Bioinformatics Advances, YYYY, 0–0
doi: 10.1093/ Bioinformatics Advances /xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Data and Text Mining

Turtling: A Time-aware Neural Topic Model on NIH Grant Data

Ruiyi Zhang[1], Ziheng Duan[1], CheYu Lee[1], Dylan Riffle[1], Martin Renqiang Min[2], Jing Zhang[1*]

[1]Department of Computer Science, University of California, Irvine, CA 92697, USA

[2]NEC Labs America, Princeton, NJ 08540, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Recent initiatives for federal grant transparency allow direct knowledge extraction from large volumes of grant texts, serving as a powerful alternative to traditional surveys. However, its computational modeling is challenging as grants are usually multifaceted with constantly-evolving topics.

Methods: We propose *Turtling*, a time-aware neural topic model with three unique characteristics. Firstly, *Turtling* employs pre-trained biomedical word embedding to extract research topics. Secondly, it leverages a probabilistic time-series model to allow smooth and coherent topic evolution. Lastly, *Turtling* leverages additional topic diversity loss and funding institute classification loss to improve topic quality and facilitate funding institute prediction.

Results: We apply *Turtling* on publicly available NIH grant text and show that it significantly outperforms other methods on topic quality metrics. We also demonstrate that *Turtling* can provide insights into research topic evolution by detecting topic trends across decades. In summary, *Turtling* may be a valuable tool for grant text analysis.

Availability: *Turtling* is freely available as an open-source software at https://github.com/aicb-ZhangLabs/Turtling

Contact: zhang.jing@uci.edu

Supplementary information: Supplementary data are available at *Bioinformatics Advances* online.

1 Introduction

Advances in machine learning algorithms and the recent initiatives for federal grant transparency have allowed direct knowledge extraction from large volumes of publicly-available online databases, potentially serving as a powerful alternative to traditional survey-based technologies. As a result, it is now possible to directly obtain quantitative and less biased grant text information that can broadly benefit scientific investigators, policy analysts, and funding agencies. Here, we aim to comprehensively navigate the funding landscape by exploring 466,730 public grant texts over the past 36 years from the National Institute of Health (NIH), the world's largest funding agency for biomedical research.

Computational modeling on NIH grant text data can be challenging for two reasons. First, NIH grant texts are usually multifaceted because they can be individually or jointly awarded from twenty-seven distinct Institutes/Centers (ICs) with overlapping priorities. Second, research topics have evolved quickly over the past decades as new technologies or health challenges have appeared (e.g., HIV and Covid pandemics in the 1980s and 2020s).

Previous researchers have leveraged topic models on NIH grant text to discover patterns reflecting latent research topics (Talley *et al.*, 2011). Topics learned from their methods are robustly correlated with specific NIH institutes, providing a basis for the discovery of interrelationships among biomedical concepts from NIH grant abstract documents. Later on, other researchers have used a labeled topic model to take the institute category information into consideration (Park *et al.*, 2016). Their work showed how text classification techniques can be used to analyze funding patterns of a specific institute. However, two problems limited the application of their models. First, training NIH data from scratch cannot capture rare word distributions. Second, while research topics have changed dramatically over the past twenty years, authors there used a static model that cannot capture temporal evolution information of research topics. Recently, some new topic modeling methods have been developed to capture topic trends in the general NLP area. (Dieng *et al.*, 2019, 2020; Blei and Lafferty, 2006; Blei *et al.*, 2003). Specifically, they use pre-trained word embeddings to improve their topic quality and probabilistic time series to allow topics to vary smoothly over time. Nevertheless, it is challenging to directly apply them to NIH grant data due to its rare biomedical terminologies and complicated institute category information.

To tackle these challenges, we propose *Turtling*, a time-aware neural topic model with multi-task losses, which encourages diverse topics and IC classification. *Turtling* has three unique characteristics compared with existing models. Firstly, *Turtling* extracts topics from biomedical word embedding space, lessening the word scarcity problem. Secondly, it

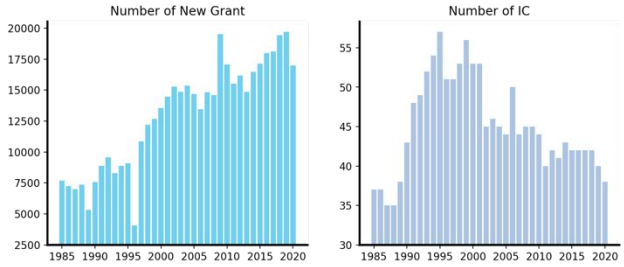

Figure 1. Statistics of the Grant dataset. Left panel shows the number of new grants every year from 1985 to 2020, and right panel shows the number of ICs every year.

leverages a probabilistic time-series model, which allows smooth and coherent topic evolution. Lastly, *Turtling* leverages additional topic diversity loss and IC classification loss to further improve extracted topic quality and topic correlation with specific NIH institutes. The losses above contribute to the extraction of diverse and high-quality topics that contain IC-specific information.

