Towards Reinterpreting Neural Topic Models via Composite Activations

Jia Peng Lim
Singapore Management University
jiapeng.lim.2021@smu.edu.sg

Hady W. Lauw
Singapore Management University
hadywlauw@smu.edu.sg

Abstract

Most Neural Topic Models (NTM) use a variational auto-encoder framework producing K topics limited to the size of the encoder's output. These topics are interpreted through the selection of the top activated words via the weights or reconstructed vector of the decoder that are directly connected to each neuron. In this paper, we present a model-free two-stage process to reinterpret NTM and derive further insights on the state of the trained model. Firstly, building on the original information from a trained NTM, we generate a pool of potential candidate "composite topics" by exploiting possible co-occurrences within the original set of topics, which decouples the strict interpretation of topics from the original NTM. This is followed by a combinatorial formulation to select a final set of composite topics, which we evaluate for coherence and diversity on a large external corpus. Lastly, we employ a user study to derive further insights on the reinterpretation process.

1 Introduction

To help us understand the latent structures within a text corpus, topic models associate each document with "topics" (Blei et al., 2003). In turn, each topic is associated with a set of words that frequently co-occur together in various documents, forming a semantically coherent grouping that fosters interpretability. Aside from the common applications in text analysis and classifications, topic models are also used in advanced downstream tasks such as in summarization (Wang et al., 2020), text generation (Wang et al., 2019), and language modelling (Lau et al., 2017). While earlier topic models are based on graphical models, more recent topic models are neural, with several based on the variational auto-encoder framework (Kingma and Welling, 2014). Traditionally, what constitutes a topic is a neuron at the encoder's output. Its association with words is typically derived from a selection of the top activated words via the weights or reconstructed vector of the decoder connected to that neuron, forming what we now interpret as a topic-word distribution.

Motivation. While such autoencoder-based topic models are adept at learning lower-dimensional representations of documents, we question the notion of one-to-one correspondence between a topic and a neuron. We postulate that this traditional view belies the natural working order of a neural model, whereby it is the *joint* activation of several neurons, rather than the singular activation of an independent neuron, that may be responsible for the generation or reconstruction of document semantics. Moreover, the traditional interpretation of only the top activations in the resultant topic-word distribution ignores the potential information that might be gleaned from the rest of the distribution. We therefore hypothesize that individual neurons are but components of a "topic" that is inherently *compositional* in nature. And, to properly interpret an autoencoder-based topic model, we need to fully utilise the topic-word distribution space to uncover such compositions of neurons that frequently co-activate to collectively represent a semantic topic.

Approach. Given a generic class of trained neural topic model (NTM) (to be defined in Section 3) with K component (original) topics, we seek to reinterpret the NTM by finding a new set of K compositional topics that are more attuned to well-accepted measures of topic interpretability (also to be specified in Section 3). Each compositional topic is a linear combination of the original component topics. Inherently, the number of potential compositional topics are combinatorially explosive. Thus, we propose a two-stage process of *candidate generation* via mining the neural activations of various documents in the original corpus for frequently co-activated neurons, followed by *candidate selection* via solving optimization problems that map to classical algorithmic formulations with well-established computational properties.

Contributions. To our best knowledge, this is

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 3688 - 3703
December 7-11, 2022 ©2022 Association for Computational Linguistics

the first work to seek a reinterpretation of an NTM via compositional topics. We reiterate that our objective is not to replace, but to derive further quantitative insights on the state of the trained model. This reinterpretation process is model-free, as validated on a number of base NTMs (see Section 7.1).

Secondly, we propose an approach that aligns the mining of compositional topics to the objective of optimizing for well-accepted notions of topic interpretability. This approach is realized through principled formulations of frequent itemset mining for candidate generation (Section 5), as well as maximum independent sets and multi-dimensional knapsack for candidate selection (Section 6).

Thirdly, through quantitative measurements of interpretability on external large corpora, we show that the compositional topics tend to perform better than the original output of NTM's (Section 7).

Finally, as our core thrust is topic interpretability, we employ a user study to derive additional insights from the reinterpretation process (Section 8).

Implementation. Our gurobipy[1] and an alternative CVXPY[2] implementation can be found at `github.com/PreferredAI/ReIntNTM`.

2 Related Work

There are many neural topic models (NTMs), a comprehensive review can be found at Zhao et al. (2021). Primarily, the focus has been on creating better models, with numerous NTMs benchmarked in Doan and Hoang (2021). More detailed descriptions of our baseline NTMs used in the experiments can be found in Section 7.1. There are also notable research efforts to derive better interpretability of NTMs, such as through works focusing on topic sparsity (Lin et al., 2019; Gupta and Zhang, 2021), and through weakly supervised training (Meng et al., 2020).

Another popular approach to topic modelling involves using graph-based NTMs such as in Shen et al. (2021), Yang et al. (2020), and Zhang and Lauw (2020) which utilizes Graph Neural Networks, and/or, leveraging on graph representations of document/word/document-word relations and also through graph representations of higher-level entity metadata. The key distinction between our work and previous stand-alone graph-based NTMs is that our model-free approach is rooted in (non-neural network) classical selection problems with the choices (component topics) represented in a graphical manner.

Finally, there are other non-neural network-based topic modelling approaches such as online mean-field variational inference (Hoffman et al., 2010) and Non-negative matrix factorization (Zhao et al., 2017).

3 Preliminaries

Neural Topic Model (NTM). Let D denote a text corpus, K the desired number of topics, and N the vocabulary. An autoencoder-based NTM τ trained on D would produce a latent layer at the output of the encoder that we denote θ. The i^{th} neuron θ_i is referred to as an *original* or *component* topic. To associate θ_i with its topic words, we examine the topic-word decoder's weights or outputs due to the sole activation of θ_i. Considering the general case where a topic-word decoder has one of more hidden layers, we set $\theta_i = 1$ with the other $\theta_j = 0 \; \forall \theta_j \in \{\theta \setminus \theta_i\}$. Passing this input through the decoder, $\forall \theta_i \in \theta$, creates a $K \times |N|$ topic-word relation matrix β. Taking the l top-activated words from each row in β produces a topic set $\mathcal{T} = \{\mathcal{T}_i\}_{i=1,\ldots,K}$ consisting of K number of l-sized word sets $\mathcal{T}_i$, using the top activated words in each row of β.

Normalised Point-wise Mutual Information (NPMI). Introduced in Bouma (2009) and evaluated for texts in Aletras and Stevenson (2013) and Lau et al. (2014), this is a popular metric used for evaluating $\mathcal{T}$. In Röder et al. (2015), it is shown that NPMI has a good correlation with human ratings and the least sensitive to changes in the windows size parameter. This metric ranges from -1, suggesting incoherence, to 1, suggesting coherence within the topic. Let n represent a word in vocabulary N.

$$npmi(n_i, n_j) = \frac{log \frac{p(n_i,n_j)}{p(n_i)p(n_j)}}{-log(p(n_i, n_j))} \tag{1}$$

$$\text{NPMI}(\mathcal{T}) = \frac{1}{K} \sum_{t \in \mathcal{T}} \frac{\sum_{n_i \in t} \sum_{\substack{n_j \in t, \\ n_j \neq n_i}} npmi(n_i, n_j)}{l(l-1)/2} \tag{2}$$

Topic Uniqueness (TU). We seek to obtain K diverse topics (each of which is coherent), rather than a repetition of the same coherent topics multiple times. An intuitive measure is to count how many unique words are collectively represented by

[1] www.gurobi.com

[2] www.cvxpy.org

Variable	Definition
τ	Trained Neural Topic Model
β	Topic-word relation matrix for τ
$\hat{\beta}$	Reinterpretation of β
C	Composite interaction matrix
D	Set of documents in a corpus
ϵ	Topic uniqueness hyper-parameter constraint
K	Hyper-parameter for number of topics in τ
N	Vocabulary of D
n	Word in N
s	Min. support hyper-parameter for Apriori
Θ	Document-topic relations matrix for D
$\theta_{d,k}$	Document-d : topic-k relations for $d \in D$
$\mathcal{T}$	Set of original component topics
$\hat{\mathcal{T}}$	Set of new composite topics
V	Possible set of composite topics
v	Composite topic in V
w_v	Weight representing coherence score of v
x_n	Binary variable that denotes selection of n
x_v	Binary variable that denotes selection of v

