A supervised topic embedding model and its application

Weiran Xu*, Koji Eguchi
Graduate School of Advanced Science and Engineering, Hiroshima University, Higashihiroshima, Hiroshima, Japan

* d210104@hiroshima-u.ac.jp

Abstract

We propose rTopicVec, a supervised topic embedding model that predicts response variables associated with documents by analyzing the text data. Topic modeling leverages document-level word co-occurrence patterns to learn latent topics of each document. While word embedding is a promising text analysis technique in which words are mapped into a low-dimensional continuous semantic space by exploiting the local word co-occurrence patterns within a small context window. Recently developed topic embedding benefits from combining those two approaches by modeling latent topics in a word embedding space. Our proposed rTopicVec and its regularized variant incorporate regression into the topic embedding model to model each document and a numerical label paired with the document jointly. In addition, our models yield topics predictive of the response variables as well as predict response variables for unlabeled documents. We evaluated the effectiveness of our models through experiments on two regression tasks: predicting stock return rates using news articles provided by Thomson Reuters and predicting movie ratings using movie reviews. Results showed that the prediction performance of our models was more accurate in comparison to three baselines with a statistically significant difference.

Introduction

Topic models are statistical machine learning models that find latent semantic structure in a corpus. They have been commonly applied as a tool for analyzing large amounts of text data in a variety of fields. Most topic models focus on the words that appear in documents. In some fields, however, a document is generally accompanied by a response, such as a movie review with a rating of the movie, or a financial news article accompanied by a financial indicator. Thus, to tackle such regression problems, we developed a supervised topic embedding model to infer latent topics predictive of the response.

Latent Dirichlet Allocation (LDA) [1] is a representative topic model which assumes a latent topic hidden behind each word in a document and infers the topics which compose the document by employing a hierarchical Bayesian structure. In LDA, the co-occurrence patterns of words can express semantic relevance when the corpus is large enough. However, the input documents in LDA are represented as bag-of-words (BoW), which have issues with high
dimensionality and sparsity since the words are represented as one-hot representations that lack the notion of similarity between similar words. For instance, according to Dieng et al. [2], the quality of LDA topics decreases as the vocabulary size grows larger. Therefore, conventional topic models like LDA find latent topics from a corpus-level perspective with a drawback of ignoring word-level features because of the one-hot representations.

A solution to the drawback is word embedding, which introduces a semantic space where words are represented as \( n \)-dimensional vectors and the distance of the words can be used to measure the similarity of them, where \( n \) is much smaller than the vocabulary size. That is, word embedding maps words with similar meanings closely in a low-dimensional space where a vocabulary with tens of thousands is embedded. The semantic space is constructed by analyzing semantic similarity denoted by word co-occurrences in a sufficiently large corpus since words with similar contexts usually have similar meanings according to the distributional hypothesis [3].

Hence topic models and word embedding methods are both constructed by analyzing word co-occurrences but in different ways. A variety of approaches aiming to capture the global semantic structure and local word-level features by combining topic models and word embedding have been proposed in the literature [2, 4–8]. Such models have been shown to outperform LDA in terms of topic quality, predictive performance, and document classification tasks. In this paper, we extend TopicVec [5], a topic embedding model that seamlessly integrates topic models and word embedding, which is based on the idea that latent topics are also included in the word embedding space by combining the notion of topics with the generative word embedding model PSDVec [9]. The topic distributions for each document in TopicVec, like LDA, are assumed to be drawn from Dirichlet priors. In addition, each word in a document is assumed to be extracted from a link function that considers both surrounding context and global topics represented by word embeddings and topic embeddings respectively.

On the other hand, the topic distributions for each document suitably discovered by topic models can be leveraged for downstream tasks like regression problems. However, the topics learned by unsupervised topic models may be inappropriate to describe the coefficients of the post-processed regression since the supervisory signals such as numerical labels associated with documents are not involved in the topics learning procedure. Therefore, there exist approaches that share the goal of learning topics and regression coefficients jointly when each document is paired with a label or a response [10–13]. However, such models using sparse representations suffer from the same problem with LDA. By replacing the one-hot representations with dense vectors, the topics and the prediction accuracy could be improved intuitively.

This paper is extended from our preliminary work [14]. To the best of our knowledge, this study is the first to extend topic embedding models like TopicVec for regression tasks. Inspired by sLDA [10], in this paper, assuming a response variable is drawn from a Gaussian whose expectation is the inner product of the expectation of the topics for the corresponding document and the regression coefficients, we propose rTopicVec as a supervised topic embedding model for modeling the link between each document and a numerical label paired with it. Moreover, by assuming Gaussian priors to the regression coefficients, we propose a regularized version rTopicVec-Ridge. A variational Bayesian inference approach is used to simultaneously learn the parameters of both models, including the regression coefficients and latent variables. We conducted two experiments to verify the effectiveness of our proposed models. The objective of the first experiment was to predict stock return rates using news articles provided by Thomson Reuters and stock prices from the Tokyo Stock Exchange. The second experiment was to predict movie rating scores using movie reviews. In comparison to baseline models, our proposed models improved prediction performance significantly and also have

Funding: K.E. received ROIS NII Open Collaborative Research (Grant Number: 22FS01), provided by the National Institute of Informatics, Research Organization of Information and Systems: <https://www.nii.ac.jp/en>. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.
the advantage of providing interpretability with latent topics for advanced regression analysis. The contributions of this work are summarized as follows:

1. We developed a supervised topic embedding model where words and topics are represented by embeddings and a regularized variant of the model. To the best of our knowledge, this work is the first to extend an unsupervised topic embedding model to a supervised one for regression tasks.

