トピック埋め込み回帰モデルを用いた株価予測

Predicting Stock Prices using Topic Embedding Regression Model

許 蔚然 [1*]　江口 浩二 [2†]
Weiran Xu[1]　Koji Eguchi[2]

[1] 神戸大学 大学院 システム情報学研究科
[1] Graduate School of System Informatics, Kobe University
[2] 広島大学 情報科学部
[2] School of Informatics and Data Science, Hiroshima University

Abstract: In this paper, we aim to predict stock prices by analyzing text data in financial articles. TopicVec is a topic embedding model that represents latent topics in a word embedding space. Here, word embedding maps words into a low-dimensional continuous embedding space by exploiting the local word collocation patterns in a small context window. On the other hand, topic modeling maps documents onto a low-dimensional topic space. Using the topic embedding model, topics underlying each document can be mapped into the word embedding space by combining word embedding and topic modeling.

The topic embedding model has not been used to address regression problem and also has not been used to predict stock prices by analyzing financial articles, to our knowledge. In this paper, by extending the topic embedding model to regression, we propose a topic embedding regression model called TopicVec-Reg to jointly model each document and a continuous label associated with the document. Our method takes financial articles as documents, each of which is associated with a stock price return as a continuous label, so that we can predict stock price returns for new unlabeled financial articles.

We evaluate the effectiveness of TopicVec-Reg through experiments in the task of stock return prediction using news articles provided by Thomson Reuters and stock prices by the Tokyo Stock Exchange. The result of closed test shows that our method brought meaningful improvement on prediction performance.

1　はじめに

金融市場で，人々が経済指標や企業に関するニュースや世の中の出来事などを考慮したうえで投資先を決定するのが一般的である．しかし，これらの情報は常に更新されていて，人の手ですべての情報を把握することは不可能であると考えられる．そこで，人々の負担軽減と意思決定を補佐するため，情報科学の手法でこれらの情報から株価などの金融指標を予測する研究が行われている．本稿では，金融記事のテキストデータを分析し，株価を予測する課題に取り組む．

テキストデータを分析するために，トピックモデルを使うのが一般的である．トピックモデルは文書コレクションに隠れた意味的な構造を発見する統計ツールであり，マーケティング，社会学，政治科学などの分野で応用されている．そのうち多くのトピックモデルは LDA(Latent Dirichlet Allocation) [1] という階層的な確率的モデルをベースとして拡張されてきた．LDA は，文書全体にわたる単語の共起パターンに従って，単語をトピックにグループ化する．コーパスが十分に大きい場合，共起パターンは単語間の意味的関連性をは反映するため，一貫性のあるトピックを発見できる．単語の出現確率はその背後にあると仮定されている潜在トピックによって決まる．そこで，単語は「one hot」（1 つの次元のみ 1 で他は 0 であるような語彙数次元ベクトル）で表現される．しかし，この単純な表現は高次元かつ非常にスパースであるため，与えられたコレクションが巨大な語彙量を持つと，LDA の性能が落ちてしまう [2].

*連絡先： 神戸大学 大学院 システム情報学研究科
〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: weland@cs25.scitec.kobe-u.ac.jp
†連絡先： 広島大学 情報科学部
〒 739-8527 東広島市鏡山 1-4-1
E-mail: eguchi@acm.org

単語の「one hot」表現を避けるために，単語埋め込み（word embedding）が提案された [3]．単語埋め込みは，単語の局所的な共起パターンを利用することにより，意味的又は構文的な関連性を符号化された単語の分散表現を学習できる．そのような分散表現を持つ単語のうち，似たような意味を持っていれば，埋め込み空間上で近い場所に位置する．

