株価の予測が行えるかを検討した．実験方法について以下の 4 段階に分けて記述する．

- ① 感情語辞書作成
- ② 感情値の時系列データ作成
- ③ 1 年間の感情値と株価の因果関係の推定
- ④ 感情値を利用した株価の予測

2.2.1 感情辞書作成

文章から感情を推定するために，単語と感情を定量的に表した感情値を結びつけた感情語辞書を作成する．構築方法は熊本他[3]とそれを応用して感情語辞書構築法を Twitter に適応した形に発展させた山本他[1]を参考にした．熊本他[3]の手法は，感情語辞書を構築するために，「ある印象を有する単語はその印象を表現する印象語群と共起しやすく，逆の印象を表現する印象語群とは共起しにくい」という仮定のもと，大量のデータに現れる任意の単語とあらかじめ定義してある感情語群との共起関係を調べ，その結果に基づいてそれぞれの単語の感情値を 1～0 の実数値で数値化している．

本研究では山本他[1]にならい『感情表現辞典』[2]にある単語を感情語群とし，感情語群との共起関係を調べるためのデータとして Yahoo!映画レビューのレビューデータを採用して，それぞれの感情（喜，哀，昂，驚，怒，怖）ごとに感情語辞書を構築した．Yahoo!映画レビューのレビューデータを使用する理由は，ツイートでは辞書を構築するための文章としては短く，レビューデータは熊本他[3]が採用した新聞の文章よりもツイートに近いからである．『感情表現辞典』[2]には（喜，哀，昂，驚，怒，怖，好，安，恥，厭）の 10 の感情を表す単語がそれぞれ感情別に掲載されている．しかし，10 の感情に分類して感情を推定した場合，感情推定値の精度が低くなることが山本他[1]により示されている．そこで山本他[1]はそれぞれの感情語の共起頻度を測定して類似する感情を統合することを提案し，（喜+好，哀，昂，驚，怖，安，恥，怒+厭）の 8 つに分類することを提案した．

また、熊本他[3]の感情語辞書構築法を使用するにはそれぞれの感情に対となる感情が必要であり、山本他[1]では対となる感情を共起分析、クラスタ分析と心理学者 Plutchik and Kellerman [6]の提唱した「感情の輪」を利用して決定した。しかし、山本他[1]の分類した8つに感情では対応関係が（喜+好⇔怒+厭、哀⇔怖、昂⇔驚、安⇔恥）となっており、全てをそれぞれ「感情の輪」にうまく対応させることができていない。そこで本研究ではより「感情の輪」の対応関係を重視し、山本他[1]が分析した感情の共起頻度と感情の輪に基づいて6つの感情に分類することにした（喜+好+安、哀、昂、驚+恥、怒+厭、怖）。以降、感情の輪の対応関係から6つに分類した感情を（喜、哀、昂、驚、怒、怖）と記述する。対応関係は（喜⇔哀、昂⇔驚、怒⇔怖）とした。