2. Our models learn topic parameters and regression parameters simultaneously and introduce word embeddings to improve the topic quality and prediction accuracy.

3. Our models outperformed three baseline models in prediction accuracy on two tasks that predict numerical labels associated with documents in two languages, respectively. The topic coherence was also improved by taking numerical labels into account.

Related work

The model TopicVec [5] we extend in this paper is a topic embedding model incorporating a generative word embedding PSDVec [9] with latent topics. There exist other models that share the idea of combining topic models with word embedding methods. All of those are unsupervised models, while our model is a supervised extension of topic embedding.

**GaussianLDA**

GaussianLDA proposed by Das et al. [4], uses pre-fitted word embeddings to benefit topic models. It replaces the categorical distributions representing topic-word distributions in LDA with multivariate Gaussian distributions so that topics and words share the same embedding space where the word embeddings are assumed to be drawn from a multivariate Gaussian centered at a topic embedding that is drawn from a multivariate Gaussian with zero mean and an inverse Wishart distribution as covariance.

**STE**

Assuming that a word may have different representations under different topics, a unified framework STE (Skip-gram Topical word Embedding) proposed by Shi et al. [7] learns latent topics and topic-specific word embeddings jointly rather than learns them separately in a two-step way. The learned word embeddings are useful to address the issue of polysemy. They proposed two variants of STE by modeling each skip-gram in two different ways depending on whether the topics behind the two words in each skip-gram are the same or not.

**GPU-DMM**

Focusing on analyzing short texts, Li et al. [6] proposed a topic model for short texts that is based on Dirichlet Mixture Model with pre-fitted word embeddings incorporated by a generalized Pólya urn (GPU) model. The word embedding trained on a large corpus can supplement short text analysis where the context and word co-occurrence are limited. The GPU model promotes the semantically related words under the same topic, which is efficient for short text analysis.

**WEI-FTM**

WEI-FTM is another topic model that uses pre-fitted word embeddings as prior knowledge to boost the topic quality when analyzing short texts proposed by Zhao et al. [8]. They assumed
that the topic distributions over words are affected by the inner product of the word embeddings and topic embeddings. They also applied sparsity-enforcing prior on topics to make each of them focus on a subset of words rather than the whole vocabulary, leading to better topic quality.

**ETM**

More recently, ETM (embedded topic model) developed by Dieng et al. [2] learns topics and the embeddings of them either fitting word embeddings jointly or using pre-fitted word embeddings by an amortized variational inference algorithm for which they replaced the Dirichlet with the logistic-normal distribution to model the topic distributions. They assumed that each word is generated according to the agreement between the word embeddings and the embedding of its assigned topic.

Regarding the supervised models that use topic representations of documents for regression problems, most of them mainly focus on extending LDA. Such models can be used to predict the label given an unlabeled document by inferring its latent topics. Mcauliffe et al. [10] proposed Supervised LDA (sLDA), in which the response paired with each document is presumed to be drawn from a Gaussian whose expectation is the product of the topic distribution of each document and the regression coefficients. MedLDA proposed by Zhu et al. [12] has a similar goal with sLDA and trains LDA with SVM by integrating the max-margin principle with the topic models. More recently, Wang et al. [13] proposed TAM, in which attention RNN is exploited to extend neural topic models for regression and classification tasks. While in our work, following sLDA, we propose rTopicVec that integrates the topic embedding model TopicVec with linear regression for regression problems and its regularized version rTopicVec-Ridge. We believe that our models should yield higher prediction accuracy than learning a regression model using the already estimated topic distributions as the explanatory variables, and also than sLDA, by involving word embeddings.

**Background**

We briefly review the generative word embedding PSDVec [9] and topic embedding TopicVec [5] as the basic background of our work. The notations used in this paper are listed in Table 1.

| Name | Description |
|------|-------------|
| $S$  | Vocabulary $\{s_1, \ldots, s_W\}$ |
| $V$  | Embedding matrix $\{v_{s_1}, \ldots, v_{s_W}\}$ |
| $D$  | Document set $\{d_1, \ldots, d_M\}$ |
| $v_{s_m}$ | Embedding of word type $s_m$ |
| $\mu$ | Tiknov regularization coefficients $\{\mu_1, \ldots, \mu_W\}$ |
| $h_{mn}$ | Bigram empirical probability of bigram $(s_m, s_n)$ |
| $a_{s_m, s_n}$ | Bigram residuals |
| $A$  | Topic embeddings |
| $r_k$ | Topic residuals |
| $w_{ij}$ | $j$-th word in document $d_i$ |
| $L_i$ | Length of document $d_i$ |
| $z_{ij}$ | Topic assignment of the $j$-th word in doc $d_i$ |
| $\phi_i$ | Mixing proportions of topics in doc $d_i$ |
| $y$  | Response variables $\{y_1, \ldots, y_M\}$ |
| $\eta$ | Regression coefficients $\{\eta_1, \ldots, \eta_K\}$ |

https://doi.org/10.1371/journal.pone.0277104.t001
PSDVec

Positive-Semidefinite Vectors (PSDVec) [9] is a generative word embedding method based on which TopicVec was developed. In PSDVec, the conditional distribution of a focus word given its context words is assumed to be factorized approximately into independent log-bilinear terms and it is defined by the following link function:

\[
P(w_{ij}|w_{i:j-1}: w_{j+1}) \approx P(w_{ij}) \exp \left\{ \sum_{j'=j}^{j-1} v_{w_{ij}}^T v_{w_{ij}} + \sum_{j'=j}^{j-1} a_{w_{ij}w_{ij}} \right\}. \tag{1}
\]

The link function connects the word embeddings with the corpus statistics. Here, the focus word \( w_{ij} \) is assumed to be generated depending on context of size \( c \). \( v_{w_{ij}}^T v_{w_{ij}} \) captures the linear correlations of two words and bigram residual \( a_{w_{ij}w_{ij}} \) captures the non-linear part.