単語の集合である文書に着目し，各単語の埋め込みが与えられたとすると，文書を「bag-of-vectors」と見なすことができる．関連性のある単語は似た方向に向き，意味的なクラスタができてその重心を意味的重心と見なせる．特定のドメインの文書集合において，いくつかの意味的なクラスタが多くの文書の中に現れる場合があるため，それらを LDA のトピックと見なすことができる．ただし，意味的クラスタの重心は埋め込み空間に存在する．この考え方に基づいて，生成的単語埋め込みモデル PSDVec [7] にトピックを加えることで，トピック埋め込みモデル TopicVec [6] が提案された．TopicVec では，各単語の生成確率は局所的な文脈と大域的なトピックに影響され，トピック埋め込みベクトルは意味的重心として単語埋め込みと同じ埋め込み空間上に存在する．LDA と同様に Dirichlet 事前分布を用いてトピック分布を正規化する．単語の生成においては LDA で使われるカテゴリカル分布の代わりに，PSDVec でのリンク関数にトピックを加えたものを使うと仮定する．これによって，単語間の意味的関連性は埋め込み空間内に余弦類似度として符号化される．TopicVec は多クラステキスト分類タスクにおいて，LDA より優れた性能をもたらすことが報告されている [6]．

TopicVec のようなトピック埋め込みモデルを回帰問題に用いた研究や，それにより金融記事記事のテキストデータから株価を予測する研究は我々の知る限り未だ行われていない．そこで，本稿は，テキストデータとそれに対応する連続値ラベルとの関係をモデル化する回帰機能を付与したトピック埋め込みモデルとして TopicVec-Reg を提案する．TopicVec-Reg での回帰機能は，Supervised Latent Dirichlet Allocation（sLDA）[4] に倣い，1 文書に対応する連続値ラベルをガウス分布に従うと仮定する．さらに，ある企業に関する金融記事をテキストデータ，その企業に対応する株価リターン率を連続値ラベルと見なすことで，金融記事から株価リターン率を予測する．学習過程は変分推論により，回帰係数と潜在トピックを同時に学習する．

本稿の提案手法 TopicVec-Reg の予測性能を評価するために，回帰係数を考慮しない TopicVec を比較対象として，各記事が発信された前後の株価リターン率を予測するクローズドテストを行う．

2 関連研究

Das ら [5] は LDA を単語埋め込みを利用するように拡張し，GaussianLDA を提案した．この手法は，離散的なテキストを連続的な埋め込みの観測に変換し，LDA を実数値データを生成するように適合させる．トピック内の単語はトピック埋め込みを期待値パラメータとする多変量ガウス分布からランダムにサンプリングされると仮定される．しかし，埋め込み間の意味的関連性を測るにはユークリッド距離が最適な選択肢とは限らないことから，これはやや強い仮定であると考えられる．

Dieng ら [2] は LDA と単語埋め込みを結合した生成的モデル ETM を提案した．単語はカテゴリカル分布から生成されると仮定される．そのカテゴリカル分布のパラメータは単語埋め込みと与えられたトピック埋め込みの内積である．ETM は巨大な文書コーパスにおいても解釈可能性のあるトピックを学習できる．さらに，文書データのストップワードを除外しなくても，良いトピック品質を保つことができる．前述の TopicVec は ETM と同じ発想を持っているが，単語はリンク関数によって生成されると仮定される．

Li ら [7] は TopicVec の前身である生成的単語埋め込み方法 PSDVec を提案した．PSDVec は，注目単語と近くにある文脈単語との共起パターンを用いて注目単語を低次元連続的埋め込み空間に写像する．文脈が与えられた単語の条件付き分布は，独立した対数双線形項に因子分解できると仮定される．単語埋め込みと回帰残差をガウス事前分布を仮定して正規化することによって過適合を防ぐ．TopicVec が PSDVec と異なる点は，TopicVec での単語の条件付き分布は，文脈単語だけではなく，トピック埋め込みによる影響も受けることにある．

Blei ら [4] はラベル付けの文書を対象とする教師ありトピックモデル sLDA を提案した．sLDA を LDA に各文書に紐づける応答変数を付け加えたものであると考えられる．応答変数は一般化線形モデルで生成されると仮定され，文書に割り当てられたトピックの期待値をその線形回帰の説明変数とする．文書と応答変数を同時にモデル化することで，新たに与えられた文書の応答変数を良く予測できる潜在トピックを推定できると期待される．この考え方に従い，本稿はトピック埋め込みモデル TopicVec に線形回帰モデルを付与することで，トピック埋め込み回帰モデル TopicVec-reg を提案する．