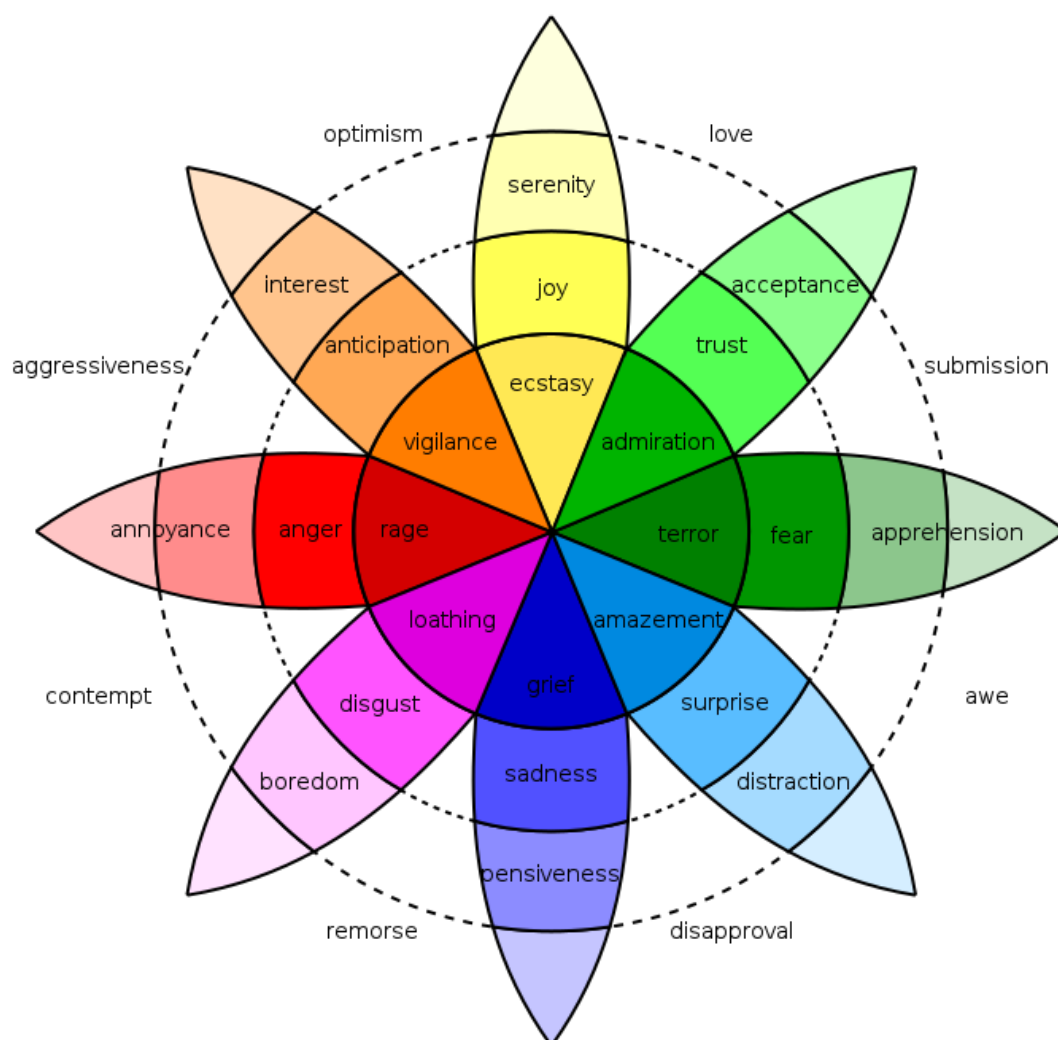


図 2：感情の輪
出所：Plutchik and Kellerman [6]

レビューデータに関しては Yahoo!映画のサイトをウェブスクレイピングすることによって取得した。集めたレビューデータの映画のタイトルと取得データ数は下図の通りである。本実験では集めたデータ数が実験結果に及ぼす影響を検討するため、映画.comに掲載されている2016年度映画興行収入の邦画、洋画それぞれトップ10の映画（計20タイトル）のレビューデータと一般社団法人日本映画製作者連盟が公開している「2016年度興収10億円以上番組」に掲載され、すでに取得したものとかぶらないタイトルのレビューデータを追加で取得したもの（計62タイトル）とでそれぞれ辞書を構築し比較した。それぞれの映画タイトルとレビューデータ数を示す。合計レビュー数は前者が27480で後者が53268であった。

表 1：2016 年度映画興行収入ランキング

ランキング	邦画		洋画	
	タイトル	データ数	タイトル	データ数
1	君の名は。	8366	スター・ウォーズ フォースの覚醒	3540
2	シン・ゴジラ	5509	ズートピア	966
3	名探偵コナン 純黒の悪夢 (ナイトメア)	829	ファインディング・ ドリー	452
4	映画妖怪ウォッチ エンマ大王と5つ の物語だニャン！	135	ペット	302
5	ONE PIECE FILM GOLD	360	オデッセイ	1340
6	信長協奏曲 (ノブナガコン ツェルト)	652	007 スペクター	914
7	映画ドラえもん 新・のび太の日本 誕生	209	アリス・イン・ワン ダーランド 時間の 旅	302
8	暗殺教室 卒業編	197	インデペンデンス・ デイ リサージェン ス	1167
9	orange オレンジ	500	シビル・ウォー キャプテン・アメリ カ	755
10	ガールズ&パン ツアー 劇場版	652	ジャングル・ブック	333

表 2：追加でデータを取得した映画タイトルとデータ数

タイトル	データ数	タイトル	データ数
映画 「聲の形」	1122	殿、利息でござる！	572
植物図鑑 運命の恋、ひろいました	441	エヴェレスト 神々の山嶺	539
デスノート Light up the NEW world	853	青空エール	283
ポケモン・ザ・ムービーXY&Z ボルケニオンと機巧のマギアナ	70	黒崎くんの言いなりになんてならない	139
映画クレヨンしんちゃん 爆睡！ユメミーワールド大突撃	410	世界から猫が消えたなら	645
HiGH&LOW THE MOVIE	326	ちはやふる -下の句-	434
母と暮せば	420	杉原千畝 スギハラチウネ	303
64-ロクヨン-前編	522	オオカミ少女と黒王子	206
64-ロクヨン-後編	473	HiGH&LOW THE RED RAIN	156
ちはやふる -上の句-	708	超高速！参勤交代 リターンズ	288
アイアムアヒーロー	1281	何者	641
さらばあぶない刑事	475	貞子vs伽椰子	467
怒り	1350	デッドプール	750
この世界の片隅に	3255	バットマン v s スーパーマン ジャスティスの誕生	937
後妻業の女	247	スーサイド・スクワッド	701
TOO YOUNG TO DIE！若くして死ぬ	717	アーロと少年	302
ミュージアム	815	インフェルノ	450
ルドルフとイッパイアッテナ	274	ジェイソン・ボーン	598
四月は君の嘘	628	ハドソン川の奇跡	797
僕だけがいない街	652	ゴーストバスターズ	499
家族はつらいよ	214	レヴェナント：蘇えりし者	828

取得したレビューデータに対して形態素解析システム JUMAN++を用いて分かち書きを行い文字列として処理しやすい形にした後、熊本他[3]の手法に倣って感情語辞書を構築した。

まずは対となる感情ごとにレビューデータがそれぞれどちらの感情に近い文章であるかを分類した。分類法に関しては各レビューデータの文章と感情語群の共起数を計測した。左右の対応する感情ごとにそれぞれ共起数を計測し、より多くの共起が確認された方の感情にそのレビューデータを感情データとして分類し、それぞれの感情ごとのレビューデータの集合（以降、レビュー集合）を作った。ただし、全く共起が確認されなかった場合や、共起数が同数であったレビューデータは省いた。