Table 1: Table of Notations

the K topics (more unique words means less repetition). TU is defined as a percentage of unique words in the topic set (Dieng et al., 2020; Bianchi et al., 2021a). This ranges from $\frac{1}{K}$ to 1, with 1 implying that each topic is unique and each word occurs once in $\mathcal{T}$.

$$TU(\mathcal{T}) = |\cup_{t \in \mathcal{T}} \{n_t \in t\}|/(l \cdot K) \quad (3)$$

4 Overview

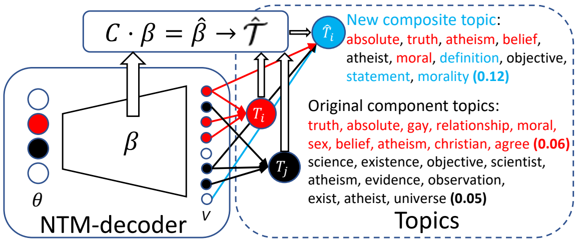

$C \cdot \beta = \hat{\beta} \rightarrow \hat{\mathcal{T}}$
$\hat{T}_i$
New composite topic:
absolute, truth, atheism, belief, atheist, moral, definition, objective, statement, morality (0.12)
Original component topics:
truth, absolute, gay, relationship, moral, sex, belief, atheism, christian, agree (0.06)
science, existence, objective, scientist, atheism, evidence, observation, exist, atheist, universe (0.05)
T_i
T_j
β
θ
V
NTM-decoder
Topics

Figure 1: In this example, the new composite topic is derived from two original component topics. Examples are drawn from ProdLDA on 20NewsGroup at $K = 50$. Coherence scores in parenthesis.

Classically, the interpretation of NTM, after training on D, is as-is via β. This assumes independence within θ and that the τ's complexity is surface-deep. Since neurons work together in a composite manner to optimize a loss function, we believe that these composite interactions C within θ has the potential to produce a better interpretation of τ. As shown in Fig. 1, we seek to find a $C \in \mathbb{R}^{L \times K}$ that interacts with $\beta^{K \times |N|}$ to form a better reinterpretation $\hat{\beta}$ to produce new topic set $\hat{\mathcal{T}}$ with K topics[3]. The sum of each row in C is constrained to 1, reflecting the components' weight in the composite topic, sufficiently representing all possible compositions.

For simplicity and without loss of generality, we consider the case where components are evenly-weighted in each composition. Modelling the compositions within β results in a binary combinatorial search space ${}_{2^K}C_K$. The difficulty of selecting the best C is further increased as it involves optimizing for two potentially-diverging objectives as there exist solutions that result in high coherence with low diversity and vice versa. Common strategies to solve for multiple objectives include min-max and weighted-sum. Cho et al. (2017) has a comprehensive survey on solving Multi-Objective Systems. We employ ϵ-programming, where we focus on NPMI objective while converting TU objective into a soft constraint.

Problem 1 (Reinterpreting NTM)**.** *Given β from a NTM τ. Find a $K \times K$ composite matrix C that produces a new reinterpretation $\hat{\beta} \in \mathbb{R}^{K \times |N|}$ where $C \cdot \beta = \hat{\beta}$. Where $\hat{\mathcal{T}}$ with K topics, $\hat{\mathcal{T}} = \{top\text{-}l(\hat{b}_i) | \hat{b}_i \in \hat{\beta}, \hat{b}_i \in \mathbb{R}^{1 \times |N|}\}$, is derived from $\hat{\beta}$ and maximizes the primary objective NPMI($\hat{\mathcal{T}}$) and secondary objective TU($\hat{\mathcal{T}}$) with soft constraints ϵ.*

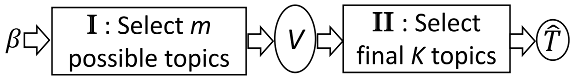


Figure 2: Two-stage reinterpretation process.

Proposed Approach. In Stage I, Topic Candidate Generation seeks to identify a pool of candidate topics V of feasible size m from the exponential number of possible compositions. In Stage II, Topic Selection uses several proposed formulations relying on ϵ to pick the final K composite topics, from V, to produce $\hat{\mathcal{T}}$ that has high NPMI and TU. We elaborate on each of these stages in the coming sections.

5 Stage I: Topic Candidate Generation Based on Neural Activation Profiles

We make the critical observation that which neurons tend to co-activate with one another can be

[3] While in general, the new topic set could be larger or smaller, for parity in this paper we set them to be the same as the original number of topics. Hence, $L = K$.

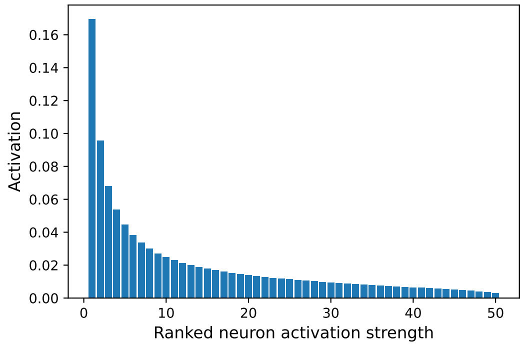


Figure 3: Activation profiles of latent neurons θ, of ProdLDA trained on 20NewsGroup at $K = 50$, sorted by decreasing activation strength. The top activated neuron strength across all documents has a mean value of 0.16. All the models included in the experiments follows a similar Pareto distribution pattern.

mined from the pattern of neural activations of the documents within a corpus. From Figure 3, the activation distribution of τ on D in layer θ is similar to a pareto distribution, with a only a few neurons being responsible for most of the activation strength. For practical purposes, we limit the size of compositions of up to five different component topics. Leveraging on D and τ, producing document-topic relations $\Theta \in \mathbb{R}^{|D|\times K}$, we can find frequently occurring compositions in D.

We can transform our current search problem to the Frequent Itemset Mining problem (FIM) (Agrawal and Srikant, 1994). The input to FIM is a set of transactions, where each transaction is a basket of items. The objective is to output all frequent itemsets, i.e., subsets of items that occur in at least s (minimum support) of the transactions.

In our context, each transaction is a document, and each item is an activated neuron. We set the minimum activation threshold κ to the fifth-largest mean activation value for Θ. For each document d, we set its $\theta_{d,k} = 1 \iff \theta_{d,k} > \kappa$ else 0, creating boolean "itemsets" (essentially baskets of co-activated neurons). Hyper-parameter minimum support s controls the size of V (setting larger values of s resulting in fewer candidates). While there are many solution approaches to FIM (Savasere et al., 1995; Toivonen, 1996), we leverage the Apriori algorithm[4] (Agrawal and Srikant, 1994).

The resulting frequent itemsets $\hat{C}$ (each itemset specifying a few co-activated neurons) generate candidate pool $V = \{\text{top-l}(b) | b \in \hat{C} \cdot \beta\}$, i.e., each $v \in V$ is a set of top-l words due to the corresponding composition of topics in an "itemset".

6 Stage II: Diversity-Constrained Coherence-Optimizing Topic Selection

We now seek to reduce V to find the final K composite topics that represent C, by optimizing for NPMI, as evaluated on D[5]. However, due to the way that diversity-oriented constraint ϵ could be formulated, this gives rise to a couple of formulation variants as outlined below.

6.1 Maximum-Weight Budget Independent Set (MWBIS)

Suppose that candidates V are vertices in a graph $G(V, E)$. An edge $(v_i, v_j) \in E$ exists if the corresponding candidate topics have more than ϵ number of similar words. To ensure diversity, we seek an independent set of unconnected vertices in G. Because we could only accommodate K topics, the independent set must be budgeted or capped in size to K. Because there are many possible K-sized independent sets, we seek the one with the maximum weight, which is the coherence score NPMI.

Mixed Integer Program. Formulating it as a mixed integer problem (MIP), we have an objective (4) with budget constraints (5) to (7). Binary x_v represents whether a topic $v \in V$ is selected, and w_v, representing NPMI of v. Constraint (5) allows us to have negative weights. Constraint (7) restrict the number of times a word can appear in $\hat{\mathcal{T}}$.

$$\max_{v} \sum_{v} x_v w_v \tag{4}$$

$$\text{s.t} \quad \sum_{v}^{V} x_v = K \tag{5}$$

$$x_i + x_j \leq 1, \forall i, j \in E \tag{6}$$

$$\sum_{j=1,\ldots,K | n \in \mathcal{T}_j} x_j \leq \epsilon + 1, \forall n \in N \tag{7}$$

This formulation is essentially a maximum-weighted budget independent set problem (MWBIS) Kalra et al. (2017), which is a variant of the well-established maximum-weighted independent set problem (MWIS), a known NP-hard problem for general graphs (Garey and Johnson, 1979). Even so, this could still be solvable for smaller

[4] http://rasbt.github.io/mlxtend

[5] This is only for training. For testing, we evaluate NPMI based on large external corpora.

graphs, particularly with the help of numerical solvers capable of approximations.