3　トピック埋め込み回帰モデル

本稿において，語彙サイズが V である D 文書のコーパスを考える．文書 d_i の j 番目の単語を $w_{ij} \in \{1, \cdots, V\}$ で表す．

3.1　TopicVec

本節では文献 [6] に示された TopicVec の概要を示す．まず，前提となる単語埋め込みモデル PSDVec [7] のリンク関数は着目する単語の埋め込み $\boldsymbol{v}_{w_c}$ と文脈単語の埋め込み $\boldsymbol{v}_{w_l}$ を接続する：

$$\begin{aligned} &P\left(w_c | w_0 : w_{c-1}\right) \\ \approx &P\left(w_c\right) \exp\left\{ \boldsymbol{v}_{w_c}^\top \sum_{l=0}^{c-1} \boldsymbol{v}_{w_l} + \sum_{l=0}^{c-1} a_{w_l w_c} \right\}. \end{aligned} \tag{1}$$

ここで，$a_{w_l w_c}$ はバイグラム残差（バイアス）と呼ばれ，$\boldsymbol{v}_{w_c}^\top \boldsymbol{v}_{w_l}$ が捕らえられない非線形部分を示す．

本稿の提案手法のベースである TopicVec の単語の条件付き分布は，PSDVec に基づいて，トピックを文脈単語と見なす：

$$\begin{aligned} &P\left(w_c | w_0 : w_{c-1}, z_c, d_i\right) \\ \approx &P\left(w_c\right) \exp\left\{ \boldsymbol{v}_{w_c}^\top \left(\sum_{l=0}^{c-1} \boldsymbol{v}_{w_l} + \boldsymbol{t}_{z_c} \right) \right. \\ &\left. + \sum_{l=0}^{c-1} a_{w_l w_c} + r_{z_c} \right\}. \end{aligned} \tag{2}$$

ここで，$\boldsymbol{t}_{z_c}$ は注目単語に割り当てられたトピックの埋め込みであり，r_{z_c} はトピックに関する残差である．

TopicVec の生成過程において，単語埋め込み $\boldsymbol{v}_{s_i}$ と残差 $a_{s_i s_j}$ はそれぞれガウス分布から抽出されると仮定される．簡単のために，それらの生成過程を省略し，トピック埋め込みの生成過程は次のようである：

1. トピック k について，半径 γ の超球から一様にトピック埋め込みを抽出する．すなわち $\boldsymbol{t}_k \sim \mathrm{Unif}(B_\gamma)$；
2. 各文書 d_i について:
   (a) ディリクレ事前確率 $\mathrm{Dir}(\boldsymbol{\alpha})$ から混合比率 $\boldsymbol{\phi}_i$ を抽出する；
   (b) j 番目の単語について:
      i. カテゴリカル分布 $\mathrm{Cat}(\boldsymbol{\phi}_i)$ からトピック割り当て z_{ij} を抽出する．
      ii. $P\left(w_{ij} | w_{i,j-c} : w_{i,j-1}, z_{ij}, d_i\right)$ に従って単語 w_{ij} を語彙集合 $\boldsymbol{S}$ から抽出する．

文書データ $\boldsymbol{D}$，単語埋め込み $\boldsymbol{V}$，バイグラム残差 $\boldsymbol{A}$，トピック埋め込み $\boldsymbol{T}$，トピック割り当て $\boldsymbol{Z}$，トピック分布 ϕ に関する対数尤度は以下のようになる：

