Topic Embedding Regression Model and its Application to Financial Texts

Weiran Xu[1] , **Koji Eguchi**[1]

[1]Graduate School of Advanced Science and Engineering, Hiroshima University
d210104@hiroshima-u.ac.jp, kxeguchi@hiroshima-u.ac.jp

Abstract

In this paper, we aim to predict stock price return rates by analyzing text data in financial news articles. A promising text analysis technique is word embedding that maps words into a low-dimensional continuous embedding space by exploiting the local word collocation patterns in a small context window. Another means of analyzing text is topic modeling that maps each document into a low-dimensional topic space. Recently developed topic embedding takes advantage of those two approaches by modeling latent topics of each document in a word embedding space. In this paper, by incorporating regression into the topic embedding model, we propose a topic embedding regression model called TopicVec-Reg to jointly model each document and a response variable associated with the document. Moreover, our method predicts the stock price return rate for unseen unlabeled financial articles. We evaluated the effectiveness of TopicVec-Reg through experiments in the task of stock return rate prediction using news articles provided by Thomson Reuters and stock prices by the Tokyo Stock Exchange. The result of closed test experiments showed that our method brought meaningful improvement on prediction performance in comparison to performing linear regression as post-processing of TopicVec. Through an open test, our method showed better prediction accuracy with a statistically significant difference.

1 Introduction

In financial markets, people generally make decisions about where to invest after taking into account economic indicators, news about companies, and events in the world. However, with the development of information technology, a large amount of text data is being transmitted every day, and it is practically difficult to keep track of all the information in the domain of interest. Therefore, as a means to support people's decision-making, some research has been conducted to predict financial indicators such as stock prices from a large amount of text data using machine learning techniques. In this paper, we analyze financial news articles that reflect business sentiment and corporate activities in particular and tackle the problem of predicting stock prices related to them.

As a tool for analyzing large amounts of text data, it is common to use topic models. The topic models are statistical machine learning models that discover semantic structures hidden in a collection of documents and have been applied in various fields. A representative topic model is Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003]. LDA assumes a latent variable that indicates the topic behind each word in a document and analyzes what topics the document is composed of by estimating the latent variable. If the size of the corpus is large enough, the co-occurrence patterns of words reflect the semantic relatedness between words, and thus appropriate topics can be found. By the way, word expressions need to be replaced in advance from the original text format to a numerical format that can be handled by machine learning. In the case of LDA, we first prepare a vocabulary, assign IDs to all word types, and represent each word in a ”one-hot” representation. However, the ”one-hot” representations suffer from problems such as high dimensionality and sparsity. Dieng et al. [2020] reported that as the size of the corpus increases, the quality of LDA's topics decreases. One way to eschew the problem of ”one-hot” representation is word embedding. Word embedding uses local co-occurrence patterns of words to learn vector representations that take into account the relevance between words. In other words, the method embeds a vocabulary of more than tens of thousands in a low-dimensional vector space. Word embedding has the property that words with similar meanings are closely mapped in the vector space.

In this paper, we focus on a topic embedding that gives a smooth combination of the topic model and the word embedding. Topic embedding discovers the semantic structure of each document with the words represented by word embeddings. Also, latent variables related to the topics are represented in the word embedding space. Based on this idea, a topic embedding model TopicVec [Li *et al.*, 2016] was proposed by adding topics to a generative word embedding model PSDVec [Li *et al.*, 2015]. In TopicVec, similar to LDA, the topic distributions are assumed to be regularized with Dirichlet priors. Also each word in a document is assumed to be drawn from a link function that takes local context and global topics into account. Topic embedding has

Proceedings of the Third Workshop on Financial Technology and Natural Language Processing
(FinNLP@IJCAI 2021), pages 15-21, Online, August 19, 2021.

been reported to outperform LDA in topic quality, handling OOV, and tasks such as document classification.

To the best of our knowledge, there is no research on using topic embedding models such as TopicVec for regression problems, or on using them to predict stock prices from text data of financial articles. In this paper, we propose TopicVec-Reg as a topic embedding model with a regression function to model the relationship between each document and a response variable associated with it. The parameters including the regression coefficients and latent variables of TopicVec-Reg can be learned simultaneously using a variational Bayesian inference method.