To verify its applicability, we have collected the *Grant* dataset, which includes 466,730 grant abstract documents and their corresponding ICs across 36 years (1985 to 2020). We tested the performance of *Turtling* against baseline methods on the extracted topic quality and IC prediction accuracy using *the Grant* dataset. Our experimental results showed that our method significantly outperformed baselines on topic coherence, diversity, and perplexity. Furthermore, we used our model to detect the topic trend across decades, providing valuable information on the evolution of research interests in the biomedical field. We then leveraged the topic proportions of a grant to predict its best-suited IC for success. We also found that grants from the same IC share similar topics in our visualizations as their topic proportion vectors were closer to each other, allowing for more interpretable predictions of IC selection given the grant abstract. In summary, our method provides an unbiased way for retrieving meaningful topics in NIH grants and its relation with NIH institutes and centers.

2 Methods

2.1 Dataset

We collect 466,730 grant abstract documents from the NIH RePORTER website offered by the NIH[1] to construct the *Grant* dataset. We download the raw text data from the RePORTER website updated on July 26th, 2022. The documents are across 36 years from 1985 to 2020. Each document is submitted to a certain Institute or Center (IC). **Figure 1** shows the number of new grants and new ICs every year. Among all ICs in our dataset, there are 62 that have been active for more than 10 years. As many grants receive funding for multiple years, we only include grants that received support for the first time.

We preprocess the *Grant* dataset by filtering out stop words and words with extremely high or low frequency. Specifically, we remove words that have a high frequency, appearing in more than 80% of a document, as well as words that have a frequency of less than 10 times in a document. We then use the Wordnet lemmatizer in NLTK to get the stem for each word (Bird and Loper, 2004). After preprocessing, we further remove documents that contain less than 10 words. In total, we obtained a vocabulary with 35,108 distinct words.

[1] https://reporter.nih.gov/

2.2 *Turtling's* Topic modeling with word embeddings

As shown in **Figure. 2**, *Turtling* adopts recent advances in probabilistic generative models of documents, such as Latent Dirichlet Allocation (LDA) and word embeddings (Dieng *et al.*, 2020; Blei *et al.*, 2003). Specifically, *Turtling* leverages vectorized word embeddings to calculate the word distribution for each topic and assumes that the semantically related word embeddings and topic embeddings are closer to each other in the embedding space (Mikolov, Sutskever, *et al.*, 2013; Mikolov, Chen, *et al.*, 2013).

Table 1. List of Symbols. We list the important symbols and notations used in this paper and briefly describe each symbol.

Symbol	Remark
$\mathrm{d_{tj}}$	BOW vector for the j-th document in year t
D_t	Document dataset at time t
θ_d	Topic Proportion of document d
β_k	Word distribution for topic k
α_k	Embedding for topic k
η_t	Prior of topic proportion at time t
z_{dn}	Topic assignment for n-th word in document d
w_{dn}	n-th word in document d
Cat	Categorical distribution
LN	Logistic normal distribution

As shown in **Table 1**, we use a vector $d_{tj} \in R^V$ to denote the bag of words (BOW) representation for the *j*-th document in year t, where V is the size of the vocabulary and t represents a specific year. We then use $D_t \in R^{N_t \times V}$ to denote the concatenation of all N_t vectors d_{tj} $(1 \le j \le N_t)$, where N_t is the number of grants for year t. Therefore, D_t is a matrix that contains BOW information for all of the grant documents in year t. We then use $D_t = \{D_1, D_2, \ldots D_T\}$ to denote our complete dataset, where T stands for the total number of years. For each BOW vector $d \in D_t$, we assign a corresponding label $y_d \in \{1, 2 \ldots M_t\}$ to the document based on the IC it was submitted to. M_t denotes the total number of ICs at a single year t.

We first consider the modeling process on a single year dataset. We define K topics $\beta_i (1 \le k \le K)$, where each topic is a word distribution over the vocabulary, and K topic embeddings $\alpha_k (1 \le k \le K)$ with the same dimension as word embeddings. The word embedding $\rho \in R^{L \times V}$ contains all of the words in the vocabulary, and L is the dimension of the embedding. We then calculate word distribution for each topic in equation (1) below.

$$\beta_k = Softmax(\rho^T \alpha_k) \; (1 \le k \le K) \# (1)$$

where $Softmax(z)_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$. In this way, it calculates the generative probability for each word in proportion to the cosine similarity between each word embedding and the topic embedding. In the document generation process, we sample each word from its corresponding topic using this generative probability.