次に感情のレビュー集合を各対となる感情ごとにそれぞれ左右として扱う。対応する感情（喜⇔哀、昂⇔驚、怒⇔怖）のうち左側の感情を表すレビュー集合を **SL**（レビュー数を **NL**）とし、右側のレビュー集合を **SR**（レビュー数を **NR**）とした。次に、それぞれのレビュー集合（**SL** もしくは **SR**）から名詞、動詞、形容詞、副詞、感動詞の単語を抽出し、単語ごとに出現頻度を数えた。このとき、ある単語 **w** の記事集合 **SL** における出現頻度を **NL(w)**、レビュー集合 **SR** における出現頻度を **NR(w)** とすると、それぞれの条件付確率は、

$$PL(w) = \frac{NL(w)}{NL} \quad , \quad PR(w) = \frac{NR(w)}{NR}$$

この **PL(w)** と **PR(w)** を用いて、左側の感情のレビュー集合に属する単語 **w** の感情値 **v(w)** を次の式で表す。

$$v(w) = \frac{PL(w) * \log_{10} NL}{PL(w) * \log_{10} NL + PR(w) * \log_{10} NR}$$

すなわち、単語 **w** の左側の感情のレビューデータ群の単語に対する条件付確率と右側の感情のレビューデータ群の単語に対する条件付確率の重み付き内分比を求め、単語 **w** の感情値として感情語辞書に登録した。なお、 $\log_{10} NL$ と $\log_{10} NR$ は重みであり、条件を満たす記事数 **NL** あるいは **NR** が多いほど大きくなるように設計されている。また山本他[1]の研究では登録する単語を感情値が 0.4 以上のもののみと記述されているが結果を比較するため 0.4 以上、0.5 以上、0.6 以上、0.7 以上、0.8 以上を閾値として単語を登録した感情語辞書をそれぞれ構築した。合計すると、6 つの感情、5 つの感情値閾値、2 つの取得レビューデータ数の感情語辞書を構築した。以下にそれぞれの感情語辞書の収録単語数を示す。

表 3：レビュー数の 27480 感情語辞書の収録単語数

感情値閾値	喜	哀	昂	驚	怒	怖
0.4	5969	6998	10206	12720	8813	11790
0.5	4499	5485	6282	9495	6367	8407
0.6	2986	4015	3057	5571	2984	5961
0.7	1284	2829	856	3976	1227	4074
0.8	342	2027	197	872	287	2160

表 4：レビュー数 53268 の感情語辞書の収録単語数

感情値閾値	喜	哀	昂	驚	怒	怖
0.4	11157	12381	15111	18703	13535	16593
0.5	8477	9283	9340	13836	9188	12042
0.6	5379	6603	4473	8065	4637	7695
0.7	2346	4374	1298	5637	1739	4996
0.8	620	2432	316	1242	423	2613

また，感情語辞書に登録された単語と感情値の例を以下に示す．

表 5：感情語辞書の単語と感情値の例

感情	単語	感情値
喜	好き	0.94
	嬉しい	0.93
	幸せな	0.84
哀	不憫	0.86
	可哀想	0.84
	孤独	0.84
昂	感動	0.96
	感激	0.95
	興奮	0.93
驚	咄然	0.98
	意外	0.96
	驚愕	0.94
怒	腹立たしい	0.84
	殺意	0.82
	憎い	0.84
怖	不安	0.92
	薄気味悪い	0.91
	臆病	0.98

2.2.2 感情値の時系列データ作成

ツイッターを提供している Twitter, Inc.からはツイッター上の情報を有効に

使用するため開発者向けに **Twitter API** が提供されている。しかし, **Twitter API** では過去一週間分のツイート情報以外取得できないため, 本研究の対象期間である 2016 年 1 月 1 日から 2016 年 12 月 31 日のツイートデータを取得することができなかった。そこでツイッターのサイト **Tweet Search** をウェブスクレイピングし, 1 日ごとに日付のみの指定で表示された上位 1000 ツイートを取得した。日本における大衆の感情値と株価の相関を調べるため, **Tweet Search** 内の「高度な検索」機能より取得するツイートの言語を日本語に設定した。

取得したツイートデータに対して 2.2.1 節で構築した感情語辞書を使用してツイートの感情値を推定した。取得した各ツイートの文章に 6 つの感情ごとに感情語辞書に一致する単語が含まれていた場合, その感情語の単語の感情と感情値をそのツイートの感情値として数えた。各ツイートの感情値を 1 日ごとに集計し, その 1 日の感情値とした。そしてそれを 2016 年 1 月 1 日から 2016 年 12 月 31 日まで行い 1 年間の感情の時系列データとした。ただし, 感情データと株価の分析を行うため, 株取引が行われていない土曜日, 日曜日, 祝日のデータは除いた。そのようにして作成した 6 つの感情の時系列データをそれぞれ比較するため正規化を行い, Z 値とした。正規化の式は下記式 3 の通りである。 X_t は感情値の時系列データを示す。6 つの感情, 5 つの感情値閾値, 2 つの取得レビューデータ数の感情語辞書を構築したので, それに対応して合計で 60 種の時系列データを作成した。