We evaluated the effectiveness of TopicVec-Reg in comparison to performing linear regression as post-processing of TopicVec through experiments in the task of stock return rate prediction using news articles provided by Thomson Reuters and stock prices by the Tokyo Stock Exchange. The result showed that our model brought meaningful improvement on prediction performance.

2 Related Work

Das et al. [2015] proposed GaussianLDA, which uses pretrained word embeddings and assumes that words in a topic are drawn from a multivariate Gaussian distribution with the topic embedding as the expectation.

More recently, Dieng et al. [2020] proposed embedded topic model (ETM), which assumes that words are generated from a categorical distribution whose parameter is the inner product of the word embeddings and the embedding of the assigned topic.

As another work on topic embedding, Li et al. [2016] proposed TopicVec based on a generative word embedding model called PSDVec [Li *et al.*, 2015]. PSDVec assumes that the conditional distribution of each word given its context can be factorized approximately into independent log-bilinear terms. In TopicVec, the conditional distribution of each word is influenced by not only its context but also the topic assigned to it. Our proposed model is positioned as an extension of PSDVec and TopicVec, which will be further described in the next section.

Blei et al. [2010] proposed a supervised topic model, Supervised Latent Dirichlet Allocation (sLDA), as a way to predict the response variable associated with each document. The response variable is assumed to be generated from a generalized linear model, such as linear regression with the expectation of the latent variables of the topics assigned to the document. By modeling the document and the response variable simultaneously, it is expected to be able to estimate the latent topics that can predict the response variable for a newly given document. However, there is a similar problem with LDA in ”one-hot” representations, as mentioned in the previous section.

Thus, in this paper, by incorporating a linear regression model into the topic embedding model TopicVec, we develop a topic embedding regression model TopicVec-Reg.

3 Background

Fisrst of all, Table 1 lists the notations used in this paper.

Name	Description
$\boldsymbol{S}$	Vocabulary$\{s_1, \cdots, s_W\}$
$\boldsymbol{V}$	Embedding martix$(\boldsymbol{v}_{s_1}, \cdots, \boldsymbol{v}_{s_W})$
$\boldsymbol{D}$	Document set$\{d_1, \cdots, d_M\}$
$\boldsymbol{v}_{s_i}$	Embedding of word s_i
$a_{s_i s_j}, \boldsymbol{A}$	Bigram residuals
$\boldsymbol{t}_k, \boldsymbol{T}$	Topic embeddings
$r_k, \boldsymbol{r}$	Topic residuals
z_{ij}	Topic assignment of the j-th word in doc d_i
$\boldsymbol{\phi}_i$	Mixing proportions of topics in doc d_i
$\boldsymbol{y}$	response variables$\{y_i, \cdots, y_M\}$
$\boldsymbol{\eta}$	regression coefficients$\{\eta_1, \cdots, \eta_K\}$

Table 1: Notations.

3.1 Word Embedding Model: PSDVec

PSDVec (Positive-Semidefinite Vectors) [Li *et al.*, 2015] is a generative word embedding method and the precursor of TopicVec. In PSDVec, given its context words, the conditional distribution of the focus word is defined by the following link function:

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

Here the focus word w_c is assumed to be generated depending on the context words of size c. $\boldsymbol{v}_{w_c}^{T} \boldsymbol{v}_{w_l}$ captures linear correlations of two words and $a_{w_c w_l}$ is the bigram residual that captures the non-linear part.

Given the hyperparameter $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_W)$ and a weight function on the bigram probability $f(h_{ij})$, the generative process for the corpus is as follows:

1. For each word s_i, draw the embedding $\boldsymbol{v}_{s_i}$ from $\mathcal{N}(\boldsymbol{0}, \frac{1}{2\mu_i}\boldsymbol{I})$;
2. For each bigram (s_i, s_j), draw $a_{s_i s_j}$ from $\mathcal{N}\left(0, \frac{1}{2f(h_{ij})}\right)$;
3. For each document d_i, draw w_{ij} from vocabulary $\boldsymbol{S}$ according to the probability defined by (1).