$$\begin{aligned} &\log p(\boldsymbol{D}, \boldsymbol{A}, \boldsymbol{V}, \boldsymbol{Z}, \boldsymbol{T}, \boldsymbol{\phi} | \boldsymbol{\alpha}, \gamma, \boldsymbol{\mu}) \\ =&C_0 - \log \mathcal{Z}(\boldsymbol{H}, \boldsymbol{\mu}) - \|\boldsymbol{A}\|_{f(\boldsymbol{H})}^2 - \sum_{i=1}^{W} \mu_i \left\|\boldsymbol{v}_{s_i}\right\|^2 \\ &+ \sum_{i=1}^{M} \left\{ \sum_{k=1}^{K} \log \phi_{ik} \left(m_{ik} + \alpha_k - 1\right) \right. \\ &+ \sum_{j=1}^{L_i} \left(r_{i,z_{ij}} + \boldsymbol{v}_{w_{ij}}^\top \left(\sum_{l=j-c}^{j-1} \boldsymbol{v}_{w_{il}} + \boldsymbol{t}_{z_{ij}} \right) \right. \\ &\left. \left. + \sum_{l=j-c}^{j-1} a_{w_{il} w_{ij}} \right) \right\}. \end{aligned} \tag{3}$$

対数尤度を最大化するために，ハイパーパラメータ $\boldsymbol{\alpha}, \gamma, \boldsymbol{\mu}$ が与えられたとし，最適な $\boldsymbol{V}, \boldsymbol{T}, p(\boldsymbol{Z}, \boldsymbol{\phi} | \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{V}, \boldsymbol{T})$ を推定する：

step1 PSDVec で $\boldsymbol{V}^*, \boldsymbol{A}^*$ を最適化する

step2 $\boldsymbol{V}^*, \boldsymbol{A}^*$ を対数尤度関数に代入し，$\boldsymbol{T}^*, p(\boldsymbol{Z}, \boldsymbol{\phi} | \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{V}, \boldsymbol{T})$ を最適化する

step2 で事後分布を解析的に計算するのは不可能であるため，変分分布 $q(\boldsymbol{Z}, \boldsymbol{\phi}; \boldsymbol{\pi}, \boldsymbol{\theta}) = q(\boldsymbol{\phi}; \boldsymbol{\theta}) q(\boldsymbol{Z}; \boldsymbol{\pi})$ で近似する．ここで，KL ダイバージェンスを導入し，推定タスクを変分下限 $\mathcal{L}(q, \boldsymbol{T})$ の最大化に置き換える：

$$\begin{aligned} &\mathrm{KL}(q \| p) \\ =&\log p(\boldsymbol{D} | \boldsymbol{T}) - \left(E_q[\log p(\boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{\phi} | \boldsymbol{T})] + \mathcal{H}(q)\right) \\ =&\log p(\boldsymbol{D} | \boldsymbol{T}) - \mathcal{L}(q, \boldsymbol{T}) \end{aligned} \tag{4}$$

ここで $p = p(\boldsymbol{Z}, \boldsymbol{\phi} | \boldsymbol{D}, \boldsymbol{T})$，$\mathcal{H}(q)$ は q のエントロピーである．変分下限 $\mathcal{L}(q, \boldsymbol{T})$ は

$$\begin{aligned} &\mathcal{L}(q, \boldsymbol{T}) \\ =&\sum_{i=1}^{M} \left\{ \sum_{k=1}^{K} \left(\sum_{j=1}^{K} \pi_{ij}^k + \alpha_k - 1 \right) \left(\psi\left(\theta_{ik}\right) - \psi\left(\theta_{i0}\right)\right) \right. \\ &\left. + \mathrm{Tr}\left(\boldsymbol{T}_i^\top \sum_{j=1}^{L_i} \boldsymbol{v}_{w_{ij}} \boldsymbol{\pi}_{ij}^\top \right) + \boldsymbol{r}_i^\top \sum_{j=1}^{L_i} \boldsymbol{\pi}_{ij} \right\} \\ &+ \mathcal{H}(q) + C_1 \end{aligned} \tag{5}$$

であり，$\mathcal{L}(q, \boldsymbol{T})$ を最大化する最適な q と $\boldsymbol{T}$ を見つけるため，GEM（Generalized Expectation-Maximization）アルゴリズムを用いる：