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

1. For each word type \( s_m \), draw the embedding \( v_{sn} \) from \( \mathcal{N}(0, \frac{1}{2m}I) \);
2. For each bigram \( (s_m, s_n) \), draw \( a_{sn} \) from \( \mathcal{N}(0, \frac{1}{2n(m+n)}) \);
3. For each document \( d_i \), draw the \( j \)-th word \( w_{ij} \) from vocabulary \( S \) according to the probability defined by (1).

We omit the derivation process here. The derived optimization objective is to fit pointwise mutual information \( \text{PMI}(s_m; s_n) = \log \frac{P(s_m, s_n)}{P(s_m)P(s_n)} \) using \( v_{sn}^T v_{sn} \) and it is optimized by a block coordinate descent algorithm.

TopicVec

TopicVec [5] was developed by taking topics into account in PSDVec described in the previous section. The conditional distribution of the focus word in TopicVec is therefore affected by its context as well as the topic assigned to the word and it is defined by the following function:

\[
P(w_{ij}|w_{i:j-1}, z_{ij}, d_i) \approx P(w_{ij}) \exp \left\{ v_{w_{ij}}^T \left( \sum_{j'=j}^{j-1} v_{w_{ij}} + t_{z_{ij}} \right) + \sum_{j'=j}^{j-1} a_{w_{ij}w_{ij}} + r_{z_{ij}} \right\}. \tag{2}
\]

Here, \( t_{z_{ij}} \) is the embedding of the topic assigned to the focus word and can be treated as one of the context words. \( r_{z_{ij}} \) is the residual of the topic \( z_{ij} \). With the link function, the distance between each word and each topic encodes the relevance of them in the embedding space. The generative process of TopicVec is as follows:

1. For each topic \( k \), randomly draw a topic embedding \( t_k \), each element of which is sampled from the standard Gaussian \( \mathcal{N}(0, 1) \).
2. For each document \( d_i \):
   a. Draw the mixing proportions \( \phi_i \) from the Dirichlet prior \( \text{Dir}(\alpha) \);
   b. For the \( j \)-th word:
      i. Draw topic assignment \( z_{ij} \) from the categorical distribution \( \text{Cat}(\phi_i) \);
ii. Draw word $w_{ij}$ from vocabulary $S$ according to $P(w_{ij}|w_{i,j-1}, z_{ij}, d_i)$.

Fig 1 presents a graphical model for the generative process above. The complete data loglikelihood of the whole corpus (the full joint log-probability of the corpus $D$, word embeddings $V$, bigram residuals $A$, topic embeddings $T$, topic assignments $Z$, and topic distributions $\phi$) can be written as

$$
\log p(D, A, V, Z, T, \phi|\alpha, \gamma, \mu) = C_0 - \sum_m \|v_m\|^2 + \sum_k \left\{ \sum_i \log \phi_k(b_{ik} + \alpha_i - 1) + \sum_j \left( r_{z_i} + v_{w_j}^{(m)} \sum_k a_{w_j z_k} \right) \right\},
$$

where $b_{ik} = \sum_j \delta(z_{ij} = k)$ indicates the number of words assigned to topic $k$. $C_0$ is constant given the hyperparameters.
Given the hyperparameters $\alpha$, $\gamma$, and $\mu$, the optimal $V$, $T$, and $p(Z, \phi|D, A, V, T)$ are estimated to maximize the loglikelihood as follows:

**Step 1** $V$ and $A$ are optimized using the original PSDVec;

**Step 2** Given optimal $V$ and $A$, the optimal $T$ and $p(Z, \phi|D, A, V, T)$ are optimized using the loglikelihood function.

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

$$\text{KL}(q||p) = \log p(D|T) - (\mathbb{E}_q[\log p(D, Z, \phi|T)] + \mathcal{H}(q))$$

(4)

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

$$L(q, T) = \sum_{i=1}^{M} \left\{ \sum_{k=1}^{K} (\sum_{j=1}^{K} \pi_{ij}^{k} + a_{k} - 1)(\psi(\theta_{ik}) - \psi(\theta_{ik}')) \right\}$$

$$+ \text{Tr}(T^T \sum_{j=1}^{I} \mathbf{v}_{ij} \mathbf{v}_{ij}^T) + r^T \sum_{j=1}^{I} \pi_{ij} + \mathcal{H}(q) + C_1.$$ 

(5)

Here, $C_1 = C_0 - \log \mathcal{Z}(H, \mu) - ||A||_2^2 - \sum_{m=1}^{W} \mu_{\alpha_m} \mathbf{w}_{\alpha_m}^2 + \sum_{i=1}^{M} \left\{ \mathbf{v}_{ij}^T \sum_{j=1}^{I} \mathbf{v}_{ij} \right\}$ is constant. Then the generalized EM algorithm is used to find the optimal $q'$ and $T'$ that maximize $L(q, T)$ as shown in Algorithm 1.