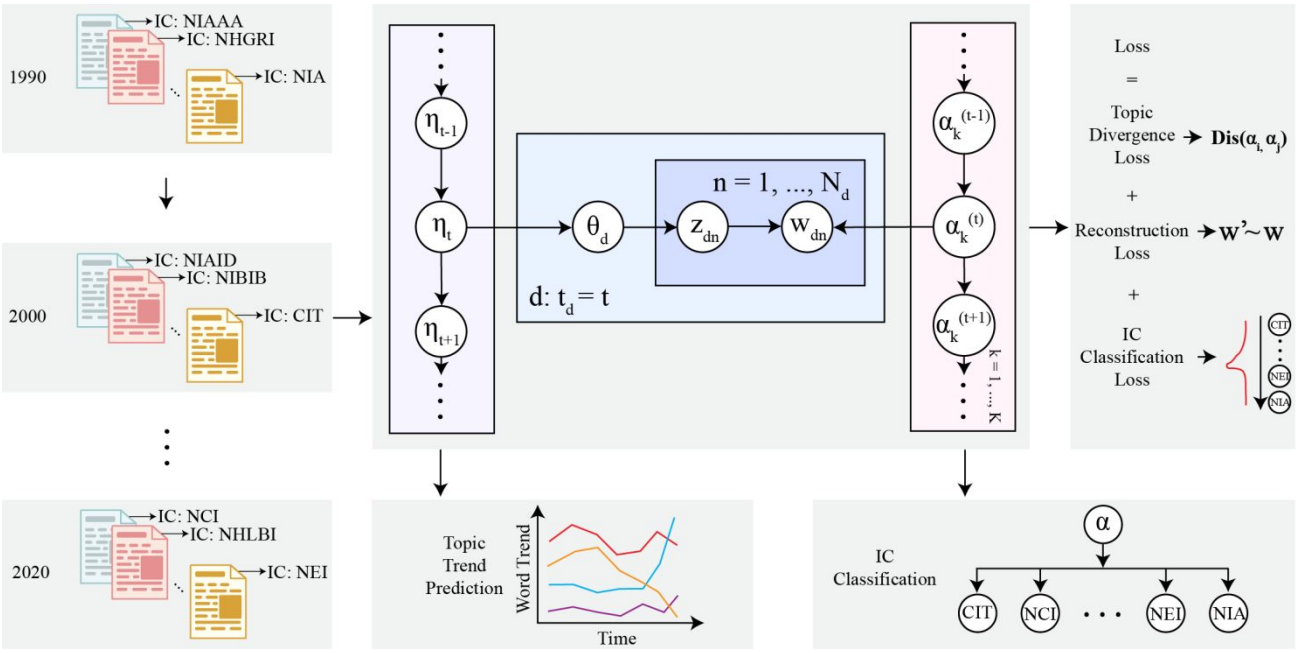

Figure 2. Flowchart of *Turtling*. *Turtling* leverages time-aware graphical topic model to extract high quality topics from grant documents across several years. The extracted topics can be used for several downstream tasks such as topic trend analysis and IC classification.

Then, we further consider a topic proportion vector θ_d with dimension K for each document, and each element of θ_d represents the probability of that topic to appear in document *d*. Formally, the generative process is as follows:

- Sample topic proportion $\theta_d \sim LN(0, I)$
- For n-th word w_{dn} in document *d*
  (a) Sample topic assignment $z_{dn} \sim Cat(\theta_d)(1 \le z_{dn} \le K)$
  Sample word $w_{dn} \sim Cat(\beta_{z_{dn}})$

where *LN* denotes the logistic normal distribution and *Cat* denotes the categorical distribution (Blei and Lafferty, 2007). z_{dn} is an integer that takes value from 1 to *K*.

2.3 Time-aware topic modeling

We then extend the method mentioned above to evolve dynamically on a multi-year dataset by allowing topics to vary smoothly over time. Within this model, the number of topics, denoted as K, remains consistent throughout all years, though the topic embeddings for each year exhibit slight variations compared to those from preceding years. Formally, for each time point *t*, *Turtling* defines a time specific topic embedding $\alpha_k^t \in R^L$. Similarly, it calculates the time specific word distribution $\beta_k^t \in R^V$ for each topic with the following formula:

$$\beta_k^t = Softmax(\rho^T \alpha_k^t)\ (1 \le k \le K) \#(2)$$

Different from the method in 2.2, the time specific topic distribution for each document θ_d^t is generated from a distribution that also evolves over time:

$$\theta_d^t \sim LN(\eta_t, \epsilon^2 I) \#(3)$$

where ϵ is a hyperparameter of the model and η_t is a latent variable that defines the prior mean of topic proportion at a specific time *t*. We assume that every η_t is a vector with dimension K generated by a random walk starting from η_{t-1} with Gaussian noise δ, so the conditional distribution of η_t given η_{t-1} is as follows:

$$p(\eta_t \mid \eta_{t-1}) = LN(\eta_{t-1}, \delta^2 I) \#(4)$$

Similarly, we assume the topic representation also evolves by random walk with Gaussian noise γ:

$$p(\alpha_k^t | \alpha_k^{t-1}) = LN(\alpha_k^{t-1}, \gamma^2 I) \#(5)$$

At time step $t = 0$, we assume both α_k^0 and η_0 follow Gaussian distribution $N(0, I)$. Thus, the generative process of *Turtling* can be summarized as:

1. Sample initial topic embeddings $\alpha_k^0 \sim N(0, I)$
2. Sample initial topic proportion mean $\eta_0 \sim N(0, I)$
3. For time step t= 1, 2…T:
   (a) Sample topic embeddings $\alpha_k^t \sim LN(\alpha_k^{t-1}, \gamma^2 I)$
   (b) Sample topic proportion mean $\eta_t \sim LN(\eta_{t-1}, \delta^2 I)$
   (c) Calculate $\beta_k^t = Softmax(\rho^T \alpha_k^t)$
4. For each document $d \in D_t$:
   (a) Sample topic proportion $\theta_d \sim LN(\eta_t, \epsilon I^2)$
   (b) For each word w_{dn} in document d:
      i. Sample topic assignment $z_{dn} \sim Cat(\theta_d)$
      ii. Sample word $w_{dn} \sim Cat(\beta_{z_{dn}}^t)$

Since *Turtling* learns topics in an embedded space, it can assign topics to words that do not appear in the training corpus as long as their embedding is given.

