A Temperature-Modified Dynamic Embedded Topic Model

Anonymous submission

No Institute Given

Abstract. Topic models are natural language processing models that can parse large collections of documents and automatically discover their main topics. However, conventional topic models fail to capture how such topics change as the collections evolve. To amend this, various researchers have proposed dynamic versions which are able to extract sequences of topics from timestamped document collections. Moreover, a recently-proposed model, the dynamic embedded topic model (DETM), joins such a dynamic analysis with the representational power of word and topic embeddings. In this paper, we propose modifying its word probabilities with a temperature parameter that controls the smoothness/sharpness trade-off of the distributions in an attempt to increase the coherence of the extracted topics. Experimental results over a selection of the COVID-19 Open Research Dataset (CORD-19), the United Nations General Debate Corpus, and the ACL Title and Abstract dataset show that the proposed model, aptly nicknamed DETM-tau, has been able to improve the model's perplexity and topic coherence for all datasets.

Keywords: Topic models · neural topic models · dynamic topic models · dynamic embedded topic model· deep neural networks.

1 Introduction

Topic models are natural language processing (NLP) models which are able to extract the main topics from a given, usually large, collection of documents. In addition, topic models are able to identify the proportions of the topics in each of the individual documents in the given collection, which can be useful for their categorization and organization. As a machine learning approach, topic models are completely unsupervised and, as such, they have proved a very useful tool for the analysis of large amounts of unstructured textual data which would be impossible to tackle otherwise. Thanks to their flexibility and ease of use, topic models have found application in domains as diverse as finance [8, 18], news [25], agriculture [9], social media [1, 18], healthcare [2, 22, 27] and many others.

Among the topic models proposed to date, latent Dirichlet allocation (LDA) [7] is broadly regarded as the most popular. Its simple, fundamental assumption is that every word in each document of the given collection is associated with a specific "topic". In turn, a topic is represented simply as a dedicated probability distribution over the words in the given vocabulary. Completed by a Dirichlet prior assumption over the topic proportions of each document, LDA has proved at the same time accurate and efficient. However, conventional topic models such as LDA are unable to analyse the sequential evolution of the topics over different time frames. This could be important, instead, for collections that exhibit substantial evolution over time. For instance, a collection of COVID-19-related articles may predominantly display topics such as "outbreak" and "patient zero" in its early stages, "lockdowns" and "vaccine development" in later stages, and "vaccination rates" and "boosters" in the present day.

To analyze the topics over time, one could in principle just partition the document collection into adequate "time slices" (e.g., months or years), and apply a conventional topic model separately over each time slice. However, this would fail to capture the continuity and the smooth transitions of the topics over time. For this reason, Lafferty and Blei in [13] have proposed a *dynamic topic model* (DTM) which is able to extract the topics from each time slice while taking into account the topics' continuity and temporal dynamics. Motivated by the representational power of word embeddings in NLP, Diang et al. in [10] have recently proposed a *dynamic embedded topic model* (DETM) which integrates DTM with embedded word representations. Since word embeddings can be pre-trained in a completely unsupervised way over large amounts of text, an embedded model such as DETM can take advantage of the information captured by the word embeddings' pre-training.

However, a common limitation for all these topic models is that they cannot be easily tuned to explore improvements of the performance evaluation measures. For this reason, in this paper we propose adding a tunable parameter (a "temperature") to the word distributions of DETM to attempt increasing the model's performance. We have tested the proposed model, aptly nicknamed *DETM-tau*, over three diverse and probing datasets: a time-sliced subset of the COVID-19 Open Research Dataset (CORD-19) [24], the United Nations (UN) General Debate Corpus [3], and the ACL Title and Abstract Dataset [5], comparing it with the best dynamic topic models from the literature such as DTM and DETM. The experimental results show that the proposed model has been able to achieve higher topic coherence and also lower test-set perplexity than both DTM and DETM in all cases.

The rest of this paper is organized as follows: the related work is presented in Section 2, including a concise review of the key topic models. DETM is recapped in Section 3.1, while the proposed approach is presented in Section 3.2. The experiments and their results are presented in Section 4. Eventually, the conclusion is given in Section 5.

