SESTM モデルによる会社四季報センチメントを用いた投資戦略の実証分析

指田 晋吾
野村アセットマネジメント株式会社
s-sashida@nomura-am.co.jp

中川 慧
野村アセットマネジメント株式会社
k-nakagawa@nomura-am.co.jp

1 はじめに

近年、機械学習技術の発展によるテキストマイニング技術の進展とともに、ファイナンス分野においても様々なテキスト情報が分析・利用されている。テキスト情報としてよく利用されるものは決算短信、有価証券報告書、新聞記事があり、これらは個々の文書の分量が多くかつ文書自体も大量にあるため、個人投資家だけではなく、証券、銀行、資産運用会社といった金融機関においても、金融テキストを自動的に解析し、業務支援や経済活動に役立てようと試みている [1, 2, 3]。とりわけ、これらのテキスト情報から抽出した情報と株価との関係が研究されており、テキストマイニングの投資における有効性が明らかになってきている [4]。例えば、[5] らは決算短信の要約から因果関係の表現を含む文を抽出し、構文パターンより因果関係を示すネットワーク (因果チェーン [6]) を構築し、そのノード間には株価のリード・ラグ効果があることを実証した。また、[7] らは決算短信とこれまで分析されていない情報媒体である会社四季報について、肯定語・否定語スコアを定義した辞書を用いて各テキストについてセンチメントスコアを算出し、それが投資に有効であることを確認している。さらに、[8] らは、テキスト内のセンチメント語を抽出する SESTM(Sentiment Extraction via Screening and Topic Modeling) モデルを構築し、そのモデルから算出されたセンチメントスコアにて非常に良好な株価リターンが得られることを確認している。本研究では、分析例が少ない会社四季報データを用いて日本株式市場において、テキスト情報を利用した投資戦略の収益機会が存在するかどうかを検証することを目的とする。分析手法としては、[7] らは四季報を対象に Bag of Words 方式を用いて共通の辞書を用いてセンチメントを付与しているのに対して、本研究では、より洗練された SESTM モデルによりセンチメントを付与し、イベントスタディ方式での実証分析を行う。

2 SESTM モデル

SESTM モデルは、[8] らがニュースからセンチメントを抽出する新しい手法を用いて構築されるモデルである。彼らが提案した手法は、データベンダーが提供するような一般的なセンチメントスコアなどで多く用いられる辞書ベースの手法とは異なり、株価リターンを直接教師として考慮するもので、リターン予測に特化している。提案手法は次の 3 つのステップから構成される。

Step 1: スクリーニング センチメントに影響を与える単語と与えない単語を分類

Step 2: センチメント評価 それら単語にトピックモデルを適用しセンチメントを評価

Step 3: スコアリング ペナルティ付きの尤度を用い文章単位のセンチメントスコアを算出

先行研究では米国株式市場を対象にダウ・ジョーンズ経済通信のニュース記事を用いて実証分析を行っており、提案手法の特徴について次の利点を挙げている。まず、計算負荷量が低いのでコンピュータリソースを多く必要としない。また、標準的な計量経済手法をベースとしており、シンプルな教師あり学習となっているので、深層学習等と比較しても解釈性が高い。そして、既存の辞書データ等は用いずに分析対象のテキストとそのテキストに対応するリターンデータのみでモデルを構築するため目的用途に特化したものを構築できることも利点として挙げている。

2.1 Notation

- n, m: 四季報の記事数、単語数
- $d_i \in \mathbb{R}_+^m$: i 番目の記事における単語数
- D: $D = [d_1, ..., d_n]^\top$ と定義される行列 $(n \times m)$

- S, N: センチメントに影響を及ぼす (及ぼさない) 単語の集合
- y_i: i 番目の記事に対するラベル (リターン)
- p_i: i 番目の記事のセンチメントスコア $p_i \in [0, 1]$
- O_+, O_-:各単語がポジティブ (ネガティブ) な確率を並べたもの (トピック)