2.4 Inference of topic proportion and topic assignment

Given a word w_{dn} in document *d* at time *t*, we then calculate the marginal likelihood of w_{dn} to optimize the parameters. As we do not know the topic

proportion θ_d and topic assignment z_{dn} in the generative process, we have to marginalize both latent variables. We first marginalize the topic proportion θ_d, so the log likelihood $p(w_{dn} \mid \alpha^t, \rho)$ is defined as:

$$p(w_{dn} \mid \alpha^t, \rho) = \int p(\theta_d) p(w_{dn} \mid \theta_d, \alpha^t, \rho) d\theta_d \#(6)$$

We then marginalize topic assignment z_{dn} to compute the conditional distribution $p(w_{dn} | \theta_d, \alpha^t, \rho)$:

$$p(w_{dn} \mid \theta_d, \alpha^t, \rho) = \sum_{k=1}^{K} p(z_{dn} = k) p(w_{dn} \mid \beta^t_{z_{dn}}) \#(7)$$

After getting the log likelihood for each word, we then get the log likelihood loss function over parameter α^t and ρ:

$$L_{lk}(\alpha, \rho) = \sum_{t=1}^{T} \sum_{d \in D_t} \sum_{w \in d} \log\left(p(w \mid \alpha^t, \rho)\right) \#(8)$$

We use amortized variational inference to approximate the posterior distribution of topic proportion θ_d for document d (Kingma and Welling, 2013). Particularly, we use neural networks μ and θ that take document d as input to predict the mean and variance of a Gaussian distribution. This Gaussian distribution is then used as the approximated posterior distribution of θ_d. Formally:

$$q_\nu(\theta_d \mid d) = LN(\mu_\nu(D), \sigma_\nu(D)) \#(9)$$

where ν denotes the parameters of the inference neural networks. We leveraged a recurrent neural network as the inference model q in our implementation. This approximate distribution can be leveraged to compute the evidence lower bound (ELBO) of the marginal log likelihood. ELBO is a function of the generative model parameters α, ρ and the variational parameters ν:

$$L_{ELBO}(\alpha, \rho, \nu) = \sum_{t=1}^{T} \sum_{d \in D_t} \left(\sum_{w \in d} E_q\left[\log\left(p(w \mid \alpha^t, \rho)\right)\right] - KL(q_\nu | p(\theta_d)) \right) \#(10)$$

We then optimize L_{ELBO} with regard to parameters (α, ρ, ν) using minibatch Monte Carlo approximation.

2.5 Topic diversity loss

Inspired by the multi-task learning method, we optimize two additional loss terms mentioned in **Section 2.5** and **Section 2.6** (Ruder, 2017). We propose a topic diversity loss to make extracted topics more informative. This loss encourages each topic representation to be far away from each other in the training process. Formally,

$$L_{TD} = \sum_{t=1}^{T} \sum_{1 \le i,j \le k} Dis(\alpha_i^t, \alpha_j^t) \#(11)$$

where $Dis(x_1, x_2)$ can be any distance metric. Specifically, we use Euclidean distance in our model.

2.6 IC classification loss

We propose an IC classification loss to let inferenced topic proportions of each document contain information for IC prediction. In the training stage, a fully connected neural network $F(x)$ takes the inferenced topic proportion θ_d as the input and outputs a probability for each IC regarding which grant document might belong to it:

$$L_{IC} = \sum_{t=1}^{T} \sum_{d \in D_t} CE(F(\theta_d), y_d) \#(12)$$

where CE represents the cross-entropy loss. We then calculate the final loss function by adding up all three losses:

$$L(\alpha, \rho, \nu) = L_{ELBO} + \lambda_1 L_{TD} + \lambda_2 L_{IC} \#(13)$$

We optimize this loss function with gradient descent to compute the optimal topic representations α, word embeddings ρ, and variational parameters ν.

2.7 Evaluation Methods

We expect a good topic model to generate topics that are interpretable and informative. Moreover, these topics should be capable of reconstructing the original word distribution. Therefore, we evaluate the performance of our topic model using metrics including topic coherence, topic diversity and test perplexity (Mimno *et al.*, 2011; Rosen-Zvi *et al.*, 2004).

Topic coherence (TC) measures the similarity of words drawn from a topic, indicating whether the topic is semantically interpretable. Formally, we compute TC for a topic by selecting the top-p words from the topic and averaging over the similarity between any pair of words:

$$TC = \frac{1}{p^2} \sum_{1 \le i,j \le p} f(w_i, w_j) \#(14)$$

where w_i, w_j are drawn from the top-p words of a topic, and f is a similarity measure. In this paper, we choose 3 different functions for f: pairwise comparison bases on context window (CA), Fitelson's confirmation measure (CP) and normalized pointwise mutual information (NPMI) (Aletras and Stevenson, 2013; Röder *et al.*, 2015).