E-Step：

$$\pi_{ij}^{k} \propto \exp\left\{\psi\left(\theta_{ik}\right) + \boldsymbol{v}_{w_{ij}}^{\top}\boldsymbol{t}_{ik} + r_{ik}\right\}, \tag{6}$$

$$\theta_{ik} = \sum_{j=1}^{L_i} \pi_{ij}^{k} + \alpha_k. \tag{7}$$

M-Step：

$$\boldsymbol{T}_i^{(l)} = \boldsymbol{T}^{(l-1)} + \lambda\left(l, L_i\right) \frac{\partial \mathcal{L}\left(q^{(l)}, \boldsymbol{T}\right)}{\partial \boldsymbol{T}_i}. \tag{8}$$

推定過程のより詳細については文献 [6] に参照されたい．

3.2 TopicVec-Reg

本節では，前節で述べた TopicVec に回帰機能を加えた提案モデル TopicVec-Reg について述べる．TocpicVec-Reg では，各文書に連続値ラベルが紐づけられ，そのラベルはガウス分布に従ってサンプリングされると仮定する．

TopicVec における文書 d_i の生成過程に，以下のような連続値ラベル（応答変数）の生成過程を加える：

3. 連続値ラベル $y|z_{i1:iN}, \boldsymbol{\eta}, \delta^2 \sim \mathrm{N}(\boldsymbol{\eta}^{\top}\bar{\boldsymbol{z}}, \delta^2)$.

ここで用いられるガウス分布の平均は回帰係数 $\boldsymbol{\eta}$ と説明変数であるその文書のトピックの期待値 $\bar{z}$ の内積である．

文書における各単語にトピックが与えられたとき，y の対数尤度の期待値は以下のようになる：

$$\begin{aligned}
&\mathrm{E}\left[\log p\left(y|\boldsymbol{Z}_{1:L_i}, \eta, \delta^2\right)\right] \\
=& -\frac{1}{2}\log\left(2\pi\delta^2\right) \\
&-\frac{1}{2\delta^2}\left(y^2 - 2y\boldsymbol{\eta}^{\top}\mathrm{E}[\bar{\boldsymbol{Z}}] + \boldsymbol{\eta}^{\top}\mathrm{E}\left[\bar{\boldsymbol{Z}}\bar{\boldsymbol{Z}}^{\top}\right]\boldsymbol{\eta}\right)
\end{aligned} \tag{9}$$

ここで，

$$\mathrm{E}[\bar{\boldsymbol{Z}}] = \bar{\boldsymbol{\pi}}_i := \frac{1}{L_i}\sum_{j=1}^{L_i}\boldsymbol{\pi}_{ij}, \tag{10}$$

$$\begin{aligned}
&\mathrm{E}\left[\bar{\boldsymbol{Z}}\bar{\boldsymbol{Z}}^{\top}\right] \\
=&\frac{1}{{L_i}^2}\left(\sum_{j=1}^{L_i}\sum_{m\neq j}\boldsymbol{\pi}_{ij}\boldsymbol{\pi}_{im}^{\top} + \sum_{j=1}^{L_i}\mathrm{diag}\left\{\boldsymbol{\pi}_j\right\}\right)
\end{aligned} \tag{11}$$

である．

TopicVec をベースにしているため，TopicVec-Reg の推定過程において，目的関数 $\mathcal{L}(q, \boldsymbol{T})$ は (9) 式を (5) 式に加えることで得られる：

$$\begin{aligned}
&\mathcal{L}(q, \boldsymbol{T}) \\
=&\sum_{i=1}^{M}\left\{\sum_{k=1}^{K}\left(\sum_{j=1}^{L_i}\boldsymbol{\pi}_{ij}^{k} + \alpha_k - 1\right)\left(\psi\left(\theta_{ik}\right) - \psi\left(\theta_{i0}\right)\right)\right. \\
&+\left(-\frac{1}{2}\log(2\pi\delta^2) - \frac{y_i^2}{2\delta^2}\right) \\
&+\mathrm{Tr}\left(T_i^{\top}\sum_{j=1}^{L_i}\boldsymbol{v}_{w_{ij}}\boldsymbol{\pi}_{ij}^{\top}\right) + \left(\boldsymbol{r}_i^{\top} + \frac{y_i\boldsymbol{\eta}^{\top}}{L_i\delta^2}\right)\sum_{j=1}^{L_i}\boldsymbol{\pi}_{ij} \\
&\left.+\left(-\boldsymbol{\eta}^{\top}\frac{1}{{L_i}^2\delta^2}\left(\sum_{j=1}^{L_i}\sum_{m\neq j}^{L_i}\boldsymbol{\pi}_{ij}\boldsymbol{\pi}_{im}^{\top} + \sum_{j=1}^{L_i}\mathrm{diag}\{\boldsymbol{\pi}_{ij}\}\right)\boldsymbol{\eta}\right)\right\} \\
&+\mathcal{H}(q) + C_1
\end{aligned} \tag{12}$$