2.2 モデルの設定

センチメント p_i で条件付けられたリターン y_i と d_i は (条件付き) 独立であると仮定する。さらに、リターン y_i とセンチメント p_i の関係は、$g(\cdot)$ を単調増加な関数として $P((y_i > 0|p_i) = g(p_i))$ と仮定する。すなわち、センチメント p_i が良いほど、リターンが正となる確率が高いと仮定する。また、関数 g の具体的な関数系は特に推定しない。最後に、m 個の単語は、それぞれセンチメントに影響を及ぼす単語 S と影響を及ぼさない単語 N に分割できるとし、それらをインデックスとした $d_{i,[S]}$ と $d_{i,[N]}$ は独立であると仮定する。さらに、センチメントが付与された単語数は、次のような多項分布に従うとする。

$$d_{i,[S]} \sim \text{Multinomial}(s_i, p_i O_+ + (1 - p_i) O_-)$$

ここで、SESTM モデルはパラメータである O_+, O_-, p_i を特定することが目的であり、これらが推定できれば、新たな記事 $i+1$ に対してセンチメント p_{i+1} を付与することができる。

2.3 パラメータの推定

SESTM モデルのパラメータである O_+, O_-, p_i の推定は次の 3 つのステップで行う。

Step 1:センチメントに影響を与える単語 S を抽出 m 内の単語 j ごとにスコア f_j を計算。

$$f_j = \frac{\text{単語 } j \text{ を含む記事数、かつ } \text{sgn}(y) = 1 \text{ の数}}{\text{単語 } j \text{ を含む記事数}}$$

次の回帰モデルを考える。

$$\text{sgn}(y_i) = f_j \cdot \text{sgn}(d_{i,j}) + \epsilon_i \quad j = 1...m$$

するとハイパーパラメータ $(\alpha_+, \alpha_-, \kappa)$ を用いて、次の $\hat{S}$ を S の推定量 (集合) とすることができる。

$$\hat{S} = \left\{ j : f_j \geq \frac{1}{2} + \alpha_+ \right\} \cup \left\{ j : f_j \leq \frac{1}{2} - \alpha_- \right\} \cap \{ j : \kappa_j \geq \kappa \}$$

Step 2:O_+, O_- の推定 p_i の推定量 $\hat{p}_i$ として次を計算する。

$$\hat{p}_i = \frac{\text{rank of } y_i \text{ in } \{y_l\}_{l=1}^n}{n}$$

以下の回帰式から、D を回帰することで $O = [O+, O-]$ を推定する。

$$\mathbb{E}\tilde{D}^\top = OW, \tag{1}$$

$$\text{where } W = \begin{bmatrix} p_1 & \dots & p_n \\ 1 - p_1 & \dots & 1 - p_n \end{bmatrix}, \tag{2}$$

$$\text{and } \tilde{D} = [\tilde{d}_1, ..., \tilde{d}_n]^\top \tag{3}$$

ここで、$\tilde{D}$ は記事 i ごとのすべてのセンチメントが付与された単語の頻度を表す。

Step 3:センチメントスコアを推定 パラメータ O_+, O_- が求まったので、新しい記事に対してセンチメントスコア p を次の最尤推定により求める。

$$\hat{p} = \arg \max_{p \in [0,1]} \hat{s}^{-1} \sum_{j=1}^{\hat{s}} d_j \log \left(p\hat{O}_{+,j} + (1 - p)\hat{O}_{-,j} \right) \tag{4}$$
$$+\lambda \log(p(1 - p))$$

s_i は S に属する単語の総数であり、$d_{i,[S]}$ は記事 i のセンチメントが付与された単語ベクトルであり観測できる。また、最尤推定においては、p_i が 0.5 をとりにくいようにペナルティ項 $\lambda \log(p(1 - p))$ が付与されている。

3 実証分析

3.1 データセット

会社四季報 (以下，四季報) は東洋経済新報社が 3 ヵ月に一度出版する発行物で、出版日は 3, 6, 9, 12 月の中旬である。四季報には、日本の株式市場の全上場企業について直近決算期の基本財務項目や投資関連指標が記載される。

また、数値以外のデータとして、会社四季報編集部による業績予想数字についての記事があり、そこには前期実績と比較した今期予想についてのコメントと、前号に掲載した予想や会社計画と比較した今期予想についてのコメントが記載される。この数値及びテキストデータは全上場企業について同様の形式でまとめられている。