2 Related Work

In this section, we review the topic models that are closely related to the proposed work, such as latent Dirichlet allocation (LDA), dynamic topic models, and topic models based on word and topic embeddings.

Let us consider a document collection, D, with an overall vocabulary containing V distinct words. In LDA, the generic n-th word in the d-th document can be noted as $w_{d,n}$, and simply treated as a categorical variable taking values in index set $[1 \dots V]$. One of the key assumptions of LDA is that each such word is uniquely assigned to a corresponding *topic*, $z_{d,n}$, which is another categorical variable taking values in set $[1 \dots K]$, where K is the number of topics that we choose to extract from the collection. In turn, each topic has an associated probability distribution over the words in the vocabulary, $\beta_k, k = 1 \dots K$, which accounts for the word frequencies typical of that specific topic. The full model of LDA can be precisely formulated and understood in terms of the following *generative model*, which is a model able to generate "synthetic" documents by orderly sampling from all the relevant distributions:

- For the d-th document, draw a K-dimensional vector, θ_d, with its topic proportions:
  - $\theta_d \sim \mathrm{Dir}(\theta_d|\alpha)$
- For each word in the d-th document:
  - Draw its topic: $z_{d,n} \sim \mathrm{Cat}(\theta_d)$
  - Draw the word from the topic's word distribution:
  - $w_{d,n} \sim \mathrm{Cat}(\beta_{z_{d,n}})$

In the above model, the first step for each document is to sample its topic proportions, θ_d, from a suitable Dirichlet distribution, $\mathrm{Dir}(\theta_d|\alpha)$. Once the topic proportions are given, the next step is to sample all of the document's words, by first sampling a topic, $z_{d,n}$, from categorical distribution [1] $\mathrm{Cat}(\theta_d)$, and then sampling the corresponding word, $w_{d,n}$, from the word distribution indexed by $z_{d,n}$, $\mathrm{Cat}(\beta_{z_{d,n}})$.

Overall, LDA is a computationally-efficient model that can be used to accurately extract the topics of a given training set of documents, and simultaneously identify the topic proportions of each individual document. LDA can also be applied to a given *test set*; in this case, the parameters of the Dirichlet distribution, α, and the word distributions, β, are kept unchanged, and only the topic proportions for the given test documents are inferred. LDA has also spawned a large number of extensions and variants, including hierarchical versions [12, 16], sequential versions [21], class-supervised versions [21], sparse versions [19, 26, 28], and many others. However, the extensions that are closely relevant to our work are the dynamic topic model (DTM) [13], the embedded topic model (ETM) [11], and the dynamic embedded topic model (DETM) [10]. We briefly review DTM and ETM hereafter, while we recap DETM in greater detail in Section 3.

DTM is a topic model that captures the evolution of the topics in a corpus of documents that is sequentially organized (typically, along the time dimension).

[1] Otherwise known as the multinomial distribution. The recent literature on variational inference seems to prefer the "categorical distribution" diction.

The corpus is first divided up into "time slices" (i.e., all the documents sharing the same time slot), and then the topics are extracted from each slice taking into account a dynamic assumption. For reasons of inference efficiency, DTM uses a logistic normal distribution, $\mathcal{LN}(\theta|\alpha)$, instead of a Dirichlet distribution to model the topic proportions of the individual documents. In addition, the samples of the logistic normal distribution are obtained by explicitly sampling a Gaussian distribution of equivalent parameters, and then applying the softmax operator, $\sigma(\cdot)$, to the Gaussian samples. The sequential dependencies between the time slices are captured by a simple dynamical model:

$$\begin{aligned} \alpha^t &\sim \mathcal{N}(\alpha^{t-1}, \delta^2 I) \\ \beta^t &\sim \mathcal{N}(\beta^{t-1}, \sigma^2 I) \end{aligned} \tag{1}$$

where α^t are the parameters of the logistic normal distribution over the topics at time t, and β^t is the matrix of all the word distributions (in logit scale), also at time t. The rest of the generative model for slice t can be expressed as:

- For the d-th document, draw its topic proportions (logit scale):
  - $\theta_d \sim \mathcal{N}(\alpha^t, a^2 I)$.
- For each word in the d-th document:
  - Draw its topic: $z_{d,n} \sim \mathrm{Cat}(\sigma(\theta_d))$
  - Draw the word from the topic's word distribution:
  - $w_{d,n} \sim \mathrm{Cat}(\sigma(\beta_{z_{d,n}}))$