**Algorithm 1** The generalized EM algorithm

**Initialize** $T$, $r$, $\theta$

repeat

**E-Step:**

$$\pi_{ij} \propto \exp\{\psi(\theta_{ik}) + \mathbf{v}_{ij}^T \mathbf{t}_{ik} + r_k\}$$

$$\theta_{ik} = \sum_{j=1}^{I} \pi_{ij} + a_{k}$$

**M-Step:**

$$T_{new} = T + \lambda \left( \ell, \sum_{i=1}^{M} L_i \right) \frac{\partial L(q, T)}{\partial T}$$

$$r = -\log (u \exp \{V^T T\})$$

until converged

**Supervised topic embedding model**

In this section, we introduce our supervised topic embedding model for regression, rTopicVec, which incorporates regression into TopicVec mentioned in the previous section, and its regularized version, rTopicVec-Ridge.

**Generative process**

In rTopicVec, we assume that the document $d_i$ and an accompanying response variable $y_i$ are generated following the generative process as follows:

1. Generate words and topic assignments of each document $d_i$ following the generative process of TopicVec;
2. Draw response variable $y_i \sim \mathcal{N}(\eta^T \tilde{Z}_i, \delta^2)$.

Step 2 is newly added here to generate a response variable given the latent topics of the document generated in step 1. The expectation of the Gaussian distribution in step 2 is the inner product of the regression coefficients $\eta$ and the expectation of topic assignments $\tilde{Z}_i$ for $d_i$. Fig 2 presents a graphical model of rTopicVec. The orange colored circles correspond to step 2 in the generative process, indicating the nodes added to the previous graphical model in Fig 1. To prevent overfitting, we propose a regularized version named rTopicVec-Ridge, by further assuming standard normal priors on the coefficients $\eta$, which is equivalent to ridge regression.

**Estimation of parameters**

For rTopicVec, we estimate the parameters including the ones for regression using the generalized EM algorithm after deriving the loglikelihood function similar to TopicVec.

---

Fig 2. Graphical model of rTopicVec.

https://doi.org/10.1371/journal.pone.0277104.g002
First, we rewrite the complete data loglikelihood in (3) to include response variables \( y = \{y_i\} \) as:
\[
\log p(D, A, V, Z, T, \phi, y|\alpha, \gamma, \mu, \eta, \delta^2)
= C_0 - \log Z(H, \mu) - \|A\|_F^2 - \sum_{m=1}^M \mu_m \|v_m\|^2
+ \sum_{l=1}^M \left\{ \sum_{k=1}^K \log \phi_k(b_k + \alpha_k - 1) - \frac{1}{2} \log (2\pi\delta^2) \right.
- \frac{1}{2\delta^2} (\nu_j^2 - 2y_j\eta^T\nu_j) \left[ + \eta^T Z, Z, \eta \right] \\
+ \sum_{l=1}^M \left( \nu_{ij}^T \left( \sum_{j=1}^{l-1} \nu_{ij} + t_{ij} \right) + \sum_{j=1}^{l-1} a_{ij} \nu_{ij} + r_{ij} \right) \right\}
\]

Then by introducing a variational distribution \( q(Z, \phi, \pi, \theta) = q(\phi, \theta)q(Z; \pi) \) as in TopicVec, the expectation of the variational distribution of the loglikelihood of the response variable \( y_i \) is obtained by
\[
\mathbb{E}_q[\log p(y_i | Z, \eta, \delta^2)] = -\frac{1}{2} \log (2\pi\delta^2)
- \frac{1}{2\delta^2} (\nu_j^2 - 2y_j\eta^T\mathbb{E}_q[Z]) + \eta^T \mathbb{E}_q[Z, Z'] \eta,
\]
where
\[
\mathbb{E}_q[Z] = \pi_i = \frac{1}{L_i} \sum_{i=1}^{L_i} \pi_i, \\
\mathbb{E}_q[Z, Z'] = \frac{1}{L_i} \left( \sum_{i=1}^{L_i} \sum_{j \neq i} \pi_j \pi_j^T + \sum_{j=1}^{L_i} \text{diag}(\pi_j) \right).
\]

Thus, the objective \( \mathcal{L}_i(q; T) \) is obtained by adding (7) to (5):
\[
\mathcal{L}_i(q; T) = \sum_{l=1}^M \left\{ \sum_{k=1}^K (\sum_{j=1}^{l-1} \pi^2 - \alpha_k - 1)(\psi(\theta_k) - \psi(\theta_a)) \right.
+ \left( \frac{1}{2} \log (2\pi\delta^2) - \frac{\nu_j^2}{2\delta^2} \right) + \text{Tr} \left( T^T \sum_{j=1}^{l-1} \nu_{ij} \pi_j^T \right) \right.
+ \left( r^T \frac{\nu_j^T \nu_j}{L_i \delta^2} \sum_{j=1}^{l-1} \pi_j \right)
+ \left. \left( -\eta^T \cdot \frac{1}{2L_i \delta^2} \left( \sum_{j=1}^{l-1} \sum_{j' \neq j} \pi_{jj'} \pi_{jj'}^T + \sum_{j=1}^{l-1} \text{diag}(\pi_j) \eta \right) \right) \right\}
+ \mathcal{H}(q) + C_i
\]

Here, \( \psi(\cdot) \) is the digamma function. \( \theta_a \) and \( T \) are updated following the corresponding equations in E-step and M-step in Algorithm 1 respectively.

The solution is obtained by setting the partial derivative w.r.t. \( \pi_{ij}^k \) to 0 after isolating the terms containing \( \pi_{ij}^k \):
\[
\pi_{ij}^k \propto \exp \left\{ \psi(\theta_a) + v_{ij}^k t_k + r_k + \frac{2\eta^T \Pi^{(k)} \eta + (\eta_k)^2}{2L_i \delta^2} \right\},
\]
where

$$\Pi^{(i)}_{ij} := \sum_{j' \neq j}^{L} \pi'_{j'} (0^{(i)}, \ldots, 1^{(k)}, \ldots, 0^{(K)})^\top$$

$$+ (0^{(i)}, \ldots, 1^{(k)}, \ldots, 0^{(K)}) \sum_{j' \neq j}^{L} \pi'_{j'}$$

is the partial derivative of $\sum_{j=1}^{L} \sum_{j' \neq j}^{L} \pi_{j'} \pi_{j'}^\top$ w.r.t. $\pi_{j'}$.