Greedy Algorithm. We introduce this simple approach, that mirrors the formulation of our MW-BIS formulation, as a fallback approach when it is infeasible to use solvers. It employs two heuristics f and g. f ensures that each $\mathcal{T}_i \in \hat{\mathcal{T}}$ is no more than ϵ-similar of each other. g ensures that each word occurs at most $\epsilon + 1$ times in $\hat{\mathcal{T}}$. The procedure iterates on V, sorted by NPMI, greedily choosing v, popped from V, if adding v to $\hat{\mathcal{T}}$ fulfils f and g. If we do not have K topics after a complete iteration, we increment ϵ, bounded by l, and repeat iteration, terminating procedure upon selecting K topics.

From Austrin et al. (2009), assuming unique games conjecture (Khot, 2002) and $P \neq NP$, they prove that there is no $\Omega(\frac{\log^2 \triangle}{\triangle})$-factor polynomial time approximation algorithm for MWIS in a degree-$\triangle$ bounded graph when $\triangle$ is sufficiently large. According to Kalra et al. (2017), the hardness result of MWIS applies to MWBIS as well. While we are unable to ensure optimal bounds for the greedy solution, it performs well empirically for a reasonable size of V (see Section 7).

6.2 Multi-Dimensional Knapsack Problem (MDKP)

In addressing diversity, the previous formulation seeks to reduce overlap between pairs of candidates. An alternative diversity constraint could be to seek some minimum number of unique words among the selected topic candidates.

Again, we maximize the similar objective (4) with budget constraint (5) and treat the number of unique words as a budget to exceed in (8). For our experiments, we set ϵ_{MDKP} to the number of unique words in the original $\mathcal{T}$, i.e., $|\{n_v \in v | v \in \mathcal{T}\}|$.

$$| \cup_{v \in V | x_v = 1} \{v\}| \geq \epsilon_{MDKP} \tag{8}$$

This formulation transforms our problem into a 0/1 Multi-dimensional Knapsack Problem (MDKP) (Martello and Toth, 1990), a NP-hard problem (Chu and Beasley, 1998). It is also noted in Laabadi et al. (2018) that available heuristics and metaheuristics approaches for MDKP did not ensure optimality.

7 Experiments

The primary objective of the following experiments is to investigate the efficacy of the terpretation process, i.e, whether the discovered composite topics via our methodology would outperform the component topics from the input NTMs (denoted *Original* in result tables) in terms of NPMI and TU.

Name	#Docs	#Words	#Labels
20NewsGroup	16,309	1,612	20
BBC-News	2,225	2,949	5
DBLP	54,595	1,513	4
M10	8,355	1,696	10

Table 2: Characteristics of text corpora used in experiments

7.1 Base Neural Topic Models

As previously asserted, our reinterpretation process is model-free, accommodating various NTMs. In this sub-section we describe the NTMs used in our experiments. There are 3 encoder parameters that we optimize for with respect to D: 1) Number and 2) Size of hidden encoder layers and 3) Dropout. For more information, refer to Appendix B.

CTM (Bianchi et al., 2021b). We chose this model as it utilises S-BERT (Reimers and Gurevych, 2019) embeddings as an additional source of information to construct a topic model. Additionally, there are other models such as (Dieng et al., 2020) that leverage on word embeddings.

NeuralLDA (Srivastava and Sutton, 2017). Introduced alongside ProdLDA, with its main difference is how its β is interpreted. For its β, the decoder's weights are further processed via batch-normalisation and softmax.

NVDM (Miao et al., 2016). It is widely used as a baseline comparison in topic modelling, and is shown to produce a topic set that has has a weaker coherence compared to other NTMs.

ProdLDA (Srivastava and Sutton, 2017). This NTM is a popular topic modelling baseline and is also used as a backbone model in CTM. Compared to NeuralLDA, ProdLDA's β does not undergo addition processing steps.

WTM (Nan et al., 2019). This model differs greatly from the other selected models as it uses Wasserstein auto-encoders (Tolstikhin et al., 2018) for topic modelling. We use the recommended hyper-parameters Dirichlet parameter of 0.1 and noise coefficient α to 0.5.

7.2 Training Corpora

We use four English language corpora from OCTIS. For more details about the preparation of the corpora, refer to Terragni et al. (2021). Aside from the quantifiable differences (Table 2), we also note that 20NewsGroup[6] and BBC-news (Lim and Buntine, 2014) have vocabularies that are considered more general and broad compared to the specialized and technical vocabularies found in M10 (Greene and Cunningham, 2006) and DBLP (Tang et al., 2008; Pan et al., 2016).

Each corpus has a predefined train/val/test split comprising of 70%/15%/15%. During the training phase, the models optimizes its loss function on the train set in an unsupervised manner. The val set is used to determine early stopping. The full corpus is used for coherence evaluation during the Topic Selection stage.

7.3 NPMI

For our NPMI calculation, we use the recommended window size of 10 to consider word co-occurrences. To score V, with $l = 10$, we evaluate for NPMI on D, using Gensim[7] (Řehůřek and Sojka, 2010) wrapper in OCTIS. These NPMI scores are then utilised to select $\hat{\mathcal{T}}$ in Topic Selection.

For a fairer evaluation against the original $\mathcal{T}$, we conducted coherence evaluation on a external large corpora, using Palmetto[8] (Röder et al., 2015), a coherence evaluation tool with its word co-occurrence index built from Wikipedia articles. We do not measure perplexity, because our reinterpretation process does not change the weights of τ, hence, τ's perplexity remains unchanged. As NPMI of topics within $\hat{\mathcal{T}}$ and $\mathcal{T}$ might not have a normal distribution, a one-sided Mann–Whitney U test (Mann and Whitney, 1947) is suitable (Hart, 2001) to evaluate the significance of the difference in NPMI between $\hat{\mathcal{T}}$ and $\mathcal{T}$.

7.4 Results

Better composite topics can be found. In most experiment instances with results for $K = 20$, shown in Table 3, we are able to discover a set of composite topics $\hat{\mathcal{T}}$ that score better in NPMI and TU on the external reference corpus, suggesting that $\hat{\mathcal{T}}$ is more coherent and has a higher generality compared to $\mathcal{T}$. The observations for $K = 20$ extends to when $K = 50$ (see Appendix C.1).

Information outside of top l words. To get a sense of how composite topics are different from the components, Table 4 shows several examples selected from ProdLDA (MDKP) on 20NewsGroup at $K = 20$. From the first example, "medical" did not appear in the top-10 words of the component surface topics. Combining all three component topics (2, 6, 12) could surface the word in this "healthcare research"-related topic. Furthermore, some words that are highly activated in the component topics, are suppressed in the composite topics. We believe this is caused by negative values in β, that may be informative. Experiments conducted with positively-constrained β yields worse results compared to unconstrained β.

Reducing redundancy. We showcase the third example in Table 4 where two unique but similar-themed component topics combine to form a better composite topic. The two component topics are excluded from the final $\hat{\mathcal{T}}$. By folding together two similar component topics, we could make room to surface other topics of other themes, improving the diversity of $\hat{\mathcal{T}}$ qualitatively.

On model collapse. When $\mathcal{T}$ contains similar topics, the composite combinations of these topics would also produce similar topics in $\hat{\mathcal{T}}$. In Table 3b, while $\mathcal{T}$ of NVDM has similar topics, we still can improve NPMI and TU in $\hat{\mathcal{T}}$, despite many candidate topics sharing similar words, However, if a topic model collapses to a single topic, it is unlikely that we can generate more topics.

Better topic set not guaranteed. This occurs when $\hat{\mathcal{T}}$ does not improve on $\mathcal{T}$ in both metrics, suggesting $\hat{\beta} \approx \beta$, such as in Table 3c, where MDKP for NeuralLDA unable to find a better $\hat{\mathcal{T}}$. Consequently, this means that we are likely to be already evaluating the best topic set that can be interpreted from the topic model.