π_{ij}^{k} の更新式を求めるには，$\mathcal{L}(q, \boldsymbol{T})$ から j 番目の単語がトピック k に割り当てられる確率 π_{ij}^{k} を含む項を抽出し，π_{ij}^{k} に関する偏微分を解く．更新式は以下のようになる：

$$\begin{aligned}
\pi_{ij}^{k} \propto \exp\Bigg\{&\psi(\theta_{ik}) + \boldsymbol{v}_{w_{ij}}^{\top}\boldsymbol{t}_{ik} + r_{ik} + \frac{y_i\eta^k}{L_i\delta^2} \\
&- \frac{\boldsymbol{\eta}^{\top}\boldsymbol{\Pi}_{i,-j}^{(k)}\boldsymbol{\eta} + (\eta^k)^2}{2{L_i}^2\delta^2}\Bigg\}.
\end{aligned} \tag{13}$$

ここで，$\boldsymbol{\Pi}_{i,-j}^{(k)}$ は $\sum_{j=1}^{L_i}\sum_{m\neq j}^{L_i}\boldsymbol{\pi}_{ij}\boldsymbol{\pi}_{im}^{\top}$ を π_{ij}^{k} に関して偏微分したものであり，

$$\begin{aligned}
\boldsymbol{\Pi}_{i,-j}^{(k)} :=& \sum_{m\neq j}^{L_i}\boldsymbol{\Pi}_{im}\,\mathrm{diag}\{0^{(1)}, \cdots, 1^{(k)}, \cdots, 0^{(K)}\} \\
&+\mathrm{diag}\{0^{(1)}, \cdots, 1^{(k)}, \cdots, 0^{(K)}\}\sum_{m\neq j}^{L_i}\boldsymbol{\Pi}_{im}
\end{aligned}$$

と定義する．$\boldsymbol{\Pi}_{im}$ は $K \times K$ の行列であり，各行は $(\pi_{im}^{1}, \pi_{im}^{2}, \cdots, \pi_{im}^{K})$ である．

θ_{ik} は（7）式の通りに更新する．

$\mathcal{L}(q, \boldsymbol{T}^{(l-1)})$ のそれぞれ $\boldsymbol{\eta}$ と δ^2 に関する偏微分を解くと，sLDA [4] と同じく $\boldsymbol{\eta}$ と δ^2 の更新式を導出できる：

$$\hat{\boldsymbol{\eta}}_{\mathrm{new}} \leftarrow \left(\mathrm{E}\left[A^{\top}A\right]\right)^{-1}\mathrm{E}\left[A\right]^{\top}\boldsymbol{y} \tag{14}$$

$$\hat{\delta}^2_{\mathrm{new}} \leftarrow \frac{1}{D}\left\{\boldsymbol{y}^{\top}\boldsymbol{y} - \boldsymbol{y}^{\top}\mathrm{E}[A]\left(\mathrm{E}\left[A^{\top}A\right]\right)^{-1}\mathrm{E}[A]^{\top}\boldsymbol{y}\right\} \tag{15}$$

ここで,

$$\mathrm{E}\left[A\right] = \frac{1}{L_i}\sum_{j=1}^{L_i}\boldsymbol{\pi}_{ij}$$

$$\mathrm{E}\left[A^\top A\right] = \sum_{i=1}^{D}\left(\frac{1}{{L_i}^2}\left(\sum_{j=1}^{L_i}\sum_{m\neq j}^{L_i}\boldsymbol{\pi}_{ij}\boldsymbol{\pi}_{im}^\top + \sum_{j=1}^{L_i}\mathrm{diag}\{\boldsymbol{\pi}_{ij}\}\right)\right)$$