Eq (7) contains the regression parameters in the learning objective. To involve the learning of the bias term in regression, we define a $M \times (K + 1)$ matrix $A$ whose row is a topic proportion vector of a document attached by a $1$: $(\overline{\pi}_i, 1)$ with the $(K + 1)$-th element corresponding to the bias. Over the whole corpus, Eq (7) can be rewritten as

$$\eta' = \text{Concat}(\eta, \eta_{\text{bias}}),$$

$$E_q[\log p(y|A, \eta', \delta^2)] = -\frac{M}{2} \log (2 \pi \delta^2) - \frac{1}{2 \delta^2} E_q[(y - A\eta')^\top (y - A\eta')],$$

where $\eta'$ is obtained by the function Concat($\cdot$) that concatenates the bias term to the end of the coefficients vector $\eta$. Taking the partial derivative w.r.t. $\eta'$ and $\delta^2$ and setting them to 0, we obtain the following to update $\eta'$ and $\delta^2$:

$$\eta'_{\text{new}} = (E_q[A^\top A])^{-1} E_q[A^\top y],$$

$$\delta^2_{\text{new}} = \frac{1}{M} \{y^\top y - y^\top E_q[A] (E_q[A^\top A])^{-1} E_q[A^\top y]\}$$

where we define $\pi_{j'} := \text{Concat}(\pi_{j'}, 1)$ to involve the bias term, and correspondingly

$$E[A] = \frac{1}{L_i} \sum_{j=1}^{L_i} \pi_{j'},$$

$$E[A^\top A] = \sum_{i=1}^{M} \left(\frac{1}{L_i} \left(\sum_{j=1}^{L_i} \sum_{j' \neq j}^{L_i} \pi_{j'} \pi_{j'}^\top + \sum_{j=1}^{L_i} \text{diag}(\pi_{j'})\right)\right).$$

For rTopicVec-Ridge, we use MAP estimation by adding $\ell_2$ regularization to Eq (10). Then similarly we obtain the following:

$$\eta'_{\text{new}} = (E_q[A^\top A] + \lambda \mathbf{I}_{\text{mod}})^{-1} E_q[A^\top y],$$

$$\delta^2_{\text{new}} = \frac{1}{M} \{y^\top y - y^\top E_q[A] (E_q[A^\top A] + \lambda \mathbf{I}_{\text{mod}})^{-1} E_q[A^\top y]\}$$

where $\lambda$ is the strength of the regularizer, and $\mathbf{I}_{\text{mod}} = \text{diag}(1^{(1)}, \ldots, 1^{(K)}, 0^{(K+1)})$ implies that the bias corresponding to the last element is excluded from the regularization.

**Experimental results**

To evaluate the prediction performance of our proposed models, we performed experiments on two prediction problems. The first is to predict stock return rates using news articles provided by Thomson Reuters and stock prices from the Tokyo Stock Exchange. The second experiment, as performed by McAuliffe et al. [10], is to predict movie rating scores using movie reviews. For each experiment, we first determined the number of topics for optimal prediction performance through validation tests with our proposed model rTopicVec.
We compared the performance of the proposed models rTopicVec and rTopicVec-Ridge with the following three baseline models:

**TopicVec+LR**: Perform linear regression as post-process using topics learned by TopicVec [5] as a baseline.

**TopicVec+Ridge**: Perform ridge regression as post-process using topics learned by TopicVec as a baseline.

**sLDA**: Supervised topic model [10] using BoW representations as a baseline.

The word embeddings $V$ for the two TopicVec-based models in the two experiments were trained by PSDVec using Japanese Wikipedia and English Wikipedia following PSDVec [9]. The Dirichlet hyperparameter $\alpha$ is fixed to $(0.1, \cdots, 0.1)$ [15] for all models in both experiments. The regularizer $\lambda$ is set to 1 for rTopicVec-Ridge and TopicVec+Ridge.

### Stock price return rates prediction

**Setup.** For text data, we used financial articles in Japanese distributed by Thomson Reuters from January 2015 to June 2017. We preprocessed the corpus by removing intractable tables and unneeded expressions, and performing morphological analysis using MeCab with mecab-ipadic-NEologd [16–18], a dictionary of neologisms and named entities, to segment words. We replaced stop words such as particles and conjunctions with $^*$ as the link function learns conditional distribution of a word in a context window. Moreover, we excluded low-frequency words occurring in fewer than five documents and documents of shorter than 50 words.

As response variables associated with the financial articles, the stock return rate of the company mentioned in each article is defined as following using the Tokyo Stock Exchange’s historical data for stock prices:

$$R = \frac{V_f - V_0}{V_f},$$

where $V_f$ is the final value on the day after the article was published and $V_0$ is the final value on the day before the article was published. When multiple companies appeared in one article, we sorted the return rates of those companies in descending order, then removed those whose absolute value was lower than the mean plus one standard deviation since the return rates of such companies may have not been affected by the content of the article. Articles that mentioned more than five companies were excluded because they likely focused on industry trends rather than specific companies.