Topic diversity (TD) penalizes the repetitive or similar topics by calculating the repetitions of topic words. We use the proportion of unique top-p words in topics to compute TD in our paper. Formally,

$$TD = \frac{N_u}{K \times p} \#(15)$$

where K is the number of topics and N_u is the number of unique words. Perplexity measures the likelihood of a topic model on a held-out test dataset.

2.8 Experimental Settings

We utilize BioWordVec as the word embeddings for our method (Zhang *et al.*, 2019). BioWordVec encompasses 200-dimensional word embeddings trained on biomedical text with a biomedical controlled vocabulary, which are more suitable to NIH grant abstract text. Note that the parameters of the word embedding layer were also updated during the training process.

We use 85% of the *Grant* dataset for training, 5% for validation and 10% for testing. For the purpose of topic quality evaluation and trend analysis, we trained *Turtling* with a topic number of K=50. We set the

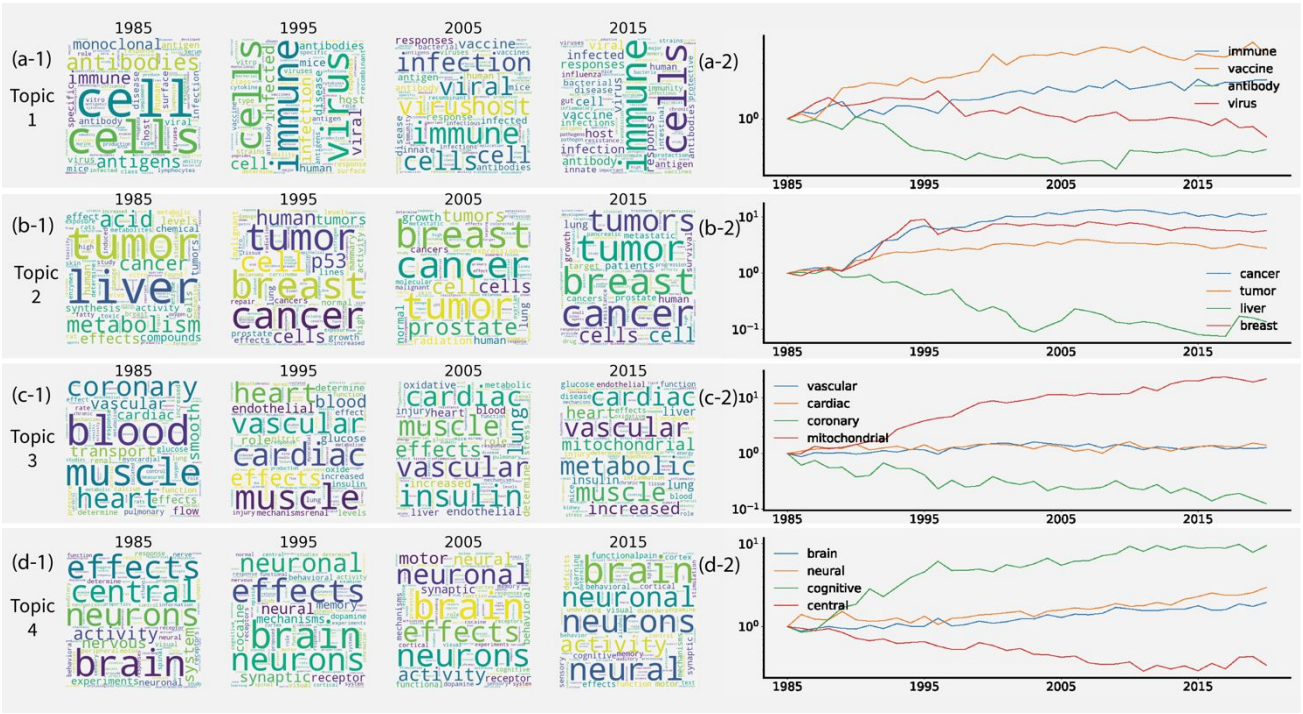

Figure 3. Wordcloud trend and keywords proportion trend for 4 topics across decades. For each topic, we selected 4 keywords and normalize their generative probability for each keyword. We then plot the normalized probability in each year from 1985 to 2020. We also select 4 specific years to create the wordcloud according to the generative probability of each topic.

learning rate of *Turtling* to be 0.001 with a small weight decay. We set the batch size to be 1024 and the dropout rate to be 0.1. We set the hyperparameters λ_1 and λ_2 in equation (13) to be 1 and 0.5. We set the hyperparameters ϵ, δ and γ in equation (3), (4), and (5) to be 0.01. We trained our model for 500 epochs on an Nvidia RTX 3090 GPU. We tested different choices of hyperparameters $K, \epsilon, \delta,$ and γ to select the best value above. Results for hyperparameters tuning are shown in **Supplementary Figure 1**.