DTM has proved capable of good empirical performance, and its inference is provided by efficient variational methods [13]. However, both LDA and DTM might lead to poor modelling in the presence of very large vocabularies, especially if the corpus is not sufficiently large to allow accurate estimation of the word probabilities. A possible mollification consists of substantially pruning the vocabulary, typically by excluding the most common and least common words. However, this carries the risk of excluding important terms a priori. The embedded topic model (ETM) [11] aims to overcome the limitations of categorical word distributions such as those of LDA and DTM by leveraging *word embeddings* [4, 15].

In ETM, each distinct word in the vocabulary is represented as a point in a standard word embedding space (typically, 300-1024D). Each topic, too, is represented as a point (a sort of "average") in the same embedding space. The compatibility between a word and a topic is then simply assessed by their dot product, and the probability of the word given the topic is expressed as in a common logistic regression classifier. The full generative model of ETM can be given as:

- For the d-th document, draw its topic proportions (logit scale):
  - $\theta_d \sim \mathcal{N}(0, I)$
- For each word in the d-th document:
  - Draw its topic: $z_{d,n} \sim \mathrm{Cat}(\sigma(\theta_d))$
  - Draw the word from the topic's word distribution:
  - $w_{d,n} \sim \mathrm{Cat}(\sigma(\rho^\top \eta_{z_{d,n}}))$

In the above, we have noted as ρ the word embedding matrix, which contains the embeddings of all the words in the given vocabulary. Assuming a dimensionality of L for the embedding space, ρ's size is $L \times V$. In turn, with notation η_k we have noted the embedding of the k-th topic. Therefore, the dot product $\rho^\top \eta_k$ evaluates to a V-dimensional vector which, suitably normalised by the softmax, returns the probabilities for the word distribution of topic k.

The ETM is a powerful topic model that joins the advantages of LDA with the well-established word embeddings. The main benefit brought by the word embeddings is that they can be robustly pre-trained using large amounts of unsupervised text from a relevant domain (potentially, even the collection itself). During training of the ETM, a user can choose to either 1) use the pre-trained word embeddings, keeping them fixed, or 2) load them as initial values, but update them during training. In alternative, a user can also choose to update the word embeddings during training, but initialise them from arbitrary or random values (in this case, not taking advantage of pre-training). Dieng *et al.* in [11] have shown that the ETM has been able to achieve higher topic coherence and diversity than LDA and other contemporary models. While the ETM, like LDA, is limited to the analysis of static corpora, it can also be extended to incorporate dynamic assumptions. This is the aim of the dynamic embedded topic model (DETM) that we describe in the following section.

3 Methodology

In this section, we first describe our baseline, the dynamic embedded topic model (3.1), and then we present the proposed approach (3.2).

3.1 The dynamic embedded topic model

The dynamic embedded topic model (DETM) joins the benefits of DTM and ETM, allowing the model to capture the topics' evolution over time while leveraging the representational power of word embeddings. The dynamic assumption over the topic proportions is the same as for the DTM:

$$\alpha^t \sim \mathcal{N}(\alpha^{t-1}, \delta^2 I) \tag{2}$$

but a dynamic prior is now assumed over the topic embeddings:

Table 1. Key sizes of the datasets used for the experiments.

Dataset	Training set	Validation set	Test set	Timestamps	Vocabulary
CORD-19TM	15,300	900	1,800	18	70,601
UNGDC	1,96,290	11,563	23,097	46	12,466
ACL	8,936	527	1,051	31	35,108

$$\eta^t \sim \mathcal{N}(\eta^{t-1}, \gamma^2 I) \tag{3}$$

The rest of the generative model for slice t is:

- For the d-th document, draw its topic proportions (logit scale):
  - $\theta_d \sim \mathcal{N}(\alpha^t, a^2 I)$.
- For each word in the d-th document:
  - Draw its topic: $z_{d,n} \sim \mathrm{Cat}(\sigma(\theta_d))$
  - Draw the word from the topic's word distribution:
  - $w_{d,n} \sim \mathrm{Cat}(\sigma(\rho^\top \eta^t_{z_{d,n}}))$