As shown in Table 2, we divided the articles into five collections: those from the first half of 2015 (H1 2015), the latter half of 2015 (H2 2015), the first half of 2016 (H1 2016), the latter half of 2016 (H2 2016), and the first half of 2017 (H1 2017). Then we further divided each of the collections into two-monthly segments and prepared four preprocessed datasets in time order. We assumed that the topics in the financial markets would change gradually over time. Therefore, we set the data for adjacent terms to overlap by one month to capture this nature. We also show in the table the number of documents contained in each term of each dataset, as well as the average number of words per document (some of the documents were short as the aforementioned were not included. Nevertheless, there is no unfairness in the comparison of the models.).

To determine the optimal number of topics to be used in the open tests, 20% of the data in each training set were randomly drawn as hold-out for validation with ten different numbers
of topics $K \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$. A validation test on the data of an overlapped month was performed with the model parameters learned by the latter of two datasets containing the data of this month, e.g., the validation test on Feb. 2016 data was performed with the model parameters learned in Term 2 rather than those learned in Term 1.

The model parameters that would be used in the open tests on Term 4 would be learned with 100% of the data in each training set and the optimal number of topics $K$ determined by the validation tests for rTopicVec. During the training procedure, for Term 1, topic embeddings $T$ were initialized randomly from the standard Gaussian distribution, and the variational parameter $\pi$ was randomly initialized following a Dirichlet distribution. For Term 2, $T$ were initialized by the ones learned on Term 1. We followed the same procedure for the following terms, but we used $T$ and $\eta$ estimated with the previous term as the initial states of $T$ and $\eta$, respectively. The regression coefficient $\eta$ was updated every five iterations, and we used the same experimental procedure for the other models as that for rTopicVec.

For models except sLDA in the open tests, with the optimal $K^*$, we used TopicVec to estimate the topics on Term 4 with $T$ learned from Term 3 as the initial $T$. While for sLDA in the open tests, with the optimal $K^*$, we used LDA to learn the topics on Term 4 with the topic-terms distributions $\beta$ learned from Term 3 as the initial $\beta$.

For all experiments, the convergence condition was that the rate of change of $\pi$ (or $\phi$ in sLDA) must be lower than 0.1% three times in a row during learning.

The response variables are predicted as following:

$$\pi'_{ij} = \text{Concat}(\pi_{ij}, 1)$$

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

To measure the performance, we used the mean squared error (MSE) between the predicted
response variables and the ground truth:

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

**Results.** Fig 3 shows the MSEs obtained by the five models as the average result of the five validation tests. The solid line which denotes rTopicVec shows that the proposed model yields higher prediction accuracy when \( K = 30 \) on average for the five validation tests. The results of TopicVec+LR when \( K > 35 \) are not shown here since the coefficients are too large due to overfitting. Besides, the two models whose regression coefficients are penalized have higher prediction accuracy when \( K \) is larger than 30.

Table 3 presents the topics with two of the highest absolute values of the regression coefficients learned by all the models on Term 3 in Dataset 2 when \( K = 30 \). Each of the topics is represented by the ten most relevant words translated from the original Japanese. The topics of TopicVec+LR are not shown here because overfitting during the linear regression learning leads to coefficients that are too large. We argue that the topics learned by our proposed models presented here are more coherent and reflect a rise/drop on stock price return rates.

Since \( K^* = 30 \) was the optimal number of topics in validation tests, we performed open tests on test sets when \( K^* = 30 \). Table 4 shows that in predicting the return rates of the unlabeled articles, one or both of the proposed models are more accurate than the baselines in four out of the five open tests. We also performed the Wilcoxon signed-rank test and the Paired-t test.
between the predicted response variables of our proposed models and those of the three baseline models. We found that four out of the five cases in which at least one of our proposed models had lower MSE values, and the p-values were less than 10%, indicating that the prediction performance of our models was more accurate than the three baseline models with a statistically significant difference. However, the MSE of rTopicVec was marginally higher than that of the baselines on Term 4 of Dataset 5, which may be due to the slight overfitting that occurred during the training process, while was mitigated by $\ell_2$ regularization shown by rTopicVec-Ridge. Thus we argue that the proposed models offers the advantage of explainability on the relationship between the latent topics and regression coefficients.

### Table 3. Top 10 words of two topics with highest absolute values of regression coefficients.

| Model          | Top 10 words                                                                 |
|----------------|----------------------------------------------------------------------------|
| rTopicVec      | coefficient: 0.092 hot, stock, material, favor, rebound, leading, buying,   |
|                | the company, dividend, announcement                                          |
|                | coefficient: -0.055 disappoint, downward revision, hot, stock, linking,      |
|                | sound, deficit, fluctuation, bring down, reverse                            |
| rTopicVec-Ridge| coefficient: 0.062 favor, linking, increased profit, forecast, contribute,  |
|                | raise, hot, stock, dividend, achievement                                      |
|                | coefficient: -0.041 disappoint, hot, stock, rebound, softness, reverse,      |
|                | sound, newspaper, report, report                                            |
| TopicVec+Ridge | coefficient: 0.104 acquisition, treasury stock, stock, self, stock,        |
|                | conduct, issue, total, upper limit, hold                                    |
|                | coefficient: -0.071 money, rebuild, procurement, sponsor, debt, insolvency,  |
|                | investment, fund, cost, group                                               |
| sLDA           | coefficient: 0.078 dollar, near, the U.S.A., euro, not, rise, domestic,      |
|                | increase in interest rates, financial institutions, the day before           |
|                | coefficient: -0.063 announcement, forecast, linking, stock, hot, downward   |
|                | revision, disappoint, deficit, forecast, revision                           |

These words are translated from the original Japanese into English.

https://doi.org/10.1371/journal.pone.0277104.t003

between the predicted response variables of our proposed models and those of the three baseline models. We found that four out of the five cases in which at least one of our proposed models had lower MSE values, and the p-values were less than 10%, indicating that the prediction performance of our models was more accurate than the three baseline models with a statistically significant difference. However, the MSE of rTopicVec was marginally higher than that of the baselines on Term 4 of Dataset 5, which may be due to the slight overfitting that occurred during the training process, while was mitigated by $\ell_2$ regularization shown by rTopicVec-Ridge. Thus we argue that the proposed models offers the advantage of explainability on the relationship between the latent topics and regression coefficients.

