

文章極性を考慮したニューステキスト分析による経済動向予測

川崎 拓海^{†1} 穴田 一^{†1}

概要：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やファンダメンタル分析、数値情報を用いたテクニカル分析などによる様々な研究が行われている。しかし数値情報だけで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、世論に目を向けることを意味し、数値情報だけでは解釈が難しいマーケティングの予測を精度高く行える可能性があると考えられる。そこで本研究では、日本語評価極性辞書と金融専門極性辞書を用いた、文の肯定否定を考慮したニューステキスト感情分析による東証株価指数(TOPIX)の株価予測を提案する。

キーワード：株価予測、テキスト分析、勾配ブースティング決定木、テキストマイニング

Economic Trend Prediction by News Analysis Using Polarity Dictionaries

TAKUMI KAWASAKI^{†1} HAJIME ANADA^{†1}

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やファンダメンタル分析、数値情報を用いたテクニカル分析などによる様々な研究が行われている。その中でも数値情報だけでなくテキスト情報に含まれているニュース記事を考慮することは、世論に目を向けることを意味し、数値情報だけでは説明が難しい市場の予測を精度高く行える可能性があると考えられる。そこで本研究では、テキストマイニング手法を用いてニュース記事から株価の上昇・下落の予測を行った。テキストマイニング手法を用いた金融予測についても様々な研究が行われているが、本研究では、新聞記事の予測前営業日と予測当日のテキストを用いて株価の上昇下落を予測した和泉らの研究[1]を基に、日本語評価極性辞書[2][3]と金融専門極性辞書[4]を用いたニューステキスト分析による東証株価指数(TOPIX)の株価予測を提案し、その有効性を確認した。

2 提案

和泉らの既存研究の全体の平均正解率は 71.4%であるが、悪い年は 56.3%と不安定である。これは単語の出現数や出現パターンのみ考慮して、単語の印象を考慮していないことが要因であると考えられ、人に良い印象を与える単語が出現すると株価が上昇し、人に悪い印象を与える単語が出現すると株価が下落すると考えた。

そこで提案手法では極性単語を用いた、文の肯定否定と出現頻度を考慮した特徴量抽出を行った。

今回極性単語を扱うにあたり、日本語評価極性辞書と金融専門極性辞書を利用した。日本語評価極性辞書とは様々な用法や名詞に対し、ネガティブ・ポジティブの二値分類している辞書である。また、金融専門極性辞書とは金融専門単語についてネガティブ・ポジティブ度を極性値として表した辞書であり、ネガティブな単語ほど負の値が大きく、ポジティブな単語ほど正の値が大きい数値データで表されている。

この2つの極性辞書を用いて、感情分析ツールの1つである Oseti を用いて文の肯定否定を考慮した極性単語の抽出を行った。Oseti とは形態素解析ツール Mccab を用いて文章極性スコアを算出するものである。単語に“せず”や“ない”等の否定が掛かっている場合、その単語の極性を反転させスコアを求める。よって Positive(Negative)の極性単語に否定が掛かっている場合 Negative(Positive)とし、肯定否定を考慮した極性単語をそれぞれ抽出した。

本研究では IT・経済ニュースの記事に対して2つの辞書から得られる極性単語を用いたネガティブ・ポジティブ分析(以下ネガポジ分析とする)による経済動向予測を提案する。まず訓練データ内において1日に数件ずつ掲載されている IT・経済ニュースの見出しから、Oseti を用いて文の肯定否定を考慮した極性単語を抽出した。Oseti に用いる金融専門極性辞書の極性値 η は $\eta > \eta_{th}$ の場合 Positive, $\eta < -\eta_{th}$ の場合 Negative と分類された極性単語を抽出した。

^{†1} 東京都市大学大学院 総合理工学研究科
Graduate School of Tokyo City University

得られた極性辞書の単語が k 回以上出現した中から株価上昇割合 θ_1 以上、株価下落割合 θ_2 以上の単語を取り出し特徴語とした。取り出された l 個の特徴語に対し、訓練期間内のテキストに特徴語が生じている場合、特徴量を1、存在しない特徴語に関しては0とし、勾配ブースティング決定木に学習させた。