The training of DETM involves maximizing the posterior distribution over the model's latent variables, $p(\theta, \eta, \alpha|D)$. However, maximizing the exact posterior is intractable. Therefore, the common approach is to approximate it with variational inference [6] using a factorized distribution, $q_v(\theta, \eta, \alpha|D)$. Its parameters, noted collectively as v, are optimized by minimizing the Kullback-Leibler (KL) divergence between the approximation and the posterior, which is equivalent to maximizing the following expectation lower bound (ELBO):

$$\mathcal{L}(v) = \mathbb{E}[\log p(\theta, \eta, \alpha, D) - \log q_v(\theta, \eta, \alpha|D)] \tag{4}$$

The implementation of q_v relies on feed-forward neural networks to predict the variational parameters, and on LSTMs to capture the temporal dependencies; we refer the reader to [10] for details.

3.2 The proposed approach: DETM-tau

The fundamental evaluation measure for a topic model is the *topic coherence* [14]. This measure looks at the "top" words in the word distribution of each topic, and counts how often they co-occur within each individual document. The assumption is that the higher the co-occurrence, the more "coherent" is the extracted topic model.

However, topic models cannot be trained to optimize the topic coherence. In the first place, the coherence is a counting measure that depends on the outcome of a ranking operation (a top-K argmax), and it is therefore not differentiable in the model's parameters. In the second place, it is evaluated globally over the entire document set. As a consequence, alternative approaches based on reinforcement learning [23] would prove excruciatingly slow, and would not be able to single out and reward the contribution of the individual documents (the so-called "credit assignment" problem [17]).

For this reason, in this work we attempt to improve the topic coherence by utilizing a softmax *with temperature* [20] in the word distributions. The inclusion of a temperature parameter can make the word distributions "sharper" (i.e. the probability mass more concentrated in the top words, for temperatures < 1) or smoother/more uniform (for temperatures > 1). We expect this to have an impact on the final word ranking, as high temperatures will make mixing more pronounced during training, while low temperatures may "freeze" the ranking to an extent. With the addition of the temperature parameter, τ, the word distributions take the form:

$$w \sim \mathrm{Cat}(\sigma(\rho^\top \eta_z / \tau)) \tag{5}$$

While parameter τ can be optimized with the training objective like all the other parameters, we prefer using a simple validation approach over a small, plausible range of values to select its optimal value.

4 Experiments and Results

4.1 Experimental set-up

For the experiments, we have used three popular document datasets: the COVID-19 Open Research Dataset (CORD-19) [24], the United Nation General Debate Corpus (UNGDC) [3] and the ACL Title and Abstract Dataset (ACL) [5]. CORD-19 is a resource about COVID-19 and related coronaviruses such as SARS and MERS, containing over 500,000 scholarly articles, of which 200,000 with full text. For our experiments, we have created a subset organized in monthly time slices between March 2020 and August 2021, limiting each slice to the first 1,000 documents in appearance order to limit the computational complexity. We refer to our subset as CORD-19TM, and we release it publicly for reproducibility of our experiments. UNGDC covers the corpus of texts of the UN General Debate statements from 1970 to 2015 annotated by country, session and year. For this dataset, we have considered yearly slices. The ACL dataset [5] includes 10,874 title and abstract pairs from the ACL Anthology Network which is a repository of computational linguistics and natural language processing articles. For this dataset, too, we have considered yearly slices, with the years spanning from 1973 to 2006 (NB: three years are missing). As in [10], the training, validation and test sets have been created by splitting the datasets into 85%, 5% and 10% splits, respectively. All the documents were preprocessed with tokenization, stemming and lemmatization, eliminating stop words and words with document frequency greater than 70% and less than 10%, as in [10].