である．A は各行が $\bar{\boldsymbol{Z}}_d^\top$ である $D\times(K+1)$ の行列である．$K+1$ 列目は回帰係数のバイアス項に対応する．

$\boldsymbol{T}$ は（8）式の通りに勾配降下法で更新する．

4 評価実験

TopicVec-Reg の予測性能を評価するために，TopicVec を比較対象として，実験を行った．

4.1 データセット

一定とされる単語埋め込み $\boldsymbol{V}^*$ と単語埋め込み残差 $\boldsymbol{A}^*$ は日本語版 Wikipedia を用いた PSDVec で学習済みのものにする．入力されるテキストデータとして，トムソン・ロイター社が 2013 年から 2017 年に発信した日本語金融記事を用いる．各記事に紐づけられた連続値ラベルは株価リターン率を，

$$\frac{\text{記事が出た翌日の終値} - \text{記事が出た前日の終値}}{\text{記事が出た前日の終値}}$$

と設定する．株価は東京証券取引所の歩み値データを利用する．

2ヶ月分の記事を学習用データとする．このようなデータを 3 セットを用意し，それぞれ独立で実験を行った．データセットの概要を表 1 に示す．

表 1: データセット

	data
set1	Nov. 2016~Dec. 2016 # of documents: 931
set2	Jan. 2017~Feb. 2017 # of documents: 912
set3	Mar. 2017~Apr. 2017 # of documents: 845

データに対して前処理を行った．テキストデータに対して，記事の中にある扱いにくい表や必要のない表現の除去した後，MeCab と新語や固有表現を取り揃えた辞書 mecab-ipadic-NEologd [8] [9] [10] を用いて形態素解析を行い，助詞や接続詞などのストップワードを「*」に置き換えた．記事の中に複数の企業が言及された場合，記事を複製してリターン率の大きい企業に残した．最後に，5 文書未満しか出現しない低頻度語と 50 単語未満の文書を除去した．

4.2 実験設定

モデル

TopicVec-Reg: 回帰パラメータの学習をトピック推定と同時に行う

TopicVec+LR: TopicVec でトピック推定した後に線形回帰を行う

TopicVec+LR を比較対象として TopicVec-Reg の評価実験を行った．TopicVec+LR では，トピック推定を TopicVec で行った後，ライブラリ scikit-learn の線形回帰モデル Linear Regression で回帰係数を推定する．

まず学習用データでトピック数が $K=\{5,10,15,20,25\}$ の 5 通りのモデル構築を行い，トピック埋め込み $\boldsymbol{T}$ と回帰係数 $\boldsymbol{\eta}$ を得る．

学習開始時の $\boldsymbol{T}$ の初期値はガウス分布に従う乱数で与え，変分パラメータ $\boldsymbol{\pi}$ に対しては一様分布に従う乱数を与えた．ハイパーパラメータ $\boldsymbol{\alpha}=(0.1,\cdots,0.1)$ とする．

テストを行うときに以下の式に従い連続値ラベルを予測する：

$$\boldsymbol{\pi}'_{ij} = (\boldsymbol{\pi}_{ij}, 1)$$

$$\hat{y}_i = \boldsymbol{\eta}_{0:K-1}^\top \mathrm{E}_q[\bar{Z}_i] + \eta_K = \boldsymbol{\eta}_{0:K}^\top \frac{1}{L_i}\sum_{j=1}^{L_i}\boldsymbol{\pi}'_{ij} \qquad (16)$$

モデル予測性能の評価指標をラベルの真値と予測値との間の平均二乗誤差 (Mean Squared Error : MSE) を用いる：

$$MSE = \frac{1}{D}\sum_{i=1}^{D}(y_i-\hat{y}_i)^2.$$

すべての実験は MSE の変化率が 0.1%以下になるときに収束と判断され終了する．

4.3 実験結果

本実験では，学習データを学習するときに，5 回おきに得られた $\boldsymbol{\pi}$ と $\boldsymbol{\eta}$ を用いて MSE を計算した．各条件における TopicVec-Reg の性能を同じ条件の TopicVec+LR と比較した．