本研究では 1999 年 09 月から 2020 年 06 月までに発行された四季報のテキストデータと紐づく企業のスペシフィック・リターンを用いてモデルの学習及び予測を行う。本研究では、スペシフィック・リターンとして Barra JPE4 モデルのファクターモデルから算出される数値を用いた。スペシフィック・リターンとは銘柄の騰落率 (トータル・リターン) から、市場全体の全銘柄に共通するあるファクター

(要因) で説明可能なリターンを取り除いた、各銘柄に固有の変動部分である。四季報データは各企業の業績について個別に記事が用意されているため、その効果の計測がしやすい銘柄固有のリターンであるスペシフィック・リターンを教師データとして用い、予測の対象とする。なお、[8] においてはスペシフィック・リターンではなく、株価リターンを用いているものの、SESTM モデルで仮定される前提はスペシフィック・リターンでも成立する。

3.2 分析手順

分析では四季報の出版日をイベントデーとするイベントスタディ [9] を行う。イベントスタディとは情報の公開等のイベントが、当該企業の市場価値にどのような影響を与えるかを分析する方法である。

モデル作成にあたり、学習データは各出版日から過去数年分の記事を利用し、教師データは出版日から翌出版日までのスペシフィックリターンの算術平均値を正規化したものを利用する。本研究では利用する過去データを 3, 4, 5 年分の 3 パターンのモデルを作成して検証した。学習には全銘柄のデータを利用し、予測する対象は TOPIX（東証株価指数）の時価総額上位 500 銘柄 (TOPIX500 の構成銘柄) とした。モデルは四季報の出版日のタイミングで作成され、その時点の四季報のテキストデータを用いて将来のスペシフィックリターンを予測する。次に予測値を 10 分位に分け、各グループについてイベントデー前後の平均スペシフィックリターン (前回出版日から翌出版日) を求める。そして、各基準日 (四季報の出版日) で得られた各分位の結果の平均を求め、その累和推移を算出する。このイベントスタディ分析にて四季報データにおける SESTM モデルの有効性を確認する。

3.3 分析結果

各学習モデルのイベントスタディの結果は表 (1) となった。累積日数は 57 営業日分である。どのモデルについても上位分位銘柄 (q1) の四季報出版日以降の平均スペシフィックリターンの累積和が低位分位銘柄 (q10) を上回っていることが分かる。イベント日以降の平均スペシフィックリターンについて、平均値が 0 と有意に異なるか t 検定 (両側検定) にて確認した。その結果、学習期間 4 年の上位の分位銘柄 (q1) について、1%水準で有意であった。学習期間 4 年のイベント日以降のスペシフィックリターンの累積和を表 (2)、イベント前後の累積和の推移を図 (1) に示す。図の横軸は営業日日数で座標 0 を四季報出版日 (イベントデー) としている。上位分位銘柄 (q1) の平均スペシフィックリターンは四季報出版日以降プラス方向に伸びていることが確認できる。以上の結果より、四季報データを用いた SESTM モデルの分析では、4 年分の学習データを用いたモデルが比較的良好な結果となり、検定結果からセンチメントがポジティブな銘柄については分類出来ているが、ネガティブな銘柄についての分類は上手く出来ていないことが確認出来る。この原因としては、[7] らの四季報の分析でも確認されている通り、四季報内の肯定語は否定語のおおよそ 2 倍を占めていることから、そのデータで学習したモデルの分位別の結果 (有意性) に差が生じたものと考えられる。

次に作成した SESTM モデルにおいて、単語毎にセンチメント評価を行い、学習期間におけるポジティブワードとネガティブワードの出現度について確認する。図 (2,3) は学習期間が 2005 年 9 月から 2020 年 3 月までの記事で学習させたモデルにて同期間の記事の単語でセンチメントを評価し、ポジティブ及びネガティブの単語出現数で単語の表示サイズを変化させたものである。結果を確認すると、図 (2) のポジティブ側の結果については"増益"、"堅調"、”利益"、”改善"など幾つかの肯定語の存在が確認出来る。一方、図 (3) のネガティブ側の結果については"減益"などの単語は確認出来るが否定語の数は多くないことが分かる。この違いは四季報の肯定語と否定語の比率に偏りがあるためだと考えられ、イベントスタディでの有意性の偏りについての原因考察とも整合する。