In **Section 3.4**, we leveraged *Turtling* for IC classification. Specifically, we leveraged the topic proportion vector as the input feature to a random forest classifier, which is lighter and more interpretable compared to models using entire documents as input. For a fair comparison, we applied the PCA method to the bag-of-words representation of each document with the same output dimension as the number of topics. We also trained a DETM model and extracted topic proportions as input features. Here, we selected 20 as the number of topics. As sometimes we expected the model to predict several possible IC selections, we computed the top-5 accuracy as well as the top-1 accuracy. We also tested the performance of a neural network classifier instead of a random forest classifier and the results are shown in **Supplementary Figure 2.**

3 Results

Here, we applied *Turtling* on the *Grant* dataset and evaluated its performance on the extracted topic quality and IC classification accuracy, as discussed in the following sections. In **Section 3.1**, we evaluate the performance of our model and compare it with baseline methods on several topic quality metrics, demonstrating that *Turtling* improves the quality of extracted topics. In **Section 3.2**, we leverage the topics extracted by *Turtling* from the *Grant* dataset to analyze the research topic trend in recent years. In **Section 3.3**, we create a topic heatmap and the topic hierarchy to intuitively show the correlation between extracted topics. In **Section 3.4**, we use the topic proportions as an input feature to predict IC labels on the test dataset, indicating that topics extracted by *Turtling* are strongly correlated with the selection of NIH institutes.

3.1 Turtling improves topic quality from NIH grant text

Table 2. Topic quality results. We compared the performance of our model with several baseline topic models on topic coherence and topic divergence.

Method	CA	CP	NPMI	TD	Perplexity
ETM	0.13	0.17	0.015	0.82	2986.8
DETM	0.10	-0.2	0	0.52	3617.9
Turtling	**0.11**	**0.15**	**0.023**	**0.86**	**3120.7**

We applied *Turtling* on the *Grant* dataset and benchmarked its performance from three different aspects. First, we compared the baseline model DETM (Dieng *et al.*, 2019) and our model using topic coherence (CA, CP and NPMI), topic diversity (TD) and tested perplexity described in detail in **Section 2.7**. We also evaluated an ETM model on one year of data without time information (Dieng *et al.*, 2020). As shown in **Table 2**, *Turtling* outperformed DETM on all metrics, especially in TD and CP. Furthermore, *Turtling* achieved comparable topic quality results with the static topic modeling method ETM. Note that ETM was evaluated on a single-year dataset which is much smaller than the complete dataset the other two methods used, as ETM cannot capture the dynamic evolution of topics. We also compared *Turtling* with a non-generative topic modeling

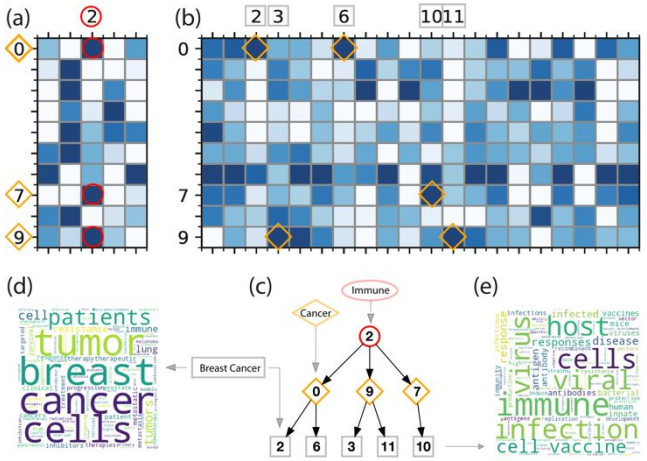

Figure 4. Heatmap and hierarchy trees for grant topics. We trained *Turtling* with 5 topics, 10 topics and 20 topics, and calculated the correlation factors between different topics.

method, BERTopic (Grootendorst, 2022). Results are shown in **Supplementary Table 1 and** *Turtling* also achieved competitive results on topic coherence and topic diversity.

3.2 Turtling highlights dynamic research topic changes over the past decades

As shown in the right part of **Figure 3**, we visualized the generative probability for some words with high generative probability in four example topics from 1985 to 2020. Note that in this plot, we normalized the generative probability for each keyword by setting the generative probability of this word in 1985 as 1 so that we can focus on the developing trend for each keyword across different years.

First, we observed clear trends of research topic and word distribution across years from our *Turtling* results. For instance, 'immune' and 'vaccine' (topic 1) related research has been increasingly attracting research attention within topic 1 since 1985 as shown in **Figure 3(a-2)**. Furthermore, within topic 2, breast cancer is one of the top increasing words, indicating significantly expanded funding opportunities in the past twenty years under this topic, as shown in **Figure 3(b-2)**. Similarly, mitochondrial and brain-related also research topics demonstrated a noticeable popularity gain in recent years. We further show the evolutionary trend of each topic of a 20-topic *Turtling* model in **Supplementary Figure 3.**

Next, we showed the temporal evolution of example words for biomedical research topics. For each of the most popular topics mentioned above, we listed some examples of top words in 1985, 1995, 2005, and 2015. To intuitively show the distribution of each word, we generated wordcloud for each topic at different time points. In wordcloud plots, larger fonts of words represent a higher generative probability of that word. The visualization results are shown in the left part of **Figure 3.**

Furthermore, we observed the keywords for each topic from the wordcloud across years. In 1985, 'blood' was a major concern in topic 3 which contains vascular related research, but 'cardiac' had been more popular since 1995. We also inferred the main topic name for each plot according to the top words in that topic. For example, given 'antibody', 'vaccine', and 'virus' in **Figure 3(a-1)**, we can infer that the research field for this topic is likely to be 'immune'.