3 つのデータセットにおける TopicVec-Reg と TopicVec-LR を用いたそれぞれ 15 通りの実験に対して MSE を算出し，さらに同じ条件における両手法の予測値に対して

Wilcoxon 符号付き順位検定を行った．実験結果として得られた MSE と p 値を表 2 に示す．太字で TopicVec-Reg が TopicVec+LR より良い予測精度を示す．太字の p 値が 5%を下回っており有意な差を示す．

表 2: 実験結果

K		set1	set2	set3
5	TopicVec-Reg	**0.016637**	**0.017398**	0.016179
	TopicVec+LR	0.016867	0.017730	0.016107
	p-value	**0**	**0**	0
10	TopicVec-Reg	0.016343	**0.015873**	**0.015685**
	TopicVec+LR	0.016220	0.016879	0.015839
	p-value	0.870397	**0.000693**	0.086139
15	TopicVec-Reg	**0.015873**	**0.015297**	**0.014299**
	TopicVec+LR	0.016273	0.016803	0.015616
	p-value	**0.020224**	**0.038315**	0.1617699
20	TopicVec-Reg	**0.013736**	**0.015790**	**0.013564**
	TopicVec+LR	0.016453	0.016785	0.014910
	p-value	**0.000565**	0.486916	**0.016891**
25	TopicVec-Reg	**0.013447**	**0.014533**	**0.012280**
	TopicVec+LR	0.015963	0.016797	0.014726
	p-value	**0.044846**	0.492524	**0.001298**

5 むすび

本稿では，金融記事から企業の株価上昇率を予測する課題に取り組み，トピック埋め込みモデル TopicVec に回帰機能を付与することで，トピック埋め込み回帰モデル TopicVec-Reg を提案した．

TopicVec-Reg では潜在トピックと回帰パラメータを同時に学習することができ，実験により潜在トピックを推定した後に回帰パラメータを推定するモデルより有意な改善を得た．

本稿ではクローズドテストにより評価を行ったが，オープンテストは今後の課題である．また，潜在トピックだけではなく，単語埋め込みを活用できるように説明変数を増やして回帰性能のさらなる改善を目指したい．

謝辞

本研究の一部は科学研究費補助金基盤研究 (B)（15H02703）の援助による．

参考文献

[1] David M Blei, Andrew Y Ng, and Michael I Jordan, ”Latent Dirichlet Allocation”, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

[2] Dieng, Adji B., Francisco JR Ruiz, and David M. Blei. ”Topic modeling in embedding spaces.” *arXiv preprint arXiv:1907.04907* (2019)

[3] Mikolov, Tomas, et al. ”Distributed representations of words and phrases and their compositionality.” *Advances in neural information processing systems.* (2013)

[4] David M Blei and Jon D. McAuliffe, ”Supervised topic models”, *Advances in neural information processing systems*, pp. 121–128 (2008)

[5] Das, Rajarshi, Manzil Zaheer, and Chris Dyer. ”Gaussian lda for topic models with word embeddings.” *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (2015)

[6] Li, Shaohua, et al. ”Generative topic embedding: a continuous representation of documents (extended version with proofs).” *arXiv preprint arXiv:1606.02979* (2016)

[7] Li, Shaohua, Jun Zhu, and Chunyan Miao. ”A generative word embedding model and its low rank positive semidefinite solution.” *arXiv preprint arXiv:1508.03826* (2015)

[8] 佐藤敏紀, 橋本泰一, 奥村学, ”単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討”, 言語処理学会第 23 回年次大会 (NLP2017), NLP2017-B6-1 (2017)

[9] 佐藤敏紀, 橋本泰一, 奥村学, ”単語分かち書き用辞書生成システム NEologd の運用 – 文書分類を例にして –”, 自然言語処理研究会研究報告, NL-229-15 (2016)

[10] Toshinori, Sato, ”Neologism dictionary based on the language resources on the Web for Mecab”(2015)