Impact of ϵ. Adjusting ϵ influences the solution space of $\hat{\mathcal{T}}$, resulting in trade-off between uniqueness and coherence. Table 5 shows that as ϵ increases, NPMI increases while TU decreases. Since different ϵ produces different $\hat{\mathcal{T}}$, we might have multiple solutions where $\hat{\mathcal{T}}$ is better than $\mathcal{T}$.

Impact of s. We tried three different ways of generating candidate pools (see Table 6) and find that in cases where $|V|$ discovered by FIM (referred to as *discovered*) is low, adding composite pairs to V generated by Apriori algorithm is a non-expensive method to increase $|V|$. However, over-generating candidates might result in topics overfitted to the training corpus. Comparing modes

[6] http://people.csail.mit.edu/jrennie/20Newsgroups/

[7] https://radimrehurek.com/gensim/models/coherencemodel.html

[8] https://aksw.org/Projects/Palmetto.html

		NPMI				TU			
	s	Original	MWBIS	MDKP	Greedy	Original	MWBIS	MDKP	Greedy
CTM	0.01	0.0624	**0.1020*****	**0.0858***	**0.104*****	0.965	**1**	**0.975**	**1**
NeuralLDA	0.01	0.0265	**0.0473***	**0.0351**	**0.0344**	0.890	**0.935**	**0.91**	**0.91**
NVDM	0.01	0.0487	**0.0738***	**0.0706**	**0.0710**	0.705	**0.795**	**0.820**	**0.86**
ProdLDA	0.01	0.0433	**0.0842****	**0.0897****	**0.081****	0.900	**0.930**	**0.950**	**0.915**
WTM	0.01	0.0565	**0.1100****	**0.108****	**0.109****	0.945	**1**	**0.955**	**1**

(a) Experiment results for 20NewsGroup with $K = 20$.

		NPMI				TU			
	s	Original	MWBIS	MDKP	Greedy	Original	MWBIS	MDKP	Greedy
CTM	0.01	0.0651	**0.0658**	**0.108***	**0.106***	0.745	**1**	**0.765**	**0.755**
NeuralLDA	0.01	0.0416	**0.0461**	**0.0527**	0.0353	0.960	**1**	**0.960**	**1**
NVDM	0.01	0.0419	NA	**0.0721*****	**0.0532**	0.385	NA	**0.425**	**0.485**
ProdLDA	0.01	0.0546	**0.0744**	**0.0934***	**0.0883***	0.810	**0.82**	**0.825**	**0.82**
WTM	0.19	0.1010	**0.1120**	**0.105**	**0.102**	0.925	**0.935**	**0.960**	**0.96**

(b) Experiment results for BBC-news with $K = 20$.

		NPMI				TU			
	s	Original	MWBIS	MDKP	Greedy	Original	MWBIS	MDKP	Greedy
CTM	0.01	0.0525	0.0450	**0.0541**	0.0435	0.83	**0.840**	**0.840**	**0.9**
NeuralLDA	0.01	0.0169	**0.0200**	NA	**0.0254**	0.96	0.865	NA	0.870
ProdLDA	0.01	0.0331	0.0166	**0.0375**	**0.0495**	0.90	**1**	**0.915**	**0.905**
WTM	0.15	-0.0581	**-0.0384**	**-0.0272***	**-0.0217***	1	1	1	1

(c) Experiment results for DBLP with $K = 20$.

		NPMI				TU			
	s	Original	MWBIS	MDKP	Greedy	Original	MWBIS	MDKP	Greedy
CTM	0.03	0.0580	**0.0732**	**0.0764**	**0.0699**	0.875	0.875	**0.915**	0.875
NeuralLDA	0.01	0.0109	0.00025	0.00285	-0.0227	0.885	0.855	**0.890**	**0.895**
ProdLDA	0.01	0.0173	**0.0469***	**0.0452****	**0.0606*****	0.725	**0.770**	**0.730**	**0.775**
WTM	0.15	0.0183	**0.0582****	**0.0553****	**0.0554****	0.965	**1**	0.965	**0.990**

(d) Experiment results for M10 with $K = 20$.

Table 3: Hyper-parameter s chosen by selecting the V with size closest to 1000. Values in bold indicate better than original baseline result. NA means unable to find a better solution than original baseline result. ***:$p < 0.01$ **:$p < 0.05$, *:$p < 0.1$ per Mann–Whitney U test. NVDM results on M10 and DBLP due to model collapse.

'pairs' (candidate topics must be composite of two components only) and 'add-pairs' (adding pairs to the discovered frequent itemsets), we can conclude that compositions of more than 2 topics can be meaningful. From our experiment results, a recommended target size $|V|$ close to 1000 is reasonable for $K = 20$ and $K = 50$, and can be revised upwards for larger values of K.

7.5 Computational Practicability

In the hundreds of experiments (shown in Figure 4), a few could not be solved within time limit with MIP gap > 0.05. These involve large V exceeding 10,000 candidate topics with ϵ set to enforce tight uniqueness constraint, i.e. $\epsilon = 0$. In Gouveia and Martins (2015), experiments on similar maximum-weight clique problems suggest that solver may be impractical when both density of graph and vertex count is high. However, setting reasonable ϵ and s to avoid such conditions, we find many feasible $\hat{\mathcal{T}}$. In any case, the Greedy approach is always capable of producing a solution.

8 User Study

We have 29 valid responses to our user study, consisting of 30 questions (14 normal and 1 verification each for two tasks below). We excluded responses that failed verification questions[9], ensuring responses of higher quality. Before starting, participants were given a short primer on coherence and reminded that there are no right or wrong answers.

Questions. Procedure of random question generation, with topics sorted alphabetically, and example questions can be found in Appendix A.

[9] A verification question would contain a 'fake' topic, e.g., "animal blood you should select this option for this question".

Set	#	Words (coherence)
$\hat{\mathcal{T}}$	2, 12, 6	**research**, medical, **treatment**, **patient**, disease, **medicine**, study, effect, **health**, fund (0.16)
$\mathcal{T}$	2	**medicine**, literature, bias, article, **research**, blood, associate, **treatment**, poster, treat (0.03)
$\mathcal{T}$	6	firearm, people, gun, **patient**, drug, bill, **health**, amendment, law, weapon (0.04)
$\mathcal{T}$	12	launch, satellite, year, mission, orbit, space, station, rocket, flight, system (0.12)
$\hat{\mathcal{T}}$	7, 9	**game**, **season**, **team**, **player**, win, **score**, **year**, play, **hockey**, **playoff** (0.18)
$\mathcal{T}$	7	**game**, **playoff**, **score**, **hockey**, fan, goal, blue, period, **season**, shot (0.10)
$\mathcal{T}$	9	good, **year**, **player**, make, time, point, **season**, average, league, **team** (0.07)
$\hat{\mathcal{T}}$	13, 15	**drive**, **card**, **disk**, **work**, **scsi**, **problem**, **driver**, hard, **ide**, controller (0.14)
$\mathcal{T}$	13	system, **disk**, **work**, run, backup, drive, memory, software, **driver**, card (0.09)
$\mathcal{T}$	15	**scsi**, **drive**, **card**, **ide**, cable, speed, **problem**, fast, boot, connector (0.08)

Table 4: Examples selected from ProdLDA (MDKP) on 20NewsGroup at $K = 20$ to demonstrate composite properties on surface topics. # - original topic ID, composite topics will have multiple. Words in topics sorted by activation strength. Words in bold denotes common words. Examples separated with double horizontal line. $\hat{\mathcal{T}}$ denotes composite topics and $\mathcal{T}$ denotes respective component topics. For full $\hat{\mathcal{T}}$ and $\mathcal{T}$, see Appendix D.

	NPMI		TU	
ϵ	MWBIS	Greedy	MWBIS	Greedy
0	**0.0663***	**0.0555**	**1**	**1**
1	**0.0842****	**0.0712***	**0.930**	**0.945**
2	**0.0928*****	**0.081****	0.875	**0.915**
3	**0.0946*****	**0.103*****	0.805	0.870

Table 5: Ablation experiment results for ProdLDA on 20NewsGroup with $K = 20$, $s = 0.01$, $|V| = 797$ across different ϵ. Baseline $\mathcal{T}$ has NPMI($\mathcal{T}$) $= 0.0423$ and $TU(\mathcal{T}) = 0.9$.