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

2.2.3 1 年間の感情値と株価の因果関係の推定

株式データは 2016 年 1 月 1 日から 2016 年 12 月 31 日の期間でそれぞれの終値のデータを株価データサイト **k-db.com** より取得し, 株価の差分系列を作成した。次に感情値の時系列データと同様上記の式で Z 値を求め正規化した。正規化を施した株価の差分系列データと感情の Z 値の時系列データに対してグレンジャー因果分析を行った。グレンジャー因果分析とは, 現在と過去の x の値のみで行った x の予測と, x と y の値による未来の x の値の予測を比較し, 後者の平均二乗誤差の方が小さくなる場合 y に x に対してのグレンジャー因果性が存在すると判断する分析手法である。本研究においては x が株価, y が感情値である。感情値が未来の株価を予測できるかの因果関係を調べるため, 1 日前から最大 7 日前の株価と感情の時系列データを利用した因果分析を行った。モデル式は以下の通りである。

$$L_1: D_t = \alpha + \sum_{i=1}^n \beta_i D_{t-i} + \epsilon_t$$

$$L_2: D_t = \alpha + \sum_{i=1}^n \beta_i D_{t-i} + \sum_{i=1}^n \gamma_i X_{t-i} + \epsilon_t$$

D_t : 時点 t の日経平均株価

X_t : 時点 t の感情値

ϵ_t : 誤差項

n : ラグの長さ (day)

グレンジャー因果分析を行うにあたって, **var** パッケージにて関数が用意されている統計ソフト **R** を使用した.

2.2.4 感情値を利用しての株価の予測

機械学習の手法である **eXtreme Gradient Boosting (XGBoost)** を用いて, 2.2.3 節の結果株価との因果関係が認められた感情値の時系列データによる株価の予測精度を検証した. **XGBoost** とは複数の弱識別器を作成して学習するブースティングという手法の一種である. このとき, 弱識別器が直列的に順番に学習されていき, それぞれの学習において前の弱識別器の学習結果が参考にされる. **XGBoost** は学習を決定木の作成によって行うようにしたもので, それぞれの学習において勾配法を使って精度を向上させている.

データ \mathbf{x} に対する予測値を返す t 番目の決定木 $f_t(\mathbf{x})$ を構築する際には下式 $L(t)$ が最小になるようにする.

$$L(t) = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \gamma T + \frac{1}{2} \lambda \|\mathbf{w}\|^2$$

このとき \mathbf{n} はデータ数であり, y_i は実測値, $\hat{y}_i^{(t-1)} + f_t(x_i)$ は予測値, 関数 l は損失関数である. 損失関数内の予測値の部分は予測値が $t-1$ 番目までの決定木を考慮して t 番目の決定木を構築していることを示している. 損失関数以降の部分は過学習を防ぐための罰則項であり, T は木の大きさを表し, γ は木の大きさに対するペナルティである. $\gamma > 0$ を設定した場合, γ の値が大きいほど T の値は小さくなり, 木の大きさが小さくなる. \mathbf{w} は決定木が返すことのできる値のベクトルであり, 例えば $T=5$ の場合 $\mathbf{w}=(w_1, w_2, w_3, w_4, w_5)$ である. すなわち \mathbf{w} は $\hat{y}_i^{(t)}$ がとることのできる値の集合である. λ は \mathbf{w} の値を調整するパラメータであり, λ の値が大きいほど \mathbf{w} の値は小さくなり, 実測値と予測値の誤差を少し

ずつ更新するので、学習に時間がかかるもののさまざまなパターンを学習することができる。しかし、さまざまなパターンを学習することは過学習につながる可能性があるので、最適な更新具合にすることが重要である。

予測を行うにあたって Python の XGBoost パッケージを使用した。今回の調査期間 1 年間のうち前半分のデータを使用して学習を行い、後半分のデータで予測精度を検証した。感情値と株価をそれぞれ 1 日前～7 日前の間で入力し、翌日の株価の終値が前日以上になるか、下落するかの予測制度をそれぞれの期間において全ての辞書において行った。ハイパーパラメータである学習率と決定木の深さの最大値は予測制度が最大になるように調整を行った。

3. 結果

3.1 グレンジャー因果分析の分析結果

グレンジャー因果分析の分析結果を下表にて示す。表の各数値は p 値であり、0.1 以下のものには*を 0.05 以下のものには**を右隣に表記した。p 値は因果関係が存在する確立を表しており、0.1 以下であれば 90%以上、0.05 以下であれば 95%以上の確率で有意である。また、左側の項目ラグ(day)はモデル式の n に相当するもので、前日から何日前までの値を用いて予測を行ったかを示している。例えば、ラグ 7 であれば前日から 7 日前までの値を用いて予測を行ったことを表す。

表 6：レビュー数 27480 で感情値 0.4 以上の辞書での分析結果

感情値>0.4						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.02321**	0.00811**	0.01018**	0.01286**	0.01486**	0.0148**
2	0.07943*	0.03547**	0.03212**	0.05195*	0.04112**	0.05793**
3	0.1215	0.03509**	0.05704*	0.07013*	0.06759*	0.08463*
4	0.07187*	0.03057**	0.05637*	0.04275**	0.04589**	0.05288*
5	0.1207	0.06544*	0.09947*	0.07838*	0.08452*	0.09395*
6	0.146	0.1046	0.1275	0.0991*	0.1046	0.11
7	0.07784*	0.01476**	0.03869**	0.0315**	0.05547*	0.03263**