Table 2. Results on the CORD-19TM dataset with 20 topics

Model	LDA	DTM	DETM	DETM-tau
Perplexity	—	—	15548.8	**14379.2**
Coherence	-0.049	0.114	0.059	**0.129**

Table 3. Results on the CORD-19TM dataset with 40 topics

Model	LDA	DTM	DETM	DETM-tau
Perplexity	—	—	14966.3	**13129.7**
Coherence	-0.047	0.081	-0.043	**0.093**

As models, we have compared the proposed DETM-tau with: the original DETM, DTM, and LDA applied separately to each individual time slice. As performance metrics, we have used the *perplexity* and the *topic coherence* which are the de-facto standards for this task. The perplexity is a measure derived from the probability assigned by the model to a document set, and should be as low as possible. It is typically measured over the test set to assess the model's generalization. The topic coherence is a measure of the co-occurrence of the "top" K words of each topic within single documents, and should be as high as possible. It is typically measured over the training set to assess the explanatory quality of the extracted topics. Several measures for the topic coherence have been proposed, and we use the NPMI coherence [14] with $K = 10$, as in [10]. As number of topics, we have chosen 20 and 40 which are commonly-used values in the literature. For the selection of the temperature parameter, τ, we have used range $[0.25 - 2.25]$ in 0.5 steps. All other hyperparameters have been left as in the corresponding original models.

4.2 Results

Tables 2 and 3 show the results over the CORD-19TM dataset with 20 and 40 topics, respectively. In terms or perplexity, the proposed DETM-tau has neatly outperformed the original DETM for both 20 and 40 topics (NB: the perplexity is not available for the LDA and DTM models). In terms of topic coherence, DETM-tau has, again, achieved the highest values. The second-best results have been achieved in both cases by DTM, while DETM and LDA have reported much lower scores. In particular, the very poor performance of LDA shows that applying a standard topic model separately on each time slice is an unsatisfactory approach, and musters further support for the use of dynamic topic models for timestamped document analysis.

Tables 4 and 5 show the results over the UNGDC and ACL datasets, respectively. For these datasets, we have not carried out experiments with DTM as it proved impractically time-consuming, and we omitted LDA outright because of its non-competitive performance. On both these datasets, too, DETM-tau has been able to achieve both lower perplexity and higher coherence than the original

Table 4. Results on the UNGDC dataset with 20 and 40 topics

Model	DETM	DETM-tau	DETM	DETM-tau
# topics	20		40	
Perplexity	3032.8	**3023.5**	2798.9	**2782.0**
Coherence	0.121	**0.129**	0.048	**0.124**

Table 5. Results on the ACL dataset with 20 and 40 topics

Model	DETM	DETM-tau	DETM	DETM-tau
# topics	20		40	
Perplexity	5536.4	**5421.1**	4360.0	**4169.6**
Coherence	0.150	**0.179**	0.153	**0.174**

Table 6. Examples of topics extracted by DETM-tau from the CORD-19TM dataset (20 topics) at different time slices.

Time slice	Examples of topics
0	zikv cytokine proinflammatory resuscitation ferritin antitumor exosomes thoracic evidencebased patienten cells infection cell virus blood disease protein tissue infected receptor patients patient health clinical care hospital months disease years therapy
10	exosomes copd frailty mgml tavi absorbance biofilm sigmaaldrich evidencebased virulence social education research health people services industry culture educational providers macrophages antibacterial antioxidant kshv mmp lmics propolis sdgs inactivation hydrogel patients studies health care patient clinical treatment disease population risk
17	nanoparticles proinflammatory bioactive antifungal inhospital coagulation angiogenesis inflammasome cells cell blood disease tissue cancer infection protein proteins metabolism patients health patient social education hospital clinical people care population

DETM. We believe that these results provide clear evidence of the importance of controlling the sharpness-smoothness trade-off of the word distributions.

To explore the sensitivity of the results to the temperature parameter, τ, Fig. 1 plots the values of the perplexity and the topic coherence of DETM-tau (CORD-19TM, 20 topics) for various values of τ, using DETM as the reference. It is clear that setting an appropriate value is important for the model's performance. However, the plots show that the proposed model has been able to outperform DETM for an ample range of values. In addition, Fig. 2 plots the values of the perplexity and the topic coherence at successive training epochs. The plots show that both metrics improve for both models as the training progresses. Given that the topic coherence is not an explicit training objective, its increase along the epochs is remarkable and gives evidence to the effective design of both models.