		NPMI		TU		
Modes	s	MWBIS	Greedy	MWBIS	Greedy	$\|V\|$
add-pairs	0.01	**0.0842****	**0.0712***	**0.930**	**0.945**	797
	0.03	**0.0643**	**0.0697***	**0.905**	**0.945**	277
	0.05	**0.0752****	**0.0751****	**0.920**	**0.930**	211
	0.07	**0.0738****	**0.0698***	**0.920**	**0.935**	198
discovered	0.01	**0.0842****	**0.0712***	**0.930**	**0.945**	797
	0.03	**0.0817****	**0.0638**	**0.920**	**0.950**	230
	0.05	**0.0588**	**0.0617**	**0.920**	**0.920**	103
	0.07	**0.0440**	**0.0436**	0.890	**0.930**	56
pairs	-	**0.0698***	**0.0698***	**0.920**	**0.935**	190

Table 6: Truncated ablation results for ProdLDA on 20NewsGroup with $K = 20$ with s for different modes of generation. Baseline $\mathcal{T}$ has NPMI($\mathcal{T}$) $= 0.0423$ and $TU(\mathcal{T}) = 0.9$. For full results, refer to Appendix C.2.

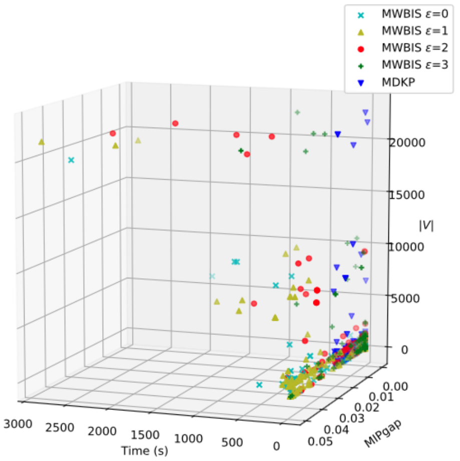


Figure 4: Three-dimensional graph detailing practicability of using solvers. Each experiment is a data point. Experiments were run on Intel Xeon Gold 6132 @ 2.60GHz with 384GB RAM.

For Task I, participants are shown a pair of composite-component topics and asked to identify which of the two is more coherent. We split the 14 questions into two groups where half of the questions contains a component topic with NPMI strictly larger than its paired composite topic, with the other half having equal or less.

For Task II, participants are shown a group of topics consisting of one composite topic and its components and asked to check which topics they think are coherent. They may select multiple options or none at all. Following which, they are asked if the composite topic is related to its components. Out of the 14 questions in Task II, 7 groups of topics will serve as control, with one of its component topic randomly swapped out with another.

Insights. In Task I, we establish that NPMI indeed has a positive correlation (Pearson's $r = 0.500^{***}$)[10] to participants' selection of their preferred topic, with greater participant's agreement in instances where NPMI difference is large. Additionally, despite the 50/50 split, in 60% of question instances, participants choose the composite topic over its component topic.

In Task II, we plot each topic shown as a point

[10] We note that our r is slightly lower than the $r = 0.653$ reported in Röder et al. (2015)

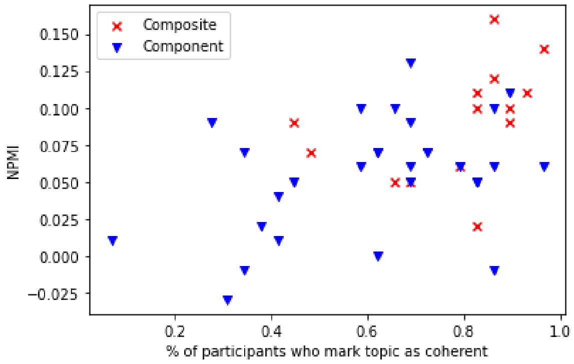


Figure 5: Plots of topics' NPMI against perceived human coherence amongst study participants.

in Figure 5. On average, composite topics have a higher consistent agreement (%), amongst participants marking it as coherent, with a mean of 78%, compared to component topics, at 61%. Additionally, in terms of composite-component relevance, 5 out of 7 treatment groups have more than 75% of participants agreeing that the composite topic is relevant to the component topics, compared to 0 out of 7 control groups for the same criteria. This reveals that while majority of the composite topics are built out of related component topics, there are also instances when non-related component topics contribute to form composite topics.

9 Conclusion

Our proposed two-stage reinterpretation process strongly demonstrates the possibility of obtaining better topic sets. Its accompanying improvements, in both computational metrics and human evaluation, highlight the necessity to view the original topic model in a composite manner to reveal a deeper interpretation. Since auto-encoder frameworks are widely used on other tasks, future investigation is required to explore and determine if this methodology can be applied to other tasks as well.

Limitations

Using composite topics for documents. We conducted a simple supervised classification task using supervised logistic regression to compare composite and original topic vectors. Classification accuracy for both vectors are very similar suggesting parity in information while being different in the interpretation of the information.

Effect of τ's K on $\hat{B}$ and $\hat{\mathcal{T}}$. Given the scope of this paper, we have not explored comparing similar NTMs with different K, i.e., comparing $\hat{\mathcal{T}}$ from a model with $K = 20$ against $\mathcal{T}$ from the same model type with $K = 50$ or higher values of K. Overcoming the current NTM's fixed K might help to generate better models tailored to evaluation for a specific number of topics and more investigation into this area is required.

$|\mathbf{V}|$ generated. For the purposes of parity, we try to keep $|V|$ at similar levels for experiments shown. However, the relationship between NTMs and generated V varies, some models might require a larger $|V|$ to showcase its full potential.

Ethics Statement

We understand that some corpus might produce topics with group of words that might cause offense due to possible sensitiveness regarding politically-charged affairs in the Middle-East. Hence, for our user study, we reviewed questions to remove or replace any topics that we think might be offensive. However, for the sake of transparency, these omitted topics are still included in the full set of topics that are listed in the Appendix D. The use of the reinterpretation process is largely dependent on the corpus that NTM τ is trained on.

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-020). We also extend our gratitude to our user study participants, as well as, our reviewers for their feedback.

References

Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499.

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany.

Per Austrin, Subhash Khot, and Muli Safra. 2009. Inapproximability of vertex cover and independent set in bounded degree graphs. In *2009 24th Annual IEEE Conference on Computational Complexity*.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 759–766, Online.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL Conference 2009*.

Jin-Hee Cho, Yating Wang, Ing-Ray Chen, Kevin S. Chan, and Ananthram Swami. 2017. A survey on modeling and optimizing multi-objective systems. *IEEE Communications Surveys & Tutorials*, 19(3):1867–1901.

P.C. Chu and J.E. Beasley. 1998. Genetic algorithm for the multidimensional knapsack problem. *Journal of Heuristics*, 4(1):63–86.

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

Thanh-Nam Doan and Tuan-Anh Hoang. 2021. Benchmarking neural topic models: An empirical study. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4363–4368, Online.

Michael R. Garey and David S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. USA.

Luis Gouveia and Pedro Martins. 2015. Solving the maximum edge-weight clique problem in sparse graphs with compact formulations. *EURO Journal on Computational Optimization*, 3(1):1–30.

Derek Greene and Padraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *ICML*, volume 148 of *ACM International Conference Proceeding Series*, pages 377–384.

Amulya Gupta and Zhu Zhang. 2021. Vector-quantization-based topic modeling. *ACM Trans. Intell. Syst. Technol.*, 12(3).

Anna Hart. 2001. Mann-whitney test is not just a test of medians: Differences in spread can be important. *BMJ (Clinical research ed.)*, 323:391–3.

Matthew Hoffman, Francis Bach, and David Blei. 2010. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 23.

Tushar Kalra, Rogers Mathew, Sudebkumar Prasant Pal, and Vijay Pandey. 2017. Maximum weighted independent sets with a budget. In *CALDAM*.

Subhash Khot. 2002. On the power of unique 2-prover 1-round games. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing - STOC '02*.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Soukaina Laabadi, Mohamed Naimi, Hassan El Amri, and Boujemâa Achchab. 2018. The 0/1 multidimensional knapsack problem and its variants: A survey of practical models and heuristic approaches. *American Journal of Operations Research*, 08(05):395–439.

Jey Han Lau, Timothy Baldwin, and Trevor Cohn. 2017. Topically driven neural language model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 355–365, Vancouver, Canada.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden.