表 7：レビュー数 27480 で感情値 0.5 以上の辞書での分析結果

感情値>0.5						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.02989**	0.00303**	0.0064**	0.05415*	0.01226**	0.03979**
2	0.09655*	0.01484**	0.01376**	0.09599*	0.02949**	0.0986*
3	0.1527	0.01135**	0.02845**	0.1088	0.05282*	0.138
4	0.09029*	0.01056**	0.03101**	0.06815*	0.0338**	0.06376*
5	0.1426	0.02386**	0.05881*	0.1181	0.06084*	0.1157
6	0.1756	0.04753**	0.08584*	0.1551	0.07889*	0.1332
7	0.1024	0.00711**	0.0257**	0.05821*	0.03738**	0.03133**

表 8：レビュー数 27480 で感情値 0.6 以上の辞書での分析結果

感情値>0.6						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.02641**	0.00582**	0.06694*	0.09947*	0.03895**	0.03238**
2	0.07798*	0.02575**	0.1995	0.09167*	0.09581*	0.06253*
3	0.1297	0.02268**	0.288	0.07906*	0.1545	0.09761*
4	0.06989*	0.02824**	0.3535	0.03977**	0.05647*	0.02531**
5	0.1113	0.04906**	0.437	0.07987*	0.1024	0.05051*
6	0.1344	0.089*	0.4924	0.1298	0.1129	0.0702*
7	0.9273*	0.02732**	0.1994	0.09498*	0.05395*	0.00696**

表 9：レビュー数 27480 で感情値 0.7 以上の辞書での分析結果

感情値>0.7						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.3454	0.01582**	0.1904	0.1237	0.1451	0.02333**
2	0.5638	0.06157*	0.4209	0.1081	0.1056	0.08014*
3	0.7216	0.07861*	0.5546	0.1657	0.097*	0.1118
4	0.3285	0.1177	0.6797	0.07389*	0.02538**	0.05864*
5	0.3456	0.1324	0.7787	0.1415	0.05418*	0.09072*
6	0.4662	0.2425	0.8004	0.2129	0.06784*	0.1282
7	0.5111	0.1833	0.8339	0.1695	0.02273**	0.05392*

表 10：レビュー数 27480 で感情値 0.8 以上の辞書での分析結果

感情値>0.8						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.3112	0.00549**	0.5012	0.7509	0.5063	0.03422**
2	0.426	0.02246**	0.5712	0.6081	0.2175	0.1125
3	0.6338	0.02651**	0.6949	0.6464	0.1565	0.1527
4	0.6715	0.05211*	0.847	0.1544	0.1468	0.1814
5	0.7192	0.0865*	0.8831	0.1411	0.2695	0.1249
6	0.8008	0.172	0.9102	0.1703	0.3092	0.1484
7	0.8463	0.1545	0.8163	0.1561	0.406	0.1649

表 11：レビュー数 53268 で感情値 0.4 以上の辞書での分析結果

感情値>0.4						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.01208**	0.0125**	0.00925**	0.02269**	0.01407**	0.01351**
2	0.03744**	0.0456**	0.02816**	0.0842*	0.04367**	0.0536*
3	0.05969*	0.05974*	0.04656**	0.116	0.07104*	0.06017*
4	0.03017**	0.06958*	0.04549**	0.06094*	0.0516*	0.04175**
5	0.05501*	0.1257	0.08259*	0.111	0.09328*	0.08357*
6	0.06882*	0.1623	0.1119	0.1331	0.128	0.1067
7	0.02727**	0.03778**	0.03003**	0.04113**	0.06273*	0.03092**

表 12：レビュー数 53268 で感情値 0.5 以上の辞書での分析結果

感情値>0.5						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.01418**	0.01152**	0.00337**	0.1087	0.01142**	0.02859**
2	0.04131**	0.03877**	0.00617**	0.1678	0.02795**	0.06423*
3	0.06508*	0.04093**	0.01248**	0.2012	0.05416*	0.07291*
4	0.02539**	0.0491**	0.0148**	0.09772*	0.0375**	0.05514*
5	0.04706**	0.07215*	0.02567**	0.1853	0.06622*	0.1122
6	0.05479*	0.09469*	0.03722**	0.2404	0.09209*	0.142
7	0.02711**	0.03461**	0.01201**	0.109	0.06223*	0.02692**

表 13：レビュー数 53268 で感情値 0.6 以上の辞書での分析結果

感情値>0.6						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.01462**	0.02455**	0.00958**	0.08413*	0.02716**	0.0239**
2	0.05243*	0.06515*	0.02344**	0.1187	0.09131*	0.04314**
3	0.07549*	0.03378**	0.03923**	0.16	0.1934	0.05927*
4	0.03129**	0.0377**	0.05131*	0.03301**	0.1835	0.04207**
5	0.06403*	0.03576**	0.09411*	0.07207*	0.2722	0.08331*
6	0.07243*	0.0525*	0.1308	0.08393*	0.3413	0.1032
7	0.04869**	0.05301*	0.1198	0.07925*	0.273	0.04351**