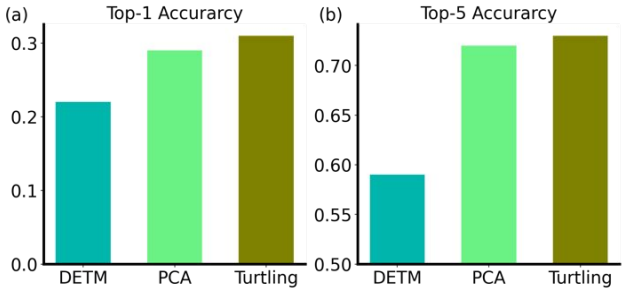

Figure 5. IC classification accuracy. We compared top-1 (a) and top-5 (b) accuracy of IC classification task using DETM, PCA and Turtling.

3.3 Turtling extracts hierarchy research topics relationships from Grant text

Next, we aim to explore the sub-fields of extracted research topics by examining connections of models trained with different topic numbers. As shown in **Figure 4**, we trained *Turtling* models with 5, 10, and 20 topics on the same collected grant text data. As a result, topics in the 5-topic model can be interpreted as broad research areas, while the sub-fields can be represented by topics in the 10- and 20-topic models. Consequently, the broad research area and subfield connections can be directly measured by the L_2 similarities of topic embeddings from different models.

We found that topic 2 in the 5-topic model is highly enriched in "immune" terminologies (red circle in **Figure 4a**, and **Figure 4b**). We explored its most closely associated sub-fields by calculating its most closely relevant topics in the subsequent 10 and 20-topic models, as shown in the heatmaps (**Figure 4. 4a-b**). For instance, topics 0, 7, and 9 in the 10-topic model showed the highest correlation with topic 2 in the 5-topic model. We can further trace down the higher resolution sub-fields in the 20 topic models by showing that topics 2 and 6, 3 and 11, and topic 10 are most connected to our subtopics in 10 topic models. We further extracted the word logo using the word frequencies in each topic and found that cancer and viral infection are important sub-fields for the "immune" topic we selected (**Figure 4c**). These results demonstrate that *Turtling*'s ability to extract hierarchical relationships between different research fields in a completely data-driven manner.

3.4 Turtling improves IC classification accuracy

Besides traditional research topic extraction tasks, an ideal grant analysis model should be able to accurately predict the funding IC and provide appropriate suggestions for future grant text data. Therefore, we further tested *Turtling's* performance on an IC classification task using the topic distributions (details in **Section 2.8**).

We benchmarked with traditional PCA and DETM models using top-1 and top-5 IC assignments. As shown in **Figure 5**, *Turtling* achieved a 31.6% top-1 accuracy, significantly higher than results from DETM and PCA (22.3% and 29.1% top-1 accuracy respectively). Furthermore, *Turtling* achieved a 73.8% top-5 accuracy which outperforms results from both methods (59.2% and 72.3% top-5 accuracy respectively). These experimental results showed that our method outperformed both of the baseline methods, demonstrating the effectiveness of using topic proportions generated by our model for IC classification.

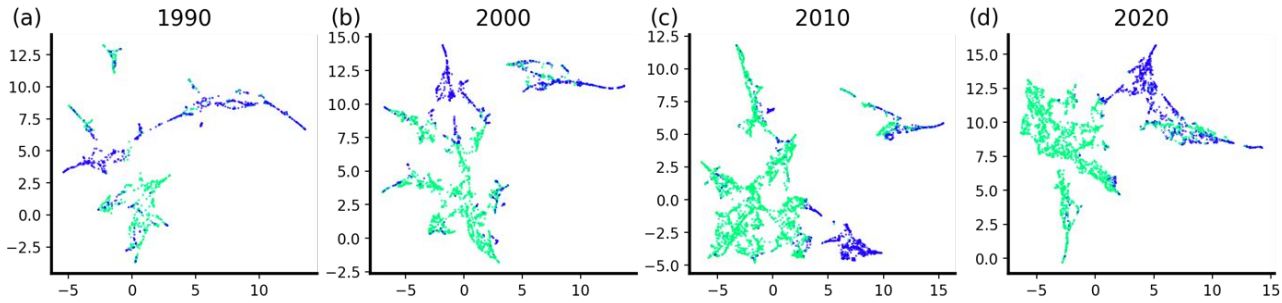

Figure 6. UMAP visualization of topic proportions. We leveraged UMAP to reduce the dimension of topic proportion vector for each document. The documents come from 'NCI' and 'NIMH' and they are represented by 2 different colors respectively.

3.5 Turtling separates documents from different ICs

To intuitively demonstrate topic proportion vectors generated by *Turtling* are separable among different ICs, we then visualized the vector of grant documents from two ICs in 1990, 2000, 2010, and 2020. We selected grants from the 'National Cancer Institute' (NCI) and the 'National Institute of Mental Health' (NIMH), as we expect the topics to vary significantly between these two ICs. We used UMAP to generate a 2-dimensional representation of topic proportion vectors for visualization (Mcinnes, 2018). The results are shown in **Figure 6.** Each dot with a certain color represents a document from a specific IC. We can observe from the plots that data points with different colors tend to form different clusters, indicating that each IC has its own topic preference.