Kar Wai Lim and Wray L. Buntine. 2014. Bibliographic analysis with the citation network topic model. In *ACML*, volume 39 of *JMLR Workshop and Conference Proceedings*.

Tianyi Lin, Zhiyue Hu, and Xin Guo. 2019. Sparsemax and relaxed wasserstein for topic sparsity. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 141–149, New York, NY, USA.

H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60.

Silvano Martello and Paolo Toth. 1990. *Knapsack Problems: Algorithms and Computer Implementations*. USA.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, Chao Zhang, and Jiawei Han. 2020. Hierarchical topic mining via joint spherical tree and text embedding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1908–1917, New York, NY, USA.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1727–1736, New York, New York, USA.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with Wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6345–6381, Florence, Italy.

Shirui Pan, Jia Wu, Xingquan Zhu, Chengqi Zhang, and Yang Wang. 2016. Tri-party deep network representation. In *IJCAI*, pages 1895–1901.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *WSDM*, pages 399–408.

Ashoka Savasere, Edward Omiecinski, and Shamkant B. Navathe. 1995. An efficient algorithm for mining association rules in large databases. In *Proceedings of the 21th International Conference on Very Large Data Bases*, VLDB '95, page 432–444, San Francisco, CA, USA.

Dazhong Shen, Chuan Qin, Chao Wang, Zheng Dong, Hengshu Zhu, and Hui Xiong. 2021. Topic modeling revisited: A document graph-based neural network perspective. In *Advances in Neural Information Processing Systems 34 - 35th Conference on Neural Information Processing Systems, NeurIPS 2021*, Advances in Neural Information Processing Systems, pages 14681–14693. Neural information processing systems foundation.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *ICLR (Poster)*.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *KDD*, pages 990–998.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. OCTIS: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270, Online.

Hannu Toivonen. 1996. Sampling large databases for association rules. In *Proceedings of the 22th International Conference on Very Large Data Bases*, VLDB '96, page 134–145, San Francisco, CA, USA.

Ilya O. Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. 2018. Wasserstein auto-encoders. In *ICLR*.

Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Topic-guided variational auto-encoder for text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 166–177, Minneapolis, Minnesota.

Zhengjue Wang, Zhibin Duan, Hao Zhang, Chaojie Wang, Long Tian, Bo Chen, and Mingyuan Zhou. 2020. Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 485–497, Online.

Liang Yang, Fan Wu, Junhua Gu, Chuan Wang, Xiaochun Cao, Di Jin, and Yuanfang Guo. 2020. Graph attention topic modeling network. In *Proceedings of The Web Conference 2020*, WWW '20, page 144–154, New York, NY, USA. Association for Computing Machinery.

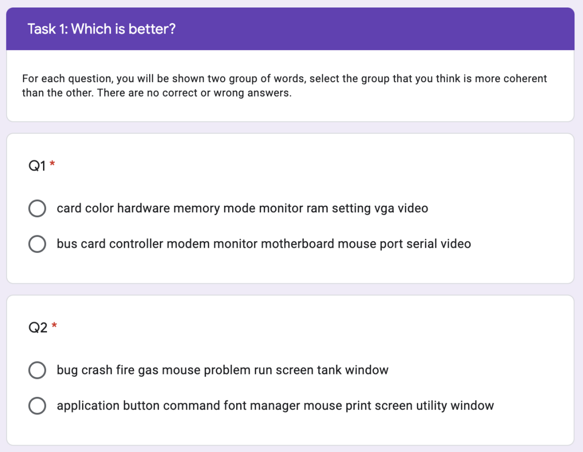
Ce Zhang and Hady W. Lauw. 2020. Topic modeling on document networks with adjacent-encoder. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6737–6745.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021. Topic modelling meets deep neural networks: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4713–4720. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Renbo Zhao, Vincent Tan, and Huan Xu. 2017. Online Nonnegative Matrix Factorization with General Divergences. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 37–45.

A User Study Appendix

Task I. To select pairs, we shuffled $\hat{\mathcal{T}}$ and select the first 7 pairs of random topics made up of one composite topic and one of its component topic where the NPMI of the composite topic is more than its random component topic. We repeat the procedure



Task 1: Which is better?

For each question, you will be shown two group of words, select the group that you think is more coherent than the other. There are no correct or wrong answers.

Q1 *

○ card color hardware memory mode monitor ram setting vga video

○ bus card controller modem monitor motherboard mouse port serial video

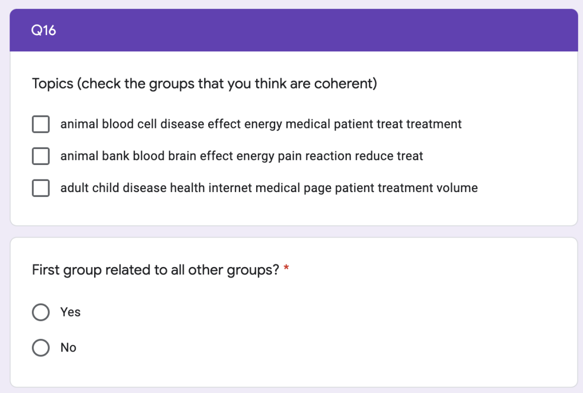
Q2 *

○ bug crash fire gas mouse problem run screen tank window

○ application button command font manager mouse print screen utility window

Figure 6: Example of a question in user study Task I.

to obtain another 7 pairs with NPMI of composite topic is lower than its random component topic. $\hat{\mathcal{T}}$ is from CTM on 20NewsGroup at $K = 50$. The options for each questions are randomized when displayed to the volunteer.



Q16

Topics (check the groups that you think are coherent)

□ animal blood cell disease effect energy medical patient treat treatment

□ animal bank blood brain effect energy pain reaction reduce treat

□ adult child disease health internet medical page patient treatment volume

First group related to all other groups? *

○ Yes

○ No

Figure 7: Example of a question in user study Task II.

Task II. We randomly select 14 groups of topics from $\hat{\mathcal{T}}$ made up of a single component and its component topics. $\hat{\mathcal{T}}$ is from ProdLDA on 20NewsGroup at $K = 50$. Of the 14 groups, we randomly choose 7 groups and replace one of its component with a random topic to create a control sample to test for composite-component similarity relations. The component topic is shown at the top of the list, followed by the component topics.

Participant recruitment. We recruited study participants from two groups of people. For the first group, we have 17 valid responses from graduates with a STEM background, physically located locally in our city. For the second group, we have 12 valid responses from a small online text-based role-playing game community, physically located around the world. On average, the responses from both group are similar.

Model	K	# neurons	# hidden layers	dropout
CTM	20	200	1	0.2
CTM	50	200	1	0.2
NeuralLDA	20	200	1	0.2
NeuralLDA	50	300	1	0.2
NVDM	20	512	2	0.0
ProdLDA	20	200	2	0.2
ProdLDA	50	200	3	0.2
WTM	20	100	2	default
WTM	50	100	2	default

(a) Parameters for NTMs for 20NewsGroup

Model	K	# neurons	# hidden layers	dropout
CTM	20	200	1	0.2
CTM	50	200	1	0.2
NeuralLDA	20	300	1	0.2
NeuralLDA	50	200	1	0.2
NVDM	20	512	2	0.0
ProdLDA	20	200	1	0.2
ProdLDA	50	200	2	0.2
WTM	20	100	2	default
WTM	50	200	1	default

(b) Parameters for NTMs for BBC-news

Model	K	# neurons	# hidden layers	dropout
CTM	20	300	2	0.2
NeuralLDA	20	200	2	0.2
ProdLDA	20	100	1	0.2
WTM	20	200	1	default

(c) Parameters for NTMs for M10

Model	K	# neurons	# hidden layers	dropout
CTM	20	300	2	0.2
NeuralLDA	20	300	1	0.2
ProdLDA	20	200	1	0.2
WTM	20	200	1	default

(d) Parameters for NTMs for DBLP

Table 7: Parameters for NTMs for 20NewsGroup

B Model Parameters and Optimization

For all NTMs, except WTM, we use OCTIS[11] bayesian optimizer to search for encoder parameters with 30 optimization iterations and 3 model runs each with selected parameters in Table 7. For all NTMs, their decoder has no hidden layers. We adapted NVDM[12] for OCTIS framework. For WTM [13], we use similar recommended parameters suggested in (Nan et al., 2019). We use default values for unmentioned parameters.