We omit the derivation process here. The derived optimization objective is to fit PMI$(s_i, s_j) = \log \frac{P(s_i, s_j)}{P(s_i)P(s_j)}$ using $\boldsymbol{v}_{s_j}^{\top} \boldsymbol{v}_{s_i}$ and it is approached by a Block Coordinate Descent algorithm.

3.2 Topic Embedding Model: TopicVec

The conditional distribution of the focus word in TopicVec [Li *et al.*, 2016] is defined by the following function:

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

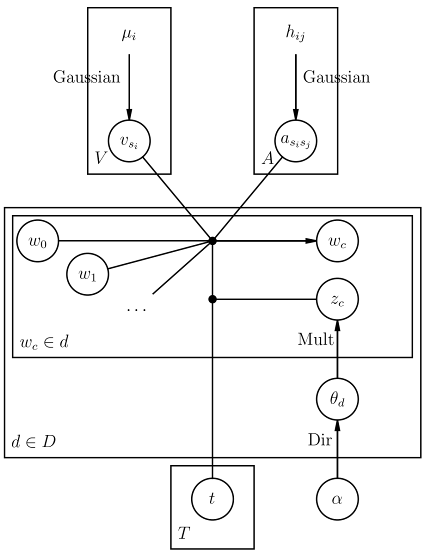

Figure 1: A graphical model of TopicVec.

Here $\boldsymbol{t}_{z_c}$ is the embedding of the topic assigned to the focus word and it can be treated as one of the context words. r_{z_c} is the residual about the topic z_c. With the link function, the relevance between the word and the topic is encoded by the cosine distance in the embedding space. The generative process of TopicVec is as follows:

1. For each topic k, draw a topic embedding uniformly from a hyper ball of radius γ, i.e. $\boldsymbol{t}_k \sim \text{Unif}(B_\gamma)$;
2. For each document d_i:
   (a) Draw the mixing proportions $\boldsymbol{\phi}_i$ from the Dirichlet prior $\text{Dir}(\boldsymbol{\alpha})$;
   (b) For the j-th word:
      i. Draw topic assignment z_{ij} from the categorical distribution $\text{Cat}(\boldsymbol{\phi}_i)$;
      ii. Draw word w_{ij} from vocabulary S according to $P\left(w_{ij} | w_{i,j-c} : w_{i,j-1}, z_{ij}, d_i\right)$.

The graphical model in Figure 1 presents the generative process above.

The complete data loglikelihood of the whole corpus: the full joint log-probability of the corpus $\boldsymbol{D}$, word embeddings $\boldsymbol{V}$, bigram residuals $\boldsymbol{A}$, topic embeddings $\boldsymbol{T}$, topic assignments $\boldsymbol{Z}$, and topic distributions $\boldsymbol{\phi}$ can be written as:

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

where $m_{ik} = \sum_{j=1}^{L_i} \delta(z_{ij} = k)$ indicates that the number of words assigned to topic k. C_0 is constant given the hyperparameters.

Given the hyperparameters $\boldsymbol{\alpha}$, γ, and $\boldsymbol{\mu}$, the optimal $\boldsymbol{V}$, $\boldsymbol{T}$, and $p(\boldsymbol{Z}, \boldsymbol{\phi} | \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{V}, \boldsymbol{T})$ are estimated to maximize the loglikelihood as follows:

Step1 $\boldsymbol{V}$ and $\boldsymbol{A}$ are obtained according to the original PSDVec;

Step2 Given $\boldsymbol{V}$ and $\boldsymbol{A}$, the loglikelihood function is used to find the optimal $\boldsymbol{T}$ and $p(\boldsymbol{Z}, \boldsymbol{\phi} | \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{V}, \boldsymbol{T})$.