### Table 4. MSE and sample standard deviation on the three test sets of stock price prediction when $K = 30$.

|                | May ~ June 2015 | Nov. ~ Dec. 2015 | May ~ June 2016 | Nov. ~ Dec. 2016 | May ~ June 2017 |
|----------------|-----------------|------------------|-----------------|------------------|-----------------|
| TopicVec+LR    | 0.00540\(^1\) | 0.00388\(^1\) | 0.00534\(^1\) | 0.00410\(^1\) | 0.00467\(^1\) |
| TopicVec+Ridge | 0.00523\(^1\) | 0.00435\(^1\) | 0.00495\(^1\) | 0.00410\(^1\) | 0.00509\(^1\) |
| sLDA           | 0.00562\(^1\) | 0.00390\(^1\) | 0.00484\(^1\) | 0.00382\(^1\) | 0.00776\(^1\) |
| rTopicVec      | 0.00493\(^1\) | 0.00310\(^1\) | 0.00474\(^1\) | 0.00383\(^1\) | 0.00818\(^1\) |
| rTopicVec-Ridge| 0.00498\(^1\) | 0.00319\(^1\) | 0.00436\(^1\) | 0.00413\(^1\) | 0.00818\(^1\) |

Bold face indicates the best performance for each test set.
- indicates that the MSE of TopicVec+LR is too large due to overfitting.
\(^1\) indicates that rTopicVec improved MSE with statistical significance using the Wilcoxon signed-rank test or the Paired-t test at the significance level of 0.1.
\(^\dagger\) indicates that rTopicVec-Ridge improved MSE with statistical significance using the Paired-t test or the Wilcoxon signed-rank test at the significance level of 0.1.

https://doi.org/10.1371/journal.pone.0277104.t004
Movie rating scores prediction

Setup. We additionally evaluated the prediction performance of our proposed models on predicting rating scores from movie reviews by 5-fold cross-validation as performed in sLDA [10]. The dataset was first used in Pang and Lee [19]. The corpus contains 5006 documents. The scores associated with the documents were transformed to approximate normality by taking logs as Blei and McAuliffe [10] did. 10% of the data randomly drawn from the dataset was used as a test set for the open test, and the remaining data was used for cross-validation, which was performed with ten different numbers of topics $K \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$ to determine the optimal number of topics $K^*$ with the smallest average MSE over the five validation tests with rTopicVec. The open test was performed using the model parameters learned with the remaining 90% of the dataset and the optimal number of topics $K^*$. We used MSE as the measure of prediction performance for the validation tests and the open test.

Results. Fig 4 shows the MSEs obtained as the average of the results of the 5-fold cross-validation. The solid line which denotes rTopicVec shows that $K = 15$ is the optimal number of topics in this prediction problem. The results of TopicVec+LR when $K > 35$ are not shown here since the coefficients are too large due to overfitting. Table 5 presents the topics with two of the highest absolute values of the regression coefficients learned by all the models on the training set when $K = 15$ except TopicVec+LR due to overfitting, and topic coherence measured by NPMI which is commonly applied to verify the topic quality. The NPMI is calculated by using reference counts from an external corpus (English Wikipedia), and is calculated by using the top ten words of each topic, averaging over all the NPMI scores of the topics. Despite

![Fig 4. Average MSE for varying number of topics K on validation sets of movie rating score prediction.](https://doi.org/10.1371/journal.pone.0277104.g004)
the averaged NPMI scores of our proposed models are lower than that of sLDA, the topic coherence increased by incorporating supervisory signals to the unsupervised TopicVec. Table 6 shows that in the open test when $K = 15$, the prediction performance of the proposed model rTopicVec is more accurate than that of the three baselines and two of them with a statistically significant difference. The lowest prediction accuracy achieved by the regularized version rTopicVec-Ridge may be due to underfitting caused by the regularization factor.

**Conclusions and discussions**

We proposed rTopicVec, a supervised topic embedding model combining a topic model in the embedding space and linear regression, and furthermore its regularized version rTopicVec-Ridge, to predict the numerical response variables labeled with documents. Through the experiments in predicting stock return rates using news articles and predicting movie ratings using movie reviews, the results showed that the prediction accuracy of our proposed models was higher than that of three baseline models since the topics learned by our proposed models are

| Table 5. Top 10 words of two topics with highest absolute values of regression coefficients and the topic coherence measured in NPMI on the training set when $K = 15$. |
|---------------------------------|
| Model (NPMI) | Top 10 words |
|-----------------|-----------------|
| rTopicVec / rTopicVec-Ridge (-0.035 / -0.046) | review, critical, task, offer, captivating, explanation, movie, director, turn, providing |
| | coefficient: 0.307 / 0.303 |
| | movie, captivating, task, director, offer, soul, debut, rarely, turn, breakdown |
| | coefficient: -0.362 / -0.406 |
| TopicVec+Ridge (-0.052) | review, offer, critical, explanation, movie, captivating, director, providing, looked, forces |
| | coefficient: 0.131 |
| | captivating, task, debut, screwball, trite melodramatic, movie’s, delightfully, perceptive, banal |
| | coefficient: -0.459 |
| sLDA (-0.031) | captivating, important, director, camera, explanation, task, material, club, twists, movie |
| | coefficient: -0.329 |
| | opinion, full, important, captivating, cuts, fear, minimum, problem, movie’s, michael |