[11] https://github.com/MIND-Lab/OCTIS

[12] referred to both https://github.com/YongfeiYan/Neural-Document-Modeling and https://github.com/ysmiao/nvdm

[13] https://github.com/awslabs/w-lda

C Additional Results Appendix

C.1 Experiment results for NTMs with $K = 50$

The tabled results for 20NewsGroup and BBC-news for NTMs with $K = 50$.

		NPMI				TU			
	s	Original	MWBIS	MDKP	Greedy	Original	MWBIS	MDKP	Greedy
CTM	0.1	0.0559	**0.0695***	**0.0948*****	**0.0836*****	0.818	**0.86**	**0.826**	**0.824**
NeuralLDA	0.01	0.0466	**0.0667****	**0.0742*****	**0.0667****	0.748	**0.786**	**0.8**	**0.782**
ProdLDA	0.1	0.0416	**0.0747*****	**0.09*****	**0.0844*****	0.748	**0.778**	**0.752**	**0.768**
WTM	0.03	0.0595	**0.0824***	**0.0977*****	**0.0939*****	0.812	**0.842**	**0.814**	**0.812**

(a) Experiment results for 20NewsGroup with $K = 50$.

		NPMI				TU			
	s	Original	MWBIS	MDKP	Greedy	Original	MWBIS	MDKP	Greedy
CTM	0.01	0.053	**0.0827****	**0.0906*****	**0.0871****	0.732	**0.742**	**0.75**	**0.738**
NeuralLDA	0.1	0.0424	**0.0427**	**0.0447**	**0.0563****	0.81	**0.876**	**0.892**	**0.85**
ProdLDA	0.07	0.0469	**0.0737***	**0.103*****	**0.0699**	0.584	**0.596**	**0.598**	**0.74**
WTM	0.2	0.0727	**0.0757**	**0.102*****	**0.088***	0.738	**0.806**	**0.758**	**0.758**

(b) Experiment results for BBC-news with $K = 50$.

Table 8: Hyper-parameter s chosen by selecting the candidate pool that has a size closest to 1000. Values in bold indicate better than original baseline result. ***:$p < 0.01$ **:$p < 0.05$, *:$p < 0.1$

C.2 Full results for ablation on s

The extended tabled results for three different modes of generations for different s.

		NPMI			TU			
Modes	s	MWBIS	MDKP	Greedy	MWBIS	MDKP	Greedy	$\lvert V\rvert$
add-pairs	0.01	**0.0842****	**0.0897****	**0.0712***	**0.930**	**0.950**	**0.945**	797
	0.03	**0.0643**	**0.0569**	**0.0697***	**0.905**	**0.935**	**0.945**	277
	0.05	**0.0752****	**0.0644**	**0.0751****	**0.920**	**0.965**	**0.930**	211
	0.07	**0.0738****	**0.0522**	**0.0698***	**0.920**	**0.955**	**0.935**	198
	0.10	**0.0785****	**0.0526**	**0.0698***	**0.920**	**0.955**	**0.935**	193
discovered	0.01	**0.0842****	**0.0897****	**0.0712***	**0.930**	**0.950**	**0.945**	797
	0.03	**0.0817****	**0.0468**	**0.0638**	**0.920**	**0.960**	**0.950**	230
	0.05	**0.0588**	NA	**0.0617**	**0.920**	NA	**0.920**	103
	0.07	**0.0440**	NA	**0.0436**	0.890	NA	**0.930**	56
	0.10	**0.0424**	NA	**0.0441**	**0.900**	NA	**0.920**	22
pairs	-	**0.0698***	**0.0561**	**0.0698***	**0.920**	**0.955**	**0.935**	190

Table 9: Ablation experiment results for ProdLDA on 20NewsGroup with $K = 20$ across candidate pools of size $|V|$ generated from different values of s and different modes of generation. Baseline $\mathcal{T}$ has NPMI($\mathcal{T}$) $= 0.0423$ and $TU(\mathcal{T}) = 0.9$, and is used to compare for significance. For MWBIS and Greedy, we select a s that produces similar TU for easier comparison.

D Full Topic Set Examples

NPMI shown are evaluated on large external corpora in Palmetto. Each composite topic is shown in terms of a listing of the component topics, e.g., composite topic (1, 17) indicates that it has been derived from combining component topic 1 and topic 17. For each topic, we show the NPMI score, as well as a list of the top-10 words.

Topic sets $\hat{\mathcal{T}}$ (MDKP) and $\mathcal{T}$ from ProdLDA on 20NewsGroup at $K = 20$

#	NPMI	Topics
New composite topics in $\hat{\mathcal{T}}$		
1, 17	0.03	people, road, town, kill, armenian, dead, soldier, woman, body, leave
3, 5	0.01	fire, compound, die, death, building, child, place, evil, tear, life
5, 8	0.07	agent, warrant, criminal, illegal, police, batf, federal, citizen, law, crime
7, 9	0.18	game, season, team, player, win, score, year, play, hockey, playoff
13, 15	0.14	drive, card, disk, work, scsi, problem, driver, hard, ide, controller
14, 19	0.07	file, window, image, color, format, display, convert, widget, program, set
2, 6, 12	0.16	research, medical, treatment, patient, disease, medicine, study, effect, health, fund
2, 11, 12	0.10	science, scientist, observation, objective, scientific, natural, theory, term, human, concept
2, 12, 16	0.08	sell, sale, price, pay, interested, cost, purchase, item, money, offer
10, 15, 16	0.04	speaker, external, connector, circuit, mhz, internal, apple, motherboard, parallel, cable
14, 16, 16	0.04	monitor, card, video, mouse, memory, meg, printer, ram, vga, resolution
15, 17, 18	0.07	engine, oil, brake, replace, car, battery, tire, plastic, shop, dealer
1, 2, 5, 11	0.06	moral, society, justify, matter, sexual, sex, defend, practice, prove, freedom
4, 6, 8, 19	0.16	internet, mail, network, address, email, privacy, access, message, newsgroup, information
4, 13, 14, 19	0.05	advance, code, compile, graphic, host, shareware, window, utility, library, application
5, 12, 17, 18	0.12	vehicle, gas, heavy, engine, tank, ride, foot, fuel, pound, weight
Common component topics in $\hat{\mathcal{T}}$ and $\mathcal{T}$		
7	0.10	game, playoff, score, hockey, fan, goal, blue, period, season, shot
8	0.06	key, clipper, chip, secure, encrypt, encryption, escrow, security, algorithm, enforcement
11	0.12	homosexual, belief, religion, truth, interpretation, nature, meaning, homosexuality, christian, human
12	0.12	launch, satellite, year, mission, orbit, space, station, rocket, flight, system
Excluded component topics from $\hat{\mathcal{T}}$ but in $\mathcal{T}$		
0	-0.03	powerful, frequently, consist, limited, earlier, deep, longer, numerous, compare, portion
1	0.10	muslim, people, israeli, genocide, village, population, turkish, jewish, government, armenian
2	0.03	medicine, literature, bias, article, research, blood, associate, treatment, poster, treat
3	0.01	people, make, time, thing, president, work, church, morning, pray, give
4	-0.02	advance, summary, reply, host, address, interested, domain, compile, email, print
5	0	batf, fire, compound, assault, knock, gas, warrant, crime, agent, criminal
6	0.04	firearm, people, gun, patient, drug, bill, health, amendment, law, weapon
9	0.07	good, year, player, make, time, point, season, average, league, team
10	-0.07	gather, pre, fair, remark, portion, critical, previously, chapter, frequently, limited
13	0.09	system, disk, work, run, backup, drive, memory, software, driver, card
14	0.05	window, screen, font, color, default, mouse, convert, event, display, problem
15	0.08	scsi, drive, card, ide, cable, speed, problem, fast, boot, connector
16	-0.04	sell, sale, price, offer, monitor, interested, shipping, video, card, condition
17	0.11	car, bike, engine, ride, tire, road, brake, start, floor, gear
18	-0.01	requirement, warning, consist, limited, submit, frequently, complaint, chain, oil, recommend
19	0.05	mail, internet, pub, site, graphic, email, file, send, format, list

Topic sets $\hat{\mathcal{T}}$ (MDKP) and $\mathcal{T}$ from ProdLDA on 20NewsGroup at $K = 50$