Since the posterior $p(\boldsymbol{Z}, \boldsymbol{\phi} | \boldsymbol{D}, \boldsymbol{T})$ is analytically intractable, the posterior is approximated by the variational distribution $q(\boldsymbol{Z}, \boldsymbol{\phi}; \boldsymbol{\pi}, \boldsymbol{\theta}) = q(\boldsymbol{\phi}; \boldsymbol{\theta}) q(\boldsymbol{Z}; \boldsymbol{\pi})$. Here, the KL divergence is introduced and the estimation task is replaced with the problem of maximizing the variational lower bound $\mathcal{L}(q, \boldsymbol{T})$:

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

Here, $\mathcal{H}(q)$ is the entropy of q. The variational lower bound $\mathcal{L}(q, \boldsymbol{T})$ is as follows:

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

Here the generalized EM algorithm is used to find the optimal q^* and $\boldsymbol{T}^*$ that maximize $\mathcal{L}(q, \boldsymbol{T})$:

E-Step

$$\pi_{ij}^k \propto \exp\left\{ \quad (\theta_{ik}) + \boldsymbol{v}_{w_{ij}}^\top \boldsymbol{t}_k + r_k \right\}, \qquad (6)$$

$$\theta_{ik} = \sum_{j=1}^{L_i} \pi_{ij}^k + \alpha_k; \qquad (7)$$

M-Step

$$\boldsymbol{T}_{\text{new}} = \boldsymbol{T} + \lambda \left(l, \sum_{i=1}^{M} L_i \right) \frac{\partial \mathcal{L}\left(q, \boldsymbol{T}\right)}{\partial \boldsymbol{T}}, \qquad (8)$$

$$\boldsymbol{r} = -\log(\boldsymbol{u} \exp\{\boldsymbol{V}^\top \boldsymbol{T}\}). \qquad (9)$$

Here, $\lambda(l, \sum_{i=1}^{M} L_i) = \frac{L_0 \lambda_0}{l \cdot \max\{\sum_{i=1}^{M} L_i, L_0\}}$ is the learning rate, l is the number of iterations in the learning process, L_0 is a predetermined threshold of the number of words, λ_0 is the initial value of λ, and $\boldsymbol{u}$ is the unigram probability of the words occurring in the corpus.

4 Topic Embedding Regression Model

In this section, we propose the topic embedding regression model, TopicVec-Reg, by incorporating regression function into TopicVec mentioned in the previous section.

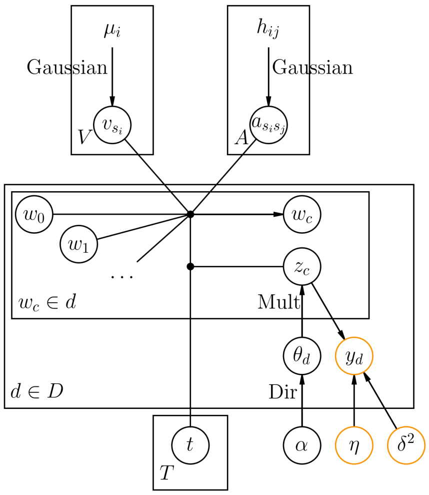

Figure 2: A graphical model of TopicVec-Reg.

4.1 Generative Process

TopicVec-Reg assumes that the document d_i and the response variable y_i associated with the document are generated following the generative process, as follows:

1. Generate each document d_i according to the generative process of TopicVec;
2. Draw response variable $y_i \sim \mathcal{N}(\boldsymbol{\eta}^\top \bar{\boldsymbol{Z}}_i, \delta^2)$.

Here, the mean of the Gaussian distribution is the inner product of the regression coefficients $\boldsymbol{\eta}$ and the expectation of latent topic assignments $\bar{\boldsymbol{Z}}_i$. Figure 2 presents a graphical model of TopicVec-Reg.

4.2 Estimation of Parameters

Similar to TopicVec, we estimate the parameters using the generalized EM algorithm after obtaining the loglikelihood function.

First, we rewrite the complete data loglikelihood in (3) to include response variables $\boldsymbol{y} = \{y_i\}$ as:

$$\begin{aligned}
&\log p(\boldsymbol{D}, \boldsymbol{A}, \boldsymbol{V}, \boldsymbol{Z}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{y} | \boldsymbol{\alpha}, \gamma, \boldsymbol{\mu}, \boldsymbol{\eta}, \delta^2) \\
=&C_0 - \log \mathcal{Z}(\boldsymbol{H}, \boldsymbol{\mu}) - \|\boldsymbol{A}\|^2_{f(\boldsymbol{H})} - \sum_{i=1}^{W} \mu_i \left\|\boldsymbol{v}_{s_i}\right\|^2 \\
&+ \sum_{i=1}^{M} \left\{ \sum_{k=1}^{K} \log \phi_{ik} \left(m_{ik} + \alpha_k - 1\right) \right. \\
&- \frac{1}{2} \log(2\pi\delta^2) - \frac{1}{2\delta^2} (y_i^2 - 2 y_i \boldsymbol{\eta}^\top \bar{\boldsymbol{Z}}_i + \boldsymbol{\eta}^\top \bar{\boldsymbol{Z}}_i \bar{\boldsymbol{Z}}_i^\top \boldsymbol{\eta}) \\
&+ \sum_{j=1}^{L_i} \left(\boldsymbol{v}_{w_{ij}}^\top \left(\sum_{l=j-c}^{j-1} \boldsymbol{v}_{w_{il}} + \boldsymbol{t}_{z_{ij}} \right) \right. \\
&\left.\left. + \sum_{l=j-c}^{j-1} a_{w_{il} w_{ij}} + r_{z_{ij}} \right) \right\}. \qquad (10)
\end{aligned}$$

Then by introducing a variational distribution $q(\boldsymbol{Z}, \boldsymbol{\phi}; \boldsymbol{\pi}, \boldsymbol{\theta}) = q(\boldsymbol{\phi}; \boldsymbol{\theta}) q(\boldsymbol{Z}; \boldsymbol{\pi})$ as in TopicVec, the expectation of the variational distribution of the loglikelihood of the response variable y_i is obtained by:

$$\begin{aligned}
&\mathbb{E}_q \left[\log p \left(y_i | \boldsymbol{Z}_i, \boldsymbol{\eta}, \delta^2 \right) \right] \\
=& -\frac{1}{2} \log \left(2\pi\delta^2 \right) \\
&- \frac{1}{2\delta^2} \left(y_i^2 - 2 y_i \boldsymbol{\eta}^\top \mathbb{E}_q[\bar{\boldsymbol{Z}}_i] + \boldsymbol{\eta}^\top \mathbb{E}_q \left[\bar{\boldsymbol{Z}}_i \bar{\boldsymbol{Z}}_i^\top \right] \boldsymbol{\eta} \right), \qquad (11)
\end{aligned}$$

where

$$\mathbb{E}_q[\bar{\boldsymbol{Z}}_i] = \bar{\boldsymbol{\pi}}_i = \frac{1}{L_i} \sum_{j=1}^{L_i} \boldsymbol{\pi}_{ij},$$

$$\mathbb{E}_q \left[\bar{\boldsymbol{Z}}_i \bar{\boldsymbol{Z}}_i^\top \right] = \frac{1}{{L_i}^2} \left(\sum_{j=1}^{L_i} \sum_{m \neq j} \boldsymbol{\pi}_{ij} \boldsymbol{\pi}_{im}^\top + \sum_{j=1}^{L_i} \operatorname{diag} \left\{ \boldsymbol{\pi}_{ij} \right\} \right).$$

Thus, the objective $\mathcal{L}_{\text{reg}}(q, \boldsymbol{T})$ is obtained by adding (11) to (5):

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

We update θ_{ik} using (7). We can obtain the solution by setting the partial derivative w.r.t. π_{ij}^k to 0 after isolating the terms containing π_{ij}^k:

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

where

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

is the partial derivative of $\sum_{j=1}^{L_i} \sum_{m \neq j}^{L_i} \boldsymbol{\pi}_{ij} \boldsymbol{\pi}_{im}^{\top}$ w.r.t. π_{ij}^{k}. $\boldsymbol{\Pi}_{im}$ is a $K \times K$ matrix whose row is $(\pi_{im}^{1}, \pi_{im}^{2}, \cdots, \pi_{im}^{K})$.