Eventually, we present a concise qualitative analysis of the extracted topics through Table 6 and Fig. 3. Table 6 shows a few examples of the topics extracted by DETM-tau from the CORD-19TM dataset (20 topics) at time slices 0, 10 and 17. Each topic is represented by its ten most frequent words. Overall, all the examples seem to enjoy good coherence and descriptive power. For instance, the first topic at time slice 0 could be titled "immune response analysis" or something

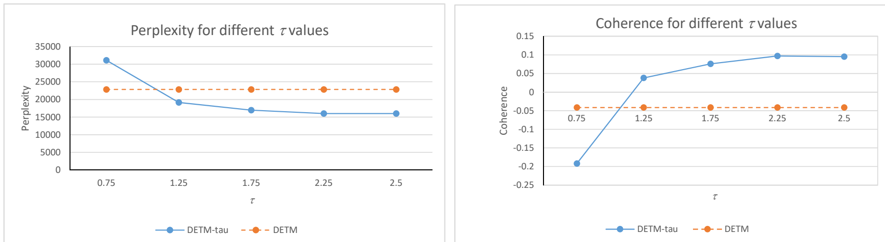

Fig. 1. Perplexity and topic coherence for DETM-tau for various values of the temperature parameter, τ (CORD-19TM, 20 topics). The value for DETM is used for comparison.

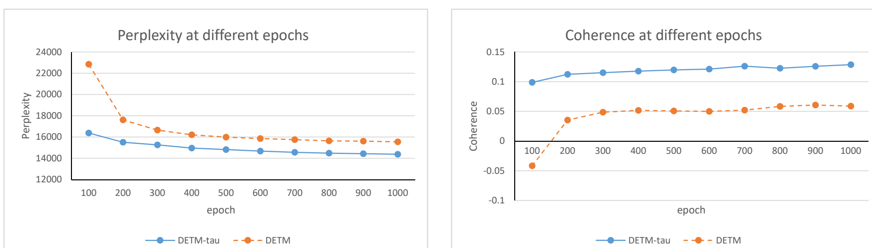

Fig. 2. Perplexity and topic coherence for DETM and DETM-tau at successive training epochs (CORD-19TM, 20 topics).

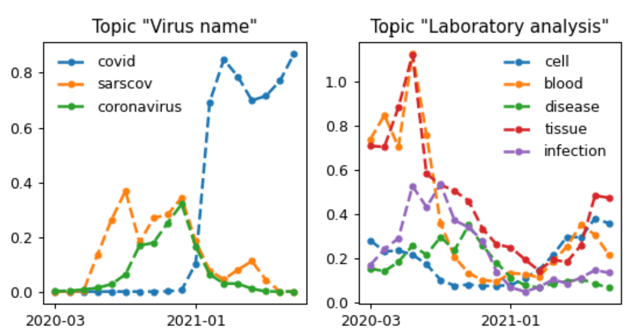

Fig. 3. Evolution of the probability of a few, selected words within their topics for the DETM-tau model with the CORD-19TM dataset, 20 topics.

akin; the last topic at time slice 17 could be titled "population health"; and so forth. Therefore, the automated categorization of the articles into such topics seem to provide a useful, and completely unsupervised, analysis. In turn, Fig. 3 shows the temporal evolution of the frequency of a few, manually selected words within their respective topics. The left-most topic, which we have labelled as "virus name", shows that referring to COVID-19 by the names "coronavirus" and "sarscov" was popular during 2020; conversely, as of January 2021, the name "covid" has become dominant. The right-most topic shows that words such as "blood", "infection" and "tissue" have decreased their in-topic frequency over time, possibly in correspondence with an increased understanding of the disease. These are just examples of the insights that can be obtained by dynamic topic models.

5 Conclusion

This paper has presented a temperature-modified dynamic embedded topic model for topic modelling of timestamped document collections. The proposed model uses a softmax with temperature over the word distributions to control their sharpness/smoothness trade-off and attempt to achieve a more effective parameterization of the overall topic model. Experiments carried out over three timestamped datasets (a subset of the CORD-19 dataset referred to as CORD-19TM, the United Nation General Debate Corpus (UNGDC) and the ACL Title and Abstract Dataset (ACL)) have showed that the proposed model, suitably nicknamed DETM-tau, has been able to outperform the original DETM model by significant margins in terms of both perplexity and topic coherence. In addition, DETM-tau has performed remarkably above the other compared models. A qualitative analysis of the results has showed that the proposed model has generally led to interpretable topics, and can offer insights into the evolution of the topics over time.