表 14：レビュー数 53268 で感情値 0.7 以上の辞書での分析結果

感情値>0.7						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.1389	0.01928**	0.03107**	0.1737	0.1164	0.0086**
2	0.3138	0.06828*	0.06873*	0.2051	0.2992	0.03119**
3	0.466	0.01597**	0.1311	0.2782	0.4685	0.05076*
4	0.4299	0.02311**	0.2151	0.09533*	0.4586	0.08246*
5	0.5184	0.01409**	0.3455	0.1578	0.3506	0.1678
6	0.5517	0.024**	0.2469	0.184	0.3713	0.1412
7	0.4073	0.0382**	0.2317	0.144	0.1938	0.1503

表 15：レビュー数 53268 で感情値 0.8 以上の辞書での分析結果

ラグ(day)	感情値>0.8					
	喜	哀	昂	驚	怒	怖
1	0.2085	0.02085**	0.06261*	0.5281	0.159	0.02957**
2	0.1956	0.07244*	0.1756	0.3579	0.3974	0.1061
3	0.2648	0.03865**	0.2679	0.593	0.4895	0.09487*
4	0.4045	0.07033*	0.3476	0.07401*	0.3855	0.1544
5	0.4296	0.155	0.5634	0.03202**	0.3012	0.3298
6	0.4898	0.2255	0.3962	0.04321**	0.4259	0.2247
7	0.3121	0.2167	0.4005	0.05155*	0.04547**	0.3004

感情値が株価を予測する指標になりえるかを研究した Bollen et al. [5]では強い因果関係が確認された感情辞書は 4 つのラグにおいて p 値が 0.05 以下であったため、本研究では 4 つ以上のラグにおいて p 値が 0.05 以下の辞書のみ強い因果関係があると判断した。よって、レビュー数 278480 の辞書では感情値の閾値が 0.4 の哀と 0.5 の哀, 昂, 怒と 0.6 の哀の辞書が, レビュー数 53268 の辞書では感情値の閾値が 0.4 の喜, 昂と 0.5 の喜, 哀, 昂と 0.6 の哀, 怖と 0.7 の哀の感情語辞書において株価との強い因果関係が認められた。特に強い因果関係があると判断したものを以下に抽出した。

表 16：レビュー数 27480 で特に強い因果関係を示す感情語辞書

ラグ(day)	感情値>0.4	感情値>0.5			感情値>0.6
	哀	哀	昂	怒	哀
1	0.008113**	0.00303**	0.0064**	0.01226**	0.00582**
2	0.03547**	0.01484**	0.01376**	0.02949**	0.02575**
3	0.03509**	0.01135**	0.02845**	0.05282*	0.02268**
4	0.03057**	0.01056**	0.03101**	0.0338**	0.02824**
5	0.06544*	0.02386**	0.05881*	0.06084*	0.04906**
6	0.1046	0.04753**	0.08584*	0.07889*	0.089*
7	0.01476**	0.00711**	0.0257**	0.03738**	0.02732**

表 17：レビュー数 53268 で特に強い因果関係を示す感情語辞書

ラグ(day)	感情値>0.4		感情値>0.5			感情値>0.6		感情値>0.7
	喜	昂	喜	哀	昂	哀	怖	哀
1	0.01208**	0.00925**	0.01418**	0.01152**	0.00337**	0.02455**	0.0239**	0.01928**
2	0.03744**	0.02816**	0.04131**	0.03877**	0.00617**	0.06515*	0.04314**	0.06828*
3	0.05969*	0.04656**	0.06508*	0.04093**	0.01248**	0.03378**	0.05927*	0.01597**
4	0.03017**	0.04549**	0.02539**	0.0491**	0.0148**	0.0377**	0.04207**	0.02311**
5	0.05501*	0.08259*	0.04706**	0.07215*	0.02567**	0.03576**	0.08331*	0.01409**
6	0.06882*	0.1119	0.05479*	0.09469*	0.03722**	0.0525*	0.1032	0.024**
7	0.02727**	0.03003**	0.02711**	0.03461**	0.01201**	0.05301*	0.04351**	0.0382**

3.2 XGBoost による予測精度算出結果

次に実際に株価の騰落を予測することができる精度を機械学習の XGBoost を使用して求めた結果を下記に示す。因果関係が認められたものに予測精度を向上させる効果があると考えられるが、例外があるか検証するために因果分析の結果に関わらず全ての感情語辞書の時系列データに対して予測精度を検証した。各感情の時系列データを含めた株価予測に関して示す前に株価のみで予測を行った場合の予測精度を示す。以下感情と株価双方の時系列データを使用した予測の結果には、株価のみで予測を行った場合の最大値 0.58 よりも高い予測精度の枠を網掛けした。

表 18：株価のみでの予測精度

ラグ(day)	株価のみ
1	0.52
2	0.580645
3	0.568
4	0.563492
5	0.559055
6	0.5625
7	0.542636

表 19：レビュー数 27480 で感情値 0.4 以上の辞書での予測精度

感情値>0.4						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.55	0.53	0.53	0.54	0.55	0.54
2	0.55	0.54	0.55	0.56	0.55	0.54
3	0.56	0.54	0.56	0.55	0.53	0.59
4	0.55	0.55	0.56	0.53	0.52	0.55
5	0.53	0.54	0.54	0.57	0.54	0.57
6	0.52	0.57	0.55	0.56	0.53	0.52
7	0.51	0.57	0.50	0.50	0.55	0.46