The terms containing $\boldsymbol{\eta}$ and δ^2 in the learning objective can be found in (11). So we define a $M \times (K+1)$ matrix A whose i-th row is $(\bar{\boldsymbol{Z}}_i, 1)$ with the $(K+1)$-th element corresponding to the bias and rewrite (11) of the whole corpus as below:

$$\begin{aligned} \boldsymbol{\eta}^{'} &= \mathrm{Concat}(\boldsymbol{\eta}, \eta_{\text{bias}}), \\ & \mathbb{E}_q[\log p(\boldsymbol{y}|A, \boldsymbol{\eta}^{'}, \delta^2)] \\ = &- \frac{M}{2} \log(2\pi\delta^2) \\ &- \frac{1}{2\delta^2} \mathbb{E}_q \left[(\boldsymbol{y} - A\boldsymbol{\eta}^{'})^{\top} (\boldsymbol{y} - A\boldsymbol{\eta}^{'}) \right], \end{aligned} \tag{14}$$

where $\boldsymbol{\eta}^{'}$ is obtained by the function Concat$(\cdot)$ that concatenates the bias term to the end of $\boldsymbol{\eta}$. Taking the derivative w.r.t. $\boldsymbol{\eta}^{'}$ and δ^2 and setting them to 0, respectively, we obtain the following:

$$\hat{\boldsymbol{\eta}}^{'}_{\text{new}} = \left(\mathbb{E}_q \left[A^{\top} A \right] \right)^{-1} \mathbb{E}_q \left[A \right]^{\top} \boldsymbol{y}, \tag{15}$$

$$\begin{aligned} & \hat{\delta}^2_{\text{new}} = \\ & \frac{1}{M} \left\{ \boldsymbol{y}^{\top} \boldsymbol{y} - \boldsymbol{y}^{\top} \mathbb{E}_q[A] \left(\mathbb{E}_q \left[A^{\top} A \right] \right)^{-1} \mathbb{E}_q[A]^{\top} \boldsymbol{y} \right\} \end{aligned} \tag{16}$$

where we define $\boldsymbol{\pi}^{'}_{ij} := \mathrm{Concat}(\boldsymbol{\pi}_{ij}, 1)$, and

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

$$\begin{aligned} & \mathbb{E}\left[A^{\top} A\right] = \\ & \sum_{i=1}^{M} \left(\frac{1}{{L_i}^2} \left(\sum_{j=1}^{L_i} \sum_{m \neq j}^{L_i} \boldsymbol{\pi}^{'}_{ij} \boldsymbol{\pi}^{'\top}_{im} + \sum_{j=1}^{L_i} \mathrm{diag}\{\boldsymbol{\pi}^{'}_{ij}\} \right) \right). \end{aligned}$$

Column $(K+1)$ corresponds to the bias term of the regression coefficient. The topic embedding $\boldsymbol{T}$ is updated by the gradient descent method as shown in (8).

5 Experimental Results

To evaluate the prediction performance of TopicVec-Reg, we conducted an evaluation experiment in comparison with TopicVec.

5.1 Dataset

As text data, we used Japanese financial articles distributed by Thomson Reuters in 2017. The stock return rate of the company mentioned in the article is tied to the article as the response variable, which is defined as

$$R = \frac{V_f - V_f^{'}}{V_f^{'}}$$

where :

V_f = final value on the day **after** the article was published

$V_f^{'}$ = final value on the day **before** the article is published

	Training Sets (Closed Test)				Test Set (Open Test)
Term	1	2	3	4	5
Dataset	Jan. ∼ Feb.	Feb. ∼ Mar.	Mar. ∼ Apr.	Apr. ∼ May	May ∼ Jun.
# of documents	720	728	726	652	673

Table 2: Overview of the datasets.

We used the Tokyo Stock Exchange's historical data for stock prices. When more than one company were mentioned in an article, we sorted the stock return rates of those companies in descending order, then removed those whose absolute value was less than the mean plus one standard deviation, because the stock return rates of such companies may have not been affected by the article. We excluded the articles each of which mentions more than 5 companies, considering that such an article is likely to focus on an industry trend rather than several specific companies.

We divided the financial articles distributed from January to June 2017 into two-monthly segments and prepared five preprocessed datasets in time order as shown in Table 2. The datasets Term 1∼4 were used as training sets for closed test, and the dataset Term 5 was used for open test.

We performed preprocessing on the text data. After removing intractable tables and unneeded expressions in the articles, we performed morphological analysis using MeCab with mecab-ipadic-NEologd [Sato, 2015] [Sato *et al.*, 2016] [Sato *et al.*, 2017], a dictionary of neologisms and unique expressions, to segment words and exclude stop words such as particles and conjunctions. Finally, low-frequency words occurring in less than 5 articles and articles of less than 50 words were removed.