To sum up, qualitative and quantitative analysis both show that the topic proportions generated by *Turtling* provide a useful and interpretable way for IC prediction tasks.

4 Discussion

In this paper, we developed *Turtling*, a time-aware topic model to analyze documents from a large grant corpus funded by the NIH. We constructed the *Grant* dataset, which contains 466,730 grant abstract documents and their corresponding ICs over the past 36 years. *Turtling* is novel with three main characteristics: the combination of biomedical word embedding and topic modeling, the time-aware nature of the graphical model, and the multi-task loss which includes topic divergence loss and IC classification loss.

We trained our model by optimizing the traditional ELBO as well as the topic diversity loss and the IC classification loss. Experimental results showed our method outperformed baseline methods on all of the metrics. We then leveraged *Turtling* to extract research topic trends from 1985 to 2020. We further demonstrated that the topic proportions generated by our method can be used for IC prediction.

In the future, we expect several extensions could be easily incorporated into our method for further performance improvement. Firstly, *Turtling* leveraged a naïve random forest classifier for IC classification, which could be substituted with more advanced deep classification models like transformers (Vaswani *et al.*, 2017). Second, pre-trained language models (PLMs) have become popular in many NLP applications (Devlin *et al.*, 2019; Peters *et al.*, 2018). Previous works have applied large PLMs to topic modeling tasks, but none of them considered the time-aware topic modeling scenario (Zhang *et al.*, 2022). As PLMs trained on biomedical text would contain large amounts of biomedical domain information, it may further improve the performance of topic models on the *Grant* dataset (Lee *et al.*, 2020). Lastly, the training process of *Turtling* is time-consuming due to its sequential inference strategy, posing a potential need for efficient inference and sampling methods.

We have implemented *Turtling* as an open-source software that is freely downloadable to the public. With the exponential growth of publicly-available grant text data, *Turtling* can be a valuable tool for investigators and funding agencies to gain research insights in a completely data driven manner.

Acknowledgement

Funding: This work was supported by the National Institutes of Health K01MH123896, R01HG012572, and R01NS128523.

References

Aletras,N. and Stevenson,M. (2013) Evaluating Topic Coherence Using Distributional Semantics. In, *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) -- Long Papers*. Association for Computational Linguistics, Potsdam, Germany, pp. 13–22.

Bird,S. and Loper,E. (2004) NLTK. In, *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions* -. Association for Computational Linguistics, Morristown, NJ, USA.

Blei,D.M. *et al.* (2003) Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.

Blei,D.M. and Lafferty,J.D. (2007) A correlated topic model of Science. *aoas*, **1**, 17–35.

Blei,D.M. and Lafferty,J.D. (2006) Dynamic topic models. In, *Proceedings of the 23rd international conference on Machine learning*, ICML '06. Association for Computing Machinery, New York, NY, USA, pp. 113–120.

Devlin,J. *et al.* (2019) BERT: Pre-training of deep bidirectional Transformers for language understanding.

Dieng,A.B. *et al.* (2019) The dynamic embedded topic model. *arXiv [cs.CL]*.

Dieng,A.B. *et al.* (2020) Topic modeling in embedding spaces. *Trans. Assoc. Comput. Linguist.*, **8**, 439–453.

Grootendorst,M. (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv [cs.CL]*.

Kingma,D.P. and Welling,M. (2013) Auto-Encoding Variational Bayes. In, *ICLR*.

Lee,J. *et al.* (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**, 1234–1240.

Mcinnes (2018) UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, **3(29)**, 861.

Mikolov,T., Sutskever,I., *et al.* (2013) Distributed Representations of Words and Phrases and their Compositionality. In, Burges,C.J. *et al.* (eds), *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Mikolov,T., Chen,K., *et al.* (2013) Efficient estimation of word representations in vector space. *arXiv [cs.CL]*.

Mimno,D. *et al.* (2011) Optimizing Semantic Coherence in Topic Models. In, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., pp. 262–272.

Park,J. *et al.* (2016) Analyzing NIH Funding Patterns over Time with Statistical Text Analysis. In, *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Peters,M. *et al.* (2018) Deep contextualized word representations. In, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, Stroudsburg, PA, USA.

Röder,M. *et al.* (2015) Exploring the Space of Topic Coherence Measures. In, *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15. Association for Computing Machinery, New York, NY, USA, pp. 399–408.

Rosen-Zvi,M. *et al.* (2004) The author-topic model for authors and documents. In, *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, UAI '04. AUAI Press, Arlington, Virginia, USA, pp. 487–494.

Ruder,S. (2017) An overview of multi-task learning in deep neural networks. *arXiv [cs.LG]*.

Talley,E.M. *et al.* (2011) Database of NIH grants using machine-learned categories and graphical clustering. *Nat. Methods*, **8**, 443–444.

Vaswani,A. *et al.* (2017) Attention is All you Need. In, Guyon,I. *et al.* (eds), *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Zhang,L. *et al.* (2022) Pre-training and fine-tuning neural topic model: A simple yet effective approach to incorporating external knowledge. In, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Stroudsburg, PA, USA.

Zhang,Y. *et al.* (2019) BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci. Data*, **6**, 52.