語彙の組み合わせの意味的統合を考慮した複合極性辞書を用いた ニューステキスト分析による株価予測

川崎拓海^{†1} 穴田一^{†2}
東京都市大学^{†1} 東京都市大学^{†2}

1 はじめに

近年、金融予測の分野ではローソク足の画像を用いた分析[1]や数値情報を用いたテクニカル分析[2]などによる様々な研究が行われている。しかし、数値情報ではなくテキスト情報も含まれているニュース記事を考慮することは、日々発表される非数値情報に目を向けることを意味し、数値情報では説明が難しいタイプの市場の予測を行える可能性があると考えられる。

そこで本研究では、新たにニュース記事から金融に関連する単語の組み合わせとその極性値を抽出する語彙獲得アルゴリズムを構築し、獲得した極性辞書を用いたニューステキスト分析による東証株価指数(TOPIX)の株価予測を提案する。

2 提案手法

和泉らの研究[3]では、日本経済新聞の記事に出現した単語を出現パターンごとに抽出し、その中でも出現頻度と株価上昇下落割合がともに閾値以上の単語を特徴語とし、予測前日の記事に特徴語が出現した際の翌日の株価の終値の上昇下落の予測を行っている。しかし、平均正解率は71.4%と高い精度にも関わらず悪い年は56.3%と不安定であった。私達は、これは単語の出現数や出現パターンのみ考慮していて、金融関連の記事に出現する単語や用語の意味を考慮していないことが要因だと考えた。そこでニュース見出し中に出現する単語に着目し、ポジティブな意味をもつ単語が出現すると株価が上昇し、ネガティブな意味をもつ単語が出現すると株価が下落すると考え、研究を進めて来た。

我々の従来研究[4]では日本語評価極性辞書[5]と金融専門極性辞書[6]に含まれる単語が前日のニュースの見出しに出現した際の株価の上昇下落の予測を行った。

しかし、既に作成されている極性辞書には、時事ニュースに出現する金融に関する重要な単語のうち存在しない単

語があるという問題や、既存の極性辞書の単語だけでは“原油価格が上昇”と”売上が上昇”の上昇の様に、ともに出現する単語によって極性が変化することを考慮できない問題があった。

そこで本研究では、見出しに出現した係り受け関係をもつ単語の組み合わせを抽出し、それぞれ組み合わせに極性値を付与することで、同じ単語でも、ともに出現する単語によって極性が変化することを考慮した極性辞書を作成し、その極性辞書に存在する組み合わせの中でもニュース見出し中の出現頻度と極性値のともに高い組み合わせを特徴語として抽出して株価の予測を行った。

2.1 語彙獲得アルゴリズム

初めに初期辞書及び抽出した複合エンタリと株価予測の学習に用いるニュース記事の見出しを用いて、複合エンタリの抽出を初期辞書の単語がニュース記事に出現する毎に行い、複合エンタリのみで構成されたリストを作成し、その後極性値を付与することで複合極性辞書を作成した。複合エンタリとは文中に存在する各文節間の係り受け関係をもつ組み合わせを抽出したものである。複合極性辞書を作成するための初期辞書の作成には金融専門極性辞書を用いた。金融専門極性辞書とは、金融関連の単語にそれぞれ-1.0~+1.0のネガティブ・ポジティブを表す極性値が付与された辞書であり、今回は金融専門極性辞書の単語の中でも極性値の絶対値が高い単語を用いて初期辞書を作成した。

まず初期辞書の単語が対象となるニュースの見出しに出現する毎に、構文解析器の1つであるCabochaを用いて、複合エンタリを抽出した。この時、抽出した複合エンタリは、Mecabを用いて形態素解析を行い、余分な品詞や記号を削除した。

全見出し分実行後、抽出した全ての複合エンタリを元に再度ニュース記事の見出しから新規の複合エンタリを抽出していく。

抽出した複合エンタリを用いて新規の複合エンタリを抽出するフローは、新規の複合エンタリが抽出されなくなるまで行い、複合エンタリのみで構成されたリストを作成した。

最後にニュース記事の見出しに抽出した複合エンタリ

Stock Price Prediction by News Analysis Using Compound Polarity
Dictionary Considering Semantic Integration of Vocabulary
Combination
Takumi Kawasaki[†], Hajime Anada[†], [†]Graduate School of Tokyo
City University