#	NPMI	Topics
New composite topics in $\hat{\mathcal{T}}$		
0, 32	0.09	address, mail, email, mailing, paper, network, list, topic, internet, advance
0, 35	0.20	space, mission, orbit, shuttle, system, launch, satellite, solar, flight, rocket
1, 3	0.09	church, christian, passage, verse, scripture, word, father, teach, refer, doctrine
1, 11	0.08	love, sin, faith, life, good, make, eternal, doctrine, hate, give
4, 25	0.10	window, font, screen, manager, expose, button, display, default, event, app
6, 36	0.04	drive, problem, speed, buy, hard, cable, fast, scsi, power, ide
7, 12	0.10	people, armenian, turkish, massacre, genocide, village, muslim, population, organize, russian
7, 38	0.07	fire, shoot, officer, batf, bullet, incident, knock, gun, wound, tank
9, 12	0.08	israeli, people, arab, jewish, state, territory, occupy, land, civil, country
10, 34	0.09	game, blue, goal, score, play, penalty, back, shot, lead, circle
13, 22	0.14	effect, treat, blood, patient, medical, energy, cell, disease, animal, treatment
14, 46	0.10	card, monitor, port, video, board, slot, motherboard, pin, external, vga
15, 22	0.01	page, guide, email, mail, interested, software, computer, daily, volume, fax
15, 35	0	bag, annual, art, book, copy, element, cover, object, title, flight
17, 24	0.05	law, public, number, key, enforcement, agency, court, amendment, encrypt, license
18, 23	0.20	team, game, season, play, player, baseball, league, playoff, fan, win
19, 31	0.12	absolute, truth, atheism, belief, atheist, moral, definition, objective, statement, morality
22, 43	0.10	medical, patient, food, doctor, treatment, health, eat, year, high, disease
23, 34	0.11	game, team, playoff, play, cap, pen, score, goal, lose, wing
24, 44	0.03	key, secret, chip, algorithm, escrow, clipper, agency, enforcement, encryption, encrypt
25, 40	0.07	window, run, file, directory, problem, manager, menu, program, application, display
25, 46	0.04	mouse, driver, card, mode, problem, video, memory, fine, window, instal
27, 38	0.09	gun, crime, criminal, illegal, violent, drug, insurance, abuse, warrant, police
28, 47	0.10	noise, battery, cycle, frequency, circuit, voltage, heat, low, band, audio
33, 42	0.02	widget, export, motif, window, resource, set, subject, include, server, client
43, 47	0.10	water, oil, temperature, weight, air, battery, heat, fuel, pressure, bike
1, 19, 31	0.14	truth, belief, absolute, christian, christianity, true, religion, human, moral, nature
2, 8, 38	-0.03	fire, batf, compound, watch, tear, gas, building, hear, death, tank
2, 11, 34	0.14	team, game, baseball, fan, play, pitch, hit, ball, bad, player
2, 23, 34	0.05	game, fan, team, play, baseball, watch, playoff, hockey, ranger, pen
3, 9, 19	0.16	homosexual, sex, homosexuality, gay, sexual, male, relationship, behavior, christian, society
4, 32, 33	0.05	advance, print, code, printer, font, convert, draw, tool, character, library
6, 16, 26	0.10	lock, engine, bike, seat, front, owner, rear, chain, wheel, paint
7, 8, 9	0.13	people, kill, civilian, child, murder, woman, innocent, rape, man, israeli
9, 17, 38	0.03	gun, batf, weapon, crime, law, assault, firearm, state, citizen, armed
14, 16, 41	0.06	sale, sell, offer, condition, cheap, price, ship, company, shipping, brand
14, 29, 36	0.02	card, disk, board, ram, video, port, modem, drive, meg, bus
15, 33, 40	0.06	graphic, processing, package, mail, database, software, object, pub, send, analysis
28, 36, 46	0.11	drive, pin, cable, card, internal, connector, connect, board, port, controller
Common component topics in $\hat{\mathcal{T}}$ and $\mathcal{T}$		
6	0.16	bike, car, drive, ride, tire, transmission, gear, engine, brake, shift
7	0.05	people, body, massacre, village, dead, town, bullet, escape, soldier, troop
8	0.02	people, time, neighbor, thing, afraid, building, mother, floor, parent, hospital
12	0.09	greek, turkish, muslim, genocide, century, jewish, armenian, international, territory, soviet
24	0.07	key, block, encrypt, secret, serial, bit, chip, session, generate, algorithm
28	0.13	audio, power, voltage, circuit, input, supply, wire, price, speaker, output
30	0.02	people, make, work, president, decision, morning, job, yesterday, talk, meeting
31	0.06	science, existence, objective, scientist, atheism, evidence, observation, exist, atheist, universe
37	0.04	vote, newsgroup, article, post, group, discussion, topic, propose, creation, response
44	0.01	government, ensure, technology, privacy, encryption, administration, industry, policy, escrow, conversation
48	0.03	quality, image, color, compression, scale, conversion, format, convert, file, shareware

Excluded component topics from $\hat{\mathcal{T}}$ but in $\mathcal{T}$		
0	0.06	post, shuttle, space, mail, posting, usenet, list, email, internet, mailing
1	0.10	church, teach, sin, love, doctrine, soul, faith, life, christian, passage
2	-0.04	fan, lot, watch, doesn, food, guess, dream, baseball, hockey, ball
3	0.05	male, homosexuality, homosexual, cite, refer, historical, law, writer, term, tradition
4	0.04	font, character, print, convert, button, advance, window, expose, printer, attribute
5	0	extend, deep, impossible, originally, permission, spread, consist, huge, tip, frequently
9	0.05	israeli, people, gay, sex, arab, law, homosexual, civilian, society, sexual
10	0.01	time, back, car, people, walk, start, blue, work, year, make
11	0	good, win, love, make, life, faith, pitcher, year, sin, team
13	0.07	energy, effect, blood, bank, reduce, pain, treat, animal, brain, reaction
14	0.01	sale, offer, card, monitor, price, video, sell, interested, board, item
15	-0.01	copy, art, graphic, bag, daily, book, sale, annual, interested, price
16	0.07	sell, sale, company, market, engine, cost, condition, launch, satellite, firm
17	0.06	firearm, license, weapon, bill, file, gun, dangerous, section, amendment, assault
18	0.10	year, good, season, team, average, player, league, draft, game, excellent
19	0.05	truth, absolute, gay, relationship, moral, sex, belief, atheism, christian, agree
20	0.01	make, people, president, time, work, military, yesterday, government, meeting, russian
21	-0.01	domain, portion, pattern, guarantee, summary, greatly, frequently, host, permission, numerous
22	0.07	patient, page, medical, health, treatment, disease, child, volume, adult, internet
23	0.11	pen, team, fan, lose, cap, baseball, playoff, game, win, play
25	0.02	window, run, problem, win, menu, main, manager, file, directory, app
26	0.02	chain, lock, clean, cut, portion, originally, travel, stay, seat, tip
27	0.02	insurance, drug, private, people, canadian, make, cost, doctor, spend, government
29	-0.01	driver, card, run, problem, instal, mouse, screen, ram, video, memory
32	0	advance, address, paper, interested, domain, email, summary, mail, fax, reply
33	0.05	tool, platform, motif, analysis, processing, widget, graphic, export, data, filter
34	0.09	game, goal, lead, score, blue, wing, tie, period, team, play
35	0.03	planet, solar, surface, earth, orbit, moon, degree, sun, mission, dark
36	0.10	drive, scsi, ide, modem, problem, transfer, system, disk, internal, apple
38	-0.03	fire, batf, gun, compound, gas, cop, auto, knock, initial, agent
39	0	suit, portion, virtually, ball, frequently, apparently, joke, supposedly, numerous, suffer
40	0.06	file, database, system, package, workstation, run, graphic, mail, function, utility
41	-0.01	portion, originally, task, virtually, external, upgrade, frequently, sale, numerous, guarantee
42	0.01	resource, variable, client, window, widget, root, make, entry, include, set
43	0.04	pressure, water, car, air, food, engine, eat, day, temperature, good
45	-0.01	portion, popular, frequently, originally, complaint, collection, permission, virtually, successful, tip
46	0.06	card, port, drive, mouse, monitor, controller, board, video, driver, pin
47	0.07	water, heat, cycle, noise, oil, weight, ride, bike, temperature, effect
49	-0.01	numerous, essentially, tip, impossible, worry, complaint, virtually, portion, frequently, suit