3 結果

提案手法の有意性を確認するためロイターニュースIT・経済ニュースの見出しを用いて、予測対象を、半年ごとに分けた2018年7月~2020年12月までのTOPIX-連動型上昇投資信託(ETF)とし、予測前営業日のニュースの見出しで上昇下落の予測を行った。訓練データの期間はそれぞれの予測日の直近の過去3年間を用いた。また、予測前日の終値と予測対象日の終値の差分をTOPIX-ETFの上昇・下落の基準とした。極性値の閾値を $\eta_{th} = 0.03$ とし、得られた極性単語で10回以上出現した単語の中から予測当日の株価の上昇割合 θ_1 が0.75以上と株価の下落割合 θ_2 が0.65以上のパターンを抽出し、特徴語として用いた。モデルのパラメータはグリットサーチを行い、最適なパラメータを選択した。

予測結果は表1の混同行列を用いて評価する

表1 混同行列の例

実際のクラス	Negative	TN(True Negative)	FN(False Positive)
	Positive	FP(False Negative)	TP(True Positive)
		Negative	Positive
		機械学習モデルの予測	

Trueは予測が正しくFalseは予測が正解のクラスと異なったことを表す。表1を元にAccuracy(正解率)やPrecision(適合率), Recall(再現率)を求め、グラフ化した結果を図1に示し、F値も求め、それぞれの結果を表2に示した。

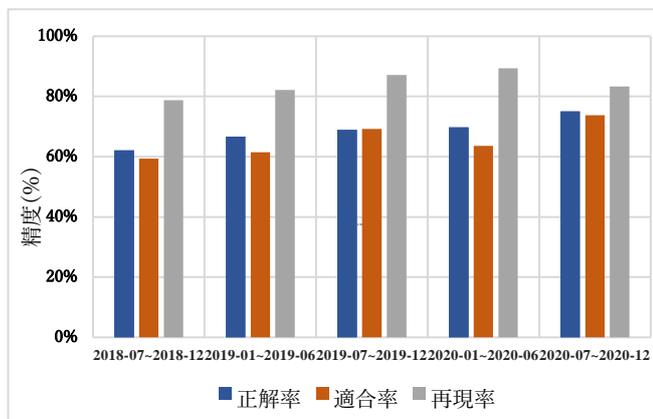


図1 混同行列を用いた結果

表2 混同行列を用いた結果

テスト期間	F 値	正解率	適合率	再現率
2018-07~2018-12	0.68	62.1%	59.4%	78.8%
2019-01~2019-06	0.67	66.7%	61.5%	82.2%
2019-07~2019-12	0.77	68.9%	69.2%	87.1%
2020-01~2020-06	0.74	69.8%	63.6%	89.4%
2020-07~2020-12	0.78	75.0%	73.8%	83.3%
全体の平均	0.73	67.5%	64.4%	84.2%

表2より、先行研究の全体の正解率が約70%に対し、提案手法の精度は低く、全体の正解率が67.5%という結果となった。しかし、悪い年の正解率でも約62%と既存手法より安定した結果を得ることができていた。

結果の詳細と考察は発表で述べる。

参考文献

- [1] 和泉凜, 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).
- [2] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一, 意見抽出のための評価表現の収集, 自然言語処理, Vol.12, No.3, pp.203-222, 2005. / Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi. Collecting Evaluative Expressions for Opinion Extraction, Journal of Natural Language Processing 12(3), 203-222 (2005).
- [3] 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刺評価極性の獲得, 言語処理学会第14回年次大会論文集, pp.584-587, 2008. / Masahiko Higashiyama, Kentaro Inui, Yuji Matsumoto. Learning Sentiment of Nouns from Selectional Preferences of Verbs and Adjectives, Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing, pp.584-587 (2008).
- [4] 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).
- [5] 中川裕志, 森辰則, 湯本紘彰: 出現頻度と接続頻度に基づく専門用語抽出, 自然言語処理, Vol. 10, No. 1, pp. 27-45 (2003).
- [6] 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刺評価極性の獲得, 言語処理学会第14回年次大会論文集, pp.584-587 (2008).