表 20：レビュー数 27480 で感情値 0.5 以上の辞書での予測精度

感情値>0.5						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.54	0.54	0.56	0.54	0.53	0.54
2	0.55	0.56	0.55	0.56	0.56	0.56
3	0.56	0.51	0.57	0.60	0.54	0.55
4	0.57	0.50	0.56	0.50	0.55	0.57
5	0.54	0.50	0.50	0.54	0.51	0.54
6	0.55	0.48	0.49	0.53	0.52	0.57
7	0.52	0.54	0.55	0.52	0.52	0.57

表 21：レビュー数 27480 で感情値 0.6 以上の辞書での予測精度

感情値>0.6						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.54	0.60	0.56	0.56	0.50	0.53
2	0.56	0.57	0.54	0.55	0.55	0.53
3	0.54	0.54	0.55	0.51	0.58	0.55
4	0.55	0.52	0.55	0.56	0.56	0.50
5	0.54	0.56	0.50	0.47	0.53	0.56
6	0.53	0.55	0.52	0.51	0.53	0.57
7	0.50	0.60	0.51	0.53	0.48	0.50

表 22：レビュー数 27480 で感情値 0.7 以上の辞書での予測精度

感情値>0.7						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.53	0.58	0.52	0.54	0.54	0.55
2	0.54	0.55	0.53	0.56	0.56	0.56
3	0.54	0.54	0.54	0.55	0.57	0.55
4	0.55	0.53	0.53	0.56	0.57	0.57
5	0.54	0.51	0.54	0.55	0.54	0.50
6	0.51	0.55	0.54	0.52	0.53	0.47
7	0.54	0.52	0.57	0.56	0.53	0.48

表 23：レビュー数 27480 で感情値 0.8 以上の辞書での予測精度

感情値>0.8						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.53	0.56	0.57	0.56	0.54	0.56
2	0.55	0.58	0.56	0.52	0.53	0.60
3	0.54	0.61	0.59	0.54	0.56	0.58
4	0.52	0.56	0.58	0.60	0.56	0.49
5	0.51	0.61	0.54	0.59	0.55	0.51
6	0.55	0.52	0.56	0.60	0.58	0.52
7	0.55	0.57	0.53	0.56	0.55	0.54

表 24：レビュー数 53268 で感情値 0.4 以上の辞書での予測精度

感情値>0.4						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.54	0.54	0.55	0.50	0.54	0.50
2	0.53	0.55	0.56	0.52	0.56	0.56
3	0.55	0.55	0.58	0.51	0.58	0.56
4	0.56	0.58	0.53	0.51	0.54	0.56
5	0.51	0.51	0.52	0.51	0.49	0.57
6	0.53	0.57	0.59	0.50	0.56	0.52
7	0.55	0.53	0.53	0.58	0.48	0.57

表 25：レビュー数 53268 で感情値 0.5 以上の辞書での予測精度

感情値>0.5						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.50	0.53	0.56	0.48	0.54	0.55
2	0.54	0.59	0.55	0.57	0.59	0.57
3	0.55	0.56	0.56	0.57	0.60	0.55
4	0.57	0.55	0.54	0.55	0.55	0.56
5	0.54	0.54	0.54	0.50	0.50	0.49
6	0.52	0.55	0.58	0.52	0.55	0.52
7	0.53	0.59	0.55	0.50	0.53	0.50

表 26：レビュー数 53268 で感情値 0.6 以上の辞書での予測精度

感情値>0.6						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.51	0.56	0.56	0.55	0.54	0.57
2	0.55	0.56	0.54	0.54	0.52	0.56
3	0.56	0.58	0.54	0.56	0.51	0.55
4	0.55	0.48	0.52	0.51	0.45	0.56
5	0.54	0.50	0.54	0.45	0.49	0.48
6	0.55	0.49	0.58	0.48	0.48	0.48
7	0.52	0.54	0.56	0.44	0.50	0.48

表 27：レビュー数 53268 で感情値 0.7 以上の辞書での予測精度

感情値>0.7						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.54	0.52	0.55	0.54	0.50	0.54
2	0.54	0.54	0.56	0.48	0.56	0.56
3	0.56	0.60	0.60	0.56	0.54	0.57
4	0.56	0.58	0.57	0.61	0.48	0.54
5	0.54	0.56	0.56	0.50	0.53	0.49
6	0.52	0.55	0.52	0.44	0.52	0.55
7	0.53	0.57	0.49	0.51	0.54	0.50

表 28：レビュー数 53268 で感情値 0.8 以上の辞書での予測精度

感情値>0.8						
ラグ(day)	喜	哀	昂	驚	怒	怖
1	0.54	0.50	0.58	0.55	0.54	0.54
2	0.58	0.52	0.60	0.55	0.52	0.52
3	0.55	0.51	0.62	0.54	0.54	0.61
4	0.56	0.50	0.59	0.56	0.56	0.54
5	0.57	0.47	0.57	0.60	0.50	0.52
6	0.58	0.46	0.54	0.58	0.45	0.53
7	0.57	0.48	0.55	0.53	0.50	0.52

グレンジャー因果分析によって強い因果関係が認められた感情語辞書の XGBoost による予測結果を下記に示す.