https://doi.org/10.1371/journal.pone.0277104.t005

| Table 6. MSE and sample standard deviation on test set of movie rating score prediction when $K = 15$. |
|---------------------------------|
| Model | MSE | SD |
|-----------------|-----------------|
| TopicVec+LR | 0.13850$^{+}$ | 0.27308 |
| TopicVec+Ridge | 0.13842$^{+}$ | 0.27359 |
| sLDA | 0.13854$^{+}$ | 0.29474 |
| rTopicVec | 0.13477$^{†}$ | 0.26940 |
| rTopicVec-Ridge | 0.13979$^{†}$ | 0.27665 |

Bold face indicates the best performance for the test set.
$^{†}$ indicates that the proposed model (rTopicVec) improved MSE with statistical significance using the Wilcoxon signed-rank test at the significance level of 0.1.
$^{‡}$ indicates that the proposed model (rTopicVec) improved MSE with statistical significance using the Paired-t test at the significance level of 0.1.

https://doi.org/10.1371/journal.pone.0277104.t006
guided to be predictive of the response variables, and the topics were more coherent than those of TopicVec measured by NPMI and more interpretable to describe a rise/drop in the response variables. In summary, our models are capable of making more accurate predictions on the numerical labels and increasing the interpretability of topics by taking account of the associated labels while reducing the dimensionality of complex text data. We argue that incorporating word embeddings brought our models better prediction accuracy and comparable interpretability than the LDA-based supervised topic model like sLDA, and that learning topics and regression simultaneously brought our models the advantage of higher interpretability and accuracy of the predictions than the models that perform linear regression as post-process. Moreover, the overfitting that occurred in rTopicVec can be alleviated by putting priors on the regression coefficients. We also noticed that there is a huge gap between training and test errors for our proposed models. To narrow this gap as well as to prevent underfitting, the optimal regularization factor needs to be more explored, which will be left for future work. Furthermore, by using the state-of-the-art word embedding based on Transformer [20], the prediction performance and the quality of topics could be improved, which will also be left for future work. Our models can be applied to various applications of predicting response variables using text. Also, our models that assumed regression tasks in this paper can easily be modified to classification tasks.

Author Contributions

Supervision: Koji Eguchi.

Writing – original draft: Weiran Xu.

Writing – review & editing: Koji Eguchi.

References

1. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003; 3 (Jan):993–1022.
2. Dieng AB, Ruiz FJ, Blei DM. Topic modelling in embedding spaces. Transactions of the Association for Computational Linguistics. 2020; 8:439–453. https://doi.org/10.1162/tacl_a_00325
3. Pilehvar MT, Camacho-Collados J. Embeddings in natural language processing: Theory and advances in vector representations of meaning. Synthesis Lectures on Human Language Technologies. 2020; 13(4):1–175. https://doi.org/10.1007/978-3-031-02177-0
4. Das R, Zaheer M, Dyer C. 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); 2015. p. 795–804.
5. Li S, Chua TS, Zhu J, Miao C. Generative topic embedding: a continuous representation of documents. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2016. p. 666–675.
6. Li C, Wang H, Zhang Z, Sun A, Ma Z. Topic modeling for short texts with auxiliary word embeddings. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval; 2016. p. 165–174.
7. Shi B, Lam W, Jameel S, Schockaert S, Lai KP. Jointly learning word embeddings and latent topics. In: Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval; 2017. p. 375–384.
8. Zhao H, Du L, Buntine W. A word embeddings informed focused topic model. In: Asian conference on machine learning. PMLR; 2017. p. 423–438.
9. Li S, Zhu J, Miao C. A Generative Word Embedding Model and its Low Rank Positive Semidefinite Solution. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; 2015. p. 1599–1609.
10. McAuliffe J, Blei D. Supervised topic models. Advances in neural information processing systems. 2007;20.
11. Ramage D, Hall D, Nallapati R, Manning CD. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 conference on empirical methods in natural language processing; 2009. p. 248–256.

12. Zhu J, Ahmed A, Xing EP. MedLDA: maximum margin supervised topic models for regression and classification. In: Proceedings of the 26th annual international conference on machine learning; 2009. p. 1257–1264.

13. Wang X, Yang Y. Neural topic model with attention for supervised learning. In: International Conference on Artificial Intelligence and Statistics. PMLR; 2020. p. 1147–1156.

14. Xu W, Eguchi K. Topic Embedding Regression Model and its Application to Financial Texts. In: Proceedings of the Third Workshop on Financial Technology and Natural Language Processing; 2021. p. 15–21.

15. Asuncion A, Welling M, Smyth P, Teh YW. On Smoothing and Inference for Topic Models. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. UAI ’09. Arlington, Virginia, USA: AUAI Press; 2009. p. 27–34.

16. Toshinori S. Neologism dictionary based on the language resources on the Web for Mecab; 2015. Available from: https://github.com/neologd/mecab-ipadic-neologd.

17. Toshinori Sato TH, Okumura M. 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). Information Processing Society of Japan; 2016. p. NL–229–15.

18. Toshinori Sato TH, Okumura M. 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. The Association for Natural Language Processing; 2017. p. NLP2017–B6–1.

19. Pang B, Lee L. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05). Ann Arbor, Michigan: Association for Computational Linguistics; 2005. p. 115–124. Available from: https://aclanthology.org/P05-1015.

20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017:30.