5.2 Experimental Setup

In our experiments, we compared two models: our proposed model (TopicVec-Reg) and a baseline model (TopicVec+LR), as follows:

TopicVec-Reg: Estimate the regression parameters and topics simultaneously.

TopicVec+LR: Perform linear regression after topic estimation with TopicVec, as a baseline.

TopicVec-Reg was compared to TopicVec+LR in the closed test with 10 different number of topics $K \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$, and then an open test was performed with the number of topics that brought the smallest average MSE in the closed test.

First, we conducted the closed test on the training set from Term 1 to Term 4 to obtain the topic embeddings $\boldsymbol{T}$ and the regression coefficient $\boldsymbol{\eta}$. Specifically, for the test on Term 1, $\boldsymbol{T}$ was randomly initialized following a Gaussian distribution, and the variational parameter $\boldsymbol{\pi}$ was randomly initialized following a uniform distribution at the beginning of learning. Then the topic embeddings $\boldsymbol{T}$ obtained on Term 1 was used as the initial $\boldsymbol{T}$ for Term 2. During training, $\boldsymbol{\eta}$ was updated

Term	Model	K=5	K=10	K=15	K=20	K=25	K=30	K=35	K=40	K=45	K=50
1	TV-Reg	0.005645	0.005229	0.004957	0.004138	0.004515	0.003904	0.004264	0.003621	0.004113	0.00355
	TV+LR	0.005775	0.005722	0.005605	0.00517	0.005179	0.005054	0.005013	0.004893	0.004837	0.004598
2	TV-Reg	0.004411	0.004315	0.004115	0.003965	0.003846	0.003764	0.003708	0.00363	0.003413	0.003524
	TV+LR	0.004519	0.004416	0.004277	0.004071	0.004015	0.004082	0.004267	0.004039	0.003824	0.003691
3	TV-Reg	0.004071	0.003916	0.003719	0.003643	0.003578	0.003496	0.003569	0.003159	0.003031	0.003278
	TV+LR	0.004104	0.003869	0.003754	0.003771	0.003591	0.003557	0.003721	0.003732	0.003492	0.00311
4	TV-Reg	0.005671	0.005388	0.004864	0.005114	0.004863	0.004551	0.004869	0.004551	0.004762	0.004436
	TV+LR	0.005691	0.005329	0.00531	0.005241	0.00514	0.004958	0.005024	0.005139	0.004999	0.004711
average	TV-Reg	0.00495	0.004712	0.004414	0.004215	0.004201	0.003929	0.004103	0.00374	0.00383	0.003697
	TV+LR	0.005022	0.004834	0.004737	0.004563	0.004481	0.004413	0.004506	0.004451	0.004288	0.004028

Table 3: MSE in the closed test.

every 5 iterations and was used to predict the response variable when the learning process converged.

We used the same experimental setup for TopicVec+LR, as for TopicVec-Reg. TopicVec is used to estimate the topics, and then the linear regression was used to estimate the regression coefficients.

The word embeddings $\boldsymbol{V}$ and the residuals $\boldsymbol{A}$ were trained by PSDVec using the Japanese Wikipedia. The hyperparameter was fixed as $\boldsymbol{\alpha} = (0.1, \cdots, 0.1)$.

In all experiments, the convergence condition was that the rate of change of $\boldsymbol{\pi}$ must be less than 0.1% for three consecutive times during learning.

We predict the stock return rate as below:

$$\hat{y}_i = \boldsymbol{\eta}^\top \mathbb{E}_q[\bar{\boldsymbol{Z}}_i] + \eta_{\text{bias}} = \boldsymbol{\eta}'^\top \frac{1}{L_i} \sum_{j=1}^{L_i} \boldsymbol{\pi}'_{ij} \tag{17}$$

We used the Mean Squared Error (MSE) between the true and predicted values of the stock return rate as the measure of prediction performance:

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

Finally, for the number of topics K with the smallest mean value of MSE in the closed test, the topic embeddings $\boldsymbol{T}$ obtained on Term 4 was used as the initial $\boldsymbol{T}$ to estimate the topics on Term 5 with TopicVec. Since the data for the first month of Term 5 was already used to train the model on Term 4, we predicted the response variables of the data of the second month of Term 5 using $\boldsymbol{\eta}$ obtained with Term 4 and then calculated the MSE between the true values and the predicted ones as the evaluation metric of the open tests.