表 29：レビュー数 27480 で因果関係が強い辞書の予測精度

ラグ(day)	感情値>0.4	感情値>0.5			感情値>0.6	株価のみ
	哀	哀	昂	怒	哀	
1	0.53	0.54	0.56	0.53	0.60	0.52
2	0.54	0.56	0.55	0.56	0.57	0.58
3	0.54	0.51	0.57	0.54	0.54	0.57
4	0.55	0.50	0.56	0.55	0.52	0.56
5	0.54	0.50	0.50	0.51	0.56	0.56
6	0.57	0.48	0.49	0.52	0.55	0.56
7	0.57	0.54	0.55	0.52	0.60	0.54

表 30：レビュー数 53268 で因果関係が強い辞書の予測精度

ラグ(day)	感情値>0.4		感情値>0.5			感情値>0.6		感情値>0.7	株価のみ
	喜	昂	喜	哀	昂	哀	怖	哀	
1	0.54	0.55	0.50	0.53	0.56	0.56	0.57	0.52	0.52
2	0.53	0.56	0.54	0.59	0.55	0.56	0.56	0.54	0.58
3	0.55	0.58	0.55	0.56	0.56	0.58	0.55	0.60	0.57
4	0.56	0.53	0.57	0.55	0.54	0.48	0.56	0.58	0.56
5	0.51	0.52	0.54	0.54	0.54	0.50	0.48	0.56	0.56
6	0.53	0.59	0.52	0.55	0.58	0.49	0.48	0.55	0.56
7	0.55	0.53	0.53	0.59	0.55	0.54	0.48	0.57	0.54

4. 考察

本研究の結果, 人間の感情と株価の間に因果関係があり, 人間の感情のうち特に「哀」は株価を予測する精度を高める傾向にあることが確認された. レビューデータを多く取得した感情語辞書では株価との因果関係が確認できる感情値時系列の数が増えたことから, レビューデータの数を増やすことが感情語辞書の精度を高めることにつながることを示唆される. ただし, 本研究の結果のみではどの程度感情語辞書構築のためのデータを集めることがよいのか, 適切なデータ量や収録する感情語の感情値閾値が存在するのかは判断できないため, 更なる研究を行う余地がある. また, Bollen et al. [5]ではグレンジャー因果分析にて高い因果関係が確認されたものが予測精度を高める結果になっているが, 今回の実験結果では高い因果関係が確認された場合でも, 株価のみで行った翌日の騰落予測の精度よりも低い予測精度の場合があり, 反対に因果関係が他と比較して高くない場合でも株価の予測精度を高めることがあるという結果になった.

株価の精度向上に関しても、Bollen et al. [5]では最高で 87.6%の精度まで株価の予測精度を向上させる結果となっているが、本研究の結果では最高でも 62%、株価のみで予測した場合よりも 4%ポイントの精度向上と、同研究で示唆されていたよりも大幅に精度が低い結果となった。Bollen et al. [5]と比較して感情分析ツールが異なる点、ツイート数の増加から同研究のように実験期間のツイート全てを取得できなかったことから、感情語辞書の精度やツイートの取得方法の改善により精度向上が見込まれる。感情語辞書からの感情値の分析においては、本研究では「人間の感情」が株価を予測する上での指標となりえるのかを検証することに重きを置き、実際にどの程度個々のツイートの感情や大衆感情を正確に分析できているか、ユーザ実験などの検証を行わなかったため、感情語辞書の分析能力検証と改良が今後の課題として挙げられる。

5. 結論

人間の感情には日経平均株価との因果関係が存在し、感情によって株価の予測精度を向上させることができることが確認された。「哀」の感情に関しては、複数の閾値、感情語辞書構築に使うデータ数の場合において因果関係と予測精度向上が確認されたことから、他の感情と比較して特に株価予測の指標としての有効性が高いことが示された。ただし、先行研究で示唆された予測精度と比較して大幅に低かったことから、感情語辞書の改良などで予測精度を向上させることができるか、また、Twitter が先行研究よりも広まった現在では Twitter からの感情分析では先行研究ほどの高い株価予測精度は見込めないのかを検証することが今後の課題として挙げられる。

参考文献

- [1] 山本湧輝, 「Twitter の感情抽出に基づくフォロイー推薦」, 甲南大学大学院自然科学研究科知能情報学専攻 修士論文 No. 181, 2016.
- [2] 中村明, 『感情表現辞典』, 東京堂出版, 1993.
- [3] 熊本忠彦, 河合由起子, 田中克己, 「回帰分析を応用したテキスト印象マイニング手法の設計と評価」, The 24th Annual Conference of the Japanese Society for Artificial Intelligence, 2010.
- [4] 佐々木理一, 「Twitter を用いた株価予測」, 工学院大学情報学部情報デザイン学科学士論文, 2015.
- [5] Johan Bollen, Huina Mao, Xiao-Jun Zeng, “Twitter mood predicts the stock

market”, CoRR abs/1010.3003, 2010.

[6] Robert Plutchik and Henry Kellerman. *Emotion: Theory, research, and experience. Vol. 1, Theories of emotion*. New York: Academic Press, 1980.