が出現した際、次式で極性値を付与し、複合極性辞書を作成する。

$$p_{i,n} = \text{sign} * \log_e \left(\frac{\text{翌日の終値}}{\text{見出し日の終値}} \right) \quad (1)$$

$$\text{極性値} = \frac{1}{n} \sum (p_{1,1} + p_{1,2} \cdots + p_{i,n}) \quad (2)$$

$p_{i,n}$ とは*i*番目に抽出した複合エントリの当日と翌日の終値の対数変化率によって求めた*n*個目の極性値である。

*sign*とは見出しの複合エントリが否定を含まない場合1、否定を含む場合-1を代入したものであり、変化率の符号を逆転させることで、否定文の時出現した複合エントリの極性値 p_i を考慮している。

(2)では獲得した極性値の平均値を求めている。

2.2 極性辞書を用いた株価予測

本研究ではIT・経済ニュースの記事に対して語彙獲得アルゴリズムを用いて構築した、複合極性辞書を用いたネガティブ・ポジティブ分析による経済動向予測手法を提案する。まず株価予測の学習期間で扱う1日に数十件ずつ掲載されているIT・経済ニュースの見出しから語彙獲得アルゴリズムを用いて複合極性辞書を作成した。そして作成した複合極性辞書の複合エントリの中でも、出現頻度と極性値の絶対値のともに高い複合エントリを特徴語とした。

提案手法では特徴語を用いてSVMを用いた学習を行う。入力では抽出した全特徴語を特徴量とし、予測前日の見出しに特徴語が存在する場合はその特徴語の極性値、存在しない特徴語に関しては0とし、SVMに学習させた。出力を予測対象日の終値 $\geq$ 予測前日の終値の場合1、予測対象日の終値 $<$ 予測前日の終値の場合0とし、予測前日と予測対象日の終値を比較して、株価が上昇するか否かを予測した。

3 結果

提案手法の有効性を確認するためロイターIT・経済ニュース2015年7月~2021年3月の見出しを用いた。

予測対象を、2020年4月~2021年3月までのTOPIX-連動型上昇投資信託(ETF)の終値の上昇・下落とし、学習期間はそれぞれの予測日の直近の過去4年半を用いた。

複合極性辞書の作成はロイターニュース2015年7月~2020年3月までの見出しを用いた。複合極性辞書の作成では、まず金融専門極性辞書に存在する極性値の絶対値が0.01以上の単語からなる初期辞書を用いて抽出し

た複合エントリを元に複合極性辞書を作成し、その極性辞書を用いて再度複合エントリを抽出して追加した。

そして作成した複合極性辞書の中でも全見出しに対し、各見出しに出現する複合エントリ数の割合が5%以上かつ極性値の絶対値が0.005以上の複合エントリを抽出し、特徴語とした。

予測結果は混同行列を用いてAccuracy(正解率)、Precision(適合率)、Recall(再現率)、F値を求め、それぞれの結果を示した。

結果と考察は発表で述べる。

参考文献

- [1] 萩尾智彦, 佐野睦夫: ローソク足を学習させた畳み込みニューラルネットワークによる仮想通貨価格予測, 第28回金融情報学研究会, pp.51-55(2022).
- [2] 片寄諒介, 吉岡真治: 複数のテクニカル指標を用いた市場動向の予測, 第34回人工知能学会, 3Rin4-12(2020).
- [3] 和泉潔, 松井藤五郎: 新聞記事の時系列テキスト分析による株式市場の動向予測, 第30回人工知能学会, 3L3-OS-16a-6(2016).
- [4] 川崎拓海, 穴田一: 文章極性を考慮したテキスト分析による経済動向予測, 第84回情報処理学会, 4M-04(2022).
- [5] 東山昌彦, 乾健太郎, 松本裕治: 述語の選択選好性に着目した名詞評価極性の獲得, 言語処理学会第14回年次大会論文集, pp.584-587(2008).
- [6] Ito T., Sakaji H., Tsubouchi K., Izumi K., Yamashita T. Text-Visualizing Neural Network Model: Understanding Online Financial Textual Data. In: Phung D., Tseng V., Webb G., Ho B., Ganji M., Rashidi L. (eds) Advances in Knowledge Discovery and Data Mining, PAKDD 2018. Lecture Notes in Computer Science, Springer, vol 10939, pp 247-259(2018).
- [7] 那須川哲哉, 金山博: 文脈一貫性を利用した極性付評価表現の語彙獲得, 情報処理学会研究報告, pp.109-116(2004).
- [8] 石垣藍睦, 沼尾雅之: Twitterからの為替予測に特化したドメイン辞書構成法の提案, 第13回情報科学技術フォーラム, RO-001(2014).
- [9] 高村大也, 乾孝司, 奥村学: スピンモデルによる単語の感情極性抽出, 情報処理学会論文誌, Vol.46, No.2, pp.627-637(2006).