References

1. David Alvarez-Melis and Martin Saveski. Topic modeling in Twitter: Aggregating tweets by conversations. In *The 10th International Conference on Web and Social Media*, pages 519–522, 2016.
2. Corey Arnold, Suzie El-Saden, Alex Bui, and Ricky Taira. Clinical case-based retrieval using latent topic analysis. *AMIA Annual Symposium Proceedings*, 2010:26–30, 2010.
3. Mikhaylov SJ. Baturo A, Dasandi N. Understanding state preferences with text as data:. In *Introducing the UN General Debate corpus. Research Politics.*, 2017.
4. Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, 2003.
5. Steven Bird, Robert Dale, Bonnie J. Dorr, Bryan Gibson, Mark T. Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R. Radev, and Yee Fan Tan. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *International Conference on Language Resources and Evaluation*, pages 1755–1759, 2008.

6. David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. pages 859–877, 2017.
7. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.
8. Mark Cecchini, Haldun Aytug, Gary J. Koehler, and Praveen Pathak. Making words work: Using financial text as a predictor of financial events. *Decis. Support Syst.*, 50(1):164–175, 2010.
9. Dmitry Devyatkin, Elena Nechaeva, Roman Suvorov, and Ilya Tikhomirov. Mapping the research landscape of agricultural sciences. *Foresight and STI Governance*, 12(1):57–76, 2018.
10. Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. The dynamic embedded topic model. 2019.
11. Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.
12. Hannah Kim, Barry Drake, Alex Endert, and Haesun Park. Architext: Interactive hierarchical topic modeling. *IEEE Trans. Vis. Comput. Graphics*, 27(9):3644–3655, 2021.
13. J. D. Lafferty and David M. Blei. The dynamic topic model. In *The 23rd International Conference on Machine learning*, pages 113–120, 2006.
14. Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 530–539, 2014.
15. Quoc V. Le and Tomás Mikolov. Distributed representations of sentences and documents. In *The 31th International Conference on Machine Learning*, volume 32, pages 1188–1196, 2014.
16. Tengfei Liu, Nevin Lianwen Zhang, and Peixian Chen. Hierarchical latent tree analysis for topic detection. *CoRR*, 8725:256–272, 2014.
17. Marvin Minsky. Steps toward artificial intelligence. *Proc. IRE*, 49(1):8–30, 1961.
18. Thien Hai Nguyen and Kiyoaki Shirai. Topic modeling based sentiment analysis on social media for stock market prediction. In *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNL 2015)*, pages 1354–1364, 2015.
19. Min Peng, Qianqian Xie, Yanchun Zhang, Hua Wang, Xiuzhen Zhang, Jimin Huang, and Gang Tian. Neural sparse topical coding. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 2332–2340, 2018.
20. John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.
21. Filipe Rodrigues, Mariana Lourenco, Bernardete Ribeiro, and Francisco C. Pereira. Learning supervised topic models for classification and regression from crowds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12):2409–2422, 2017.
22. Efsun Sarioglu, Hyeong-Ah Choi, and Kabir Yadav. Clinical report classification using natural language processing and topic modeling. In *The 11th International Conference on Machine Learning and Applications*, volume 2, pages 204–209, 2012.
23. Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Second edition, 2018.

24. Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. CORD-19: The COVID-19 Open Research Dataset. In *1st Workshop on NLP for COVID-19 at ACL 2020*, volume 1, pages 1–12, 2020.
25. Guixian Xu, Yueting Meng, Zhan Chen, Xiaoyu Qiu, Changzhi Wang, and Haishen Yao. Research on topic detection and tracking for online news texts. *IEEE Access*, 7:58407–58418, 2019.
26. Aonan Zhang, Jun Zhu, and Bo Zhang. Sparse online topic models. In *The 22nd International World Wide Web Conference (WWW 2013)*, pages 1489–1500, 2013.
27. Rui Zhang, Serguei Pakhomov, Sophia Gladding, Michael Aylward, Emily Borman-Shoap, and Genevieve Melton. Automated assessment of medical training evaluation text. *AMIA Annual Symposium Proceedings*, 2012:1459–68, 2012.
28. Jun Zhu and Eric P. Xing. Sparse topical coding. In *The 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 831–838, 2011.