5.3 Results

Table 3 shows the MSEs obtained as the result of the closed test on the training sets and their average values. The numbers in blue show that the proposed model has better prediction accuracy than the baseline model.

TV-Reg	**0.01851** ± 0.09196
TV+LR	0.01917 ± 0.08924

Table 4: MSE and standard deviation in the open test when K=50.

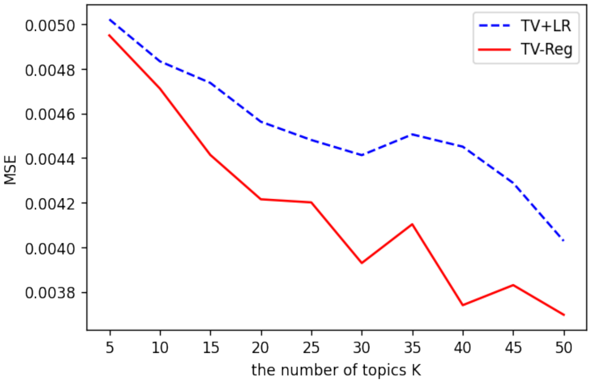

Figure 3: Average MSE for the varying number of topics K.

As noted in Table 3, TopicVec-Reg has better prediction accuracy than TopicVec+LR when $K \in \{5, 15, 20, 25, 30, 35, 40, 45\}$. Moreover, Figure 3 shows that TopicVec-Reg has better prediction accuracy than TopicVec+LR on all of K on average.

Since the average of MSE was the smallest when $K = 50$, we performed an open test on Term 5 when $K = 50$. Table 4 shows the result of MSE and standard deviation in the open test. We further carried out the Wilcoxon signed-rank test for the open-test result and then observed that the p-value was less then 1%, indicating that TopicVec-Reg has better prediction accuracy than TopicVec+LR with a statistically significant difference.

6 Conclusions

In this paper, we proposed TopicVec-Reg, a topic embedding regression model combining TopicVec and linear regression, with the aim of predicting the stock return rate of the company mentioned in each financial article. The result of the closed test on the training datasets showed that our proposed model has better prediction accuracy on average than TopicVec+LR, and statistically significant improvement in prediction accuracy was also observed in the open test under the best number of topics. More detailed evaluation, such as compared with some other models, is left for the future work.

Acknowledgments: This research was supported in part by ROIS NII Open Collaborative Research 2021-21FS01.

References

[Blei and McAuliffe, 2010] David M Blei and Jon D McAuliffe. Supervised topic models. *arXiv preprint arXiv:1003.0783*, 2010.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[Das *et al.*, 2015] Rajarshi Das, Manzil Zaheer, and Chris Dyer. Gaussian lda for topic models with word embeddings. In *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)*, pages 795–804, 2015.

[Dieng *et al.*, 2020] Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

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

[Li *et al.*, 2016] Shaohua Li, Tat-Seng Chua, Jun Zhu, and Chunyan Miao. Generative topic embedding: a continuous representation of documents (extended version with proofs). *arXiv preprint arXiv:1606.02979*, 2016.

[Sato *et al.*, 2016] Toshinori Sato, Taiichi Hashimoto, and Manabu Okumura. Operation of a word segmentation dictionary generation system called neologd (in japanese). In *Information Processing Society of Japan, Special Interest Group on Natural Language Processing (IPSJ-SIGNL)*, pages NL–229–15. Information Processing Society of Japan, 2016.

[Sato *et al.*, 2017] Toshinori Sato, Taiichi Hashimoto, and Manabu Okumura. Implementation of a word segmentation dictionary called mecab-ipadic-neologd and study on how to use it effectively for information retrieval (in japanese). In *Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing*, pages NLP2017–B6–1. The Association for Natural Language Processing, 2017.

[Sato, 2015] Toshinori Sato. Neologism dictionary based on the language resources on the web for mecab, 2015.