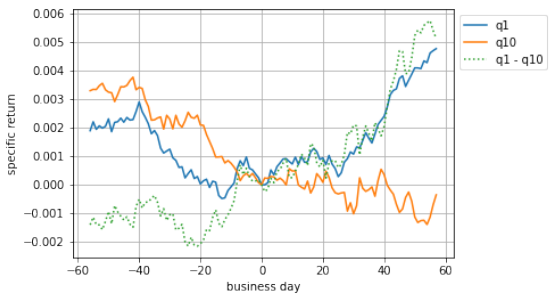

図 1 学習期間 4 年でのスペシフィックリターンの類和推移

4 まとめと今後の課題

本研究では会社四季報データを用いて SESTM モデルにてセンチメント評価を行い、イベント・スタ

表 1　各学習モデルのイベントデー以降の累積和リターン (%). 累積日数: 57 日. *は 10%有意、**は 5%有意、***は 1%有意を示す.

	学習期間		
分位	3 年	4 年	5 年
q1	0.32	0.48 ***	0.30
q10	0.03	-0.04	-0.07

表 2　学習期間 4 年のモデルでのイベントデー以降の累積和リターン (%) の推移

	分位	
累積日数	q1	q10
10	0.07	0.05
20	0.09	0.01
30	0.11	-0.10
40	0.24	0.04
50	0.41	-0.11
57	0.48	-0.04

ディによる有効性評価を行った。その結果、学習データで 4 年分の記事を用いたモデルにて比較的良好な結果を得ることが確認できた。分位分けした結果では高分位銘柄についてはスペシフィックリターンの有意性が確認出来たものの、低分位銘柄については確認出来なかった。これは四季報データ自体の肯定語の占める割合が多いことが起因していると考えられる。今後の課題としては、データ量の多い金融ニュースデータを利用した場合の分析や、比較的肯定後、否定語の割合に差がない決算短信のデータを用いた検証などが挙げられる。

図 2　2005 年 9 月から 5 年分の記事におけるポジティブワードの出現度

図 3　2005 年 9 月から 5 年分の記事におけるネガティブワードの出現度

参考文献

[1]高野海斗, 酒井浩之, 坂地泰紀, 和泉潔, 岡田奈奈, 水内利和. 株主招集通知における議案タイトルとその分類及び開始ページの推定システム. 自然言語処理, Vol. 25, No. 1, pp. 3–31, 2018.

[2]Tomoki Ito, Hiroki Sakaji, Kiyoshi Izumi, Kota Tsubouchi, and Tatsuo Yamashita. Ginn: gradient interpretable neural networks for visualizing financial texts. *International Journal of Data Science and Analytics*, Vol. 9, No. 4, pp. 431–445, 2020.

[3]高野海斗, 酒井浩之, 中川慧. 学習データの自動生成による深層学習を用いた株主招集通知の重要ページ抽出. 人工知能学会論文誌, Vol. 36, No. 1, pp. WI2–G_1 – –19, 2021.

[4]和泉潔, 坂地泰紀, 伊藤友貴, 伊藤諒. 金融テキストマイニングの最新技術動向 (特集 ai の金融応用 (実践編)). 証券アナリストジャーナル= Securities analysts journal, Vol. 55, No. 10, pp. 28–36, 2017.

[5]Nakagawa Kei, Sashida Shingo, Sakaji Hiroki, and Izumi Kiyoshi. Economic causal chain and predictable stock returns. In *2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 655–660. IEEE, 2019.

[6]Kiyoshi Izumi and Hiroki Sakaji. Economic causal-chain search using text mining technology. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pp. 61–65, 2019.

[7]山本零, 川代尚哉, 栗田昌孝. 決算短信と四季報テキスト情報の投資戦略への利用可能性検証. ジャフィー・ジャーナル, Vol. 18, pp. 46–62, 2020.

[8]Zheng Tracy Ke, Bryan T Kelly, and Dacheng Xiu. Predicting returns with text data. Technical report, National Bureau of Economic Research, 2019.

[9]John Y Campbell, John J Champbell, John W Campbell, Andrew W Lo, Andrew W Lo, and A Craig MacKinlay. *The econometrics of financial markets*. princeton University press, 1997.