# Scale-Invariant Infinite Hierarchical Topic Model

**Shusei Eshima**
Department of Government, Harvard University
1737 Cambridge Street
Cambridge, MA 02138, USA
shuseieshima@g.harvard.edu

**Daichi Mochihashi**
The Institute of Statistical Mathematics
10-3 Midori-cho, Tachikawa City
Tokyo, Japan
daichi@ism.ac.jp

## Abstract

Hierarchical topic models have been employed to organize a large number of diverse topics from corpora into a latent tree structure. However, existing models yield fragmented topics with overlapping themes whose expected probability becomes exponentially smaller along the depth of the tree. To solve this intrinsic problem, we propose a scale-invariant infinite hierarchical topic model (ihLDA). The ihLDA adaptively adjusts the topic creation to make the expected topic probability decay considerably slower than that in existing models. Thus, it facilitates the estimation of deeper topic structures encompassing diverse topics in a corpus. Furthermore, the ihLDA extends a widely used tree-structured prior (Adams et al., 2010) in a hierarchical Bayesian way, which enables drawing an infinite topic tree from the base tree while efficiently sampling the topic assignments for the words. Experiments demonstrate that the ihLDA has better topic uniqueness and hierarchical diversity than existing approaches, including state-of-the-art neural models.

## 1 Introduction

Topic models (Blei et al., 2003b; Blei and Lafferty, 2006; Chang and Blei, 2010; Roberts et al., 2016) have been used to summarize, annotate, and categorize documents. Recent advances in large-scale topic models have enabled the estimation of thousands of topics to accommodate various concepts in a large corpus (Li et al., 2014; Yu et al., 2015; Yuan et al., 2015; Chen et al., 2016), requiring users to interpret numerous topics.

Hierarchical topic models have been proposed to improve the topic organization by learning the latent topic hierarchy (Blei et al., 2003a, 2010; Adams et al., 2010; Kim et al., 2012; Paisley et al., 2015; Isonuma et al., 2020; Chen et al., 2021). However, these hierarchical topic models will create a fragmented tree structure with the probabilities of a substantial number of topics becoming exponentially smaller. These topics typically have few assigned words and similar word distributions. Recent hierarchical topic models with neural architectures (Isonuma et al., 2020; Chen et al., 2021) have the same issue of topic fragmentation and use a fixed number of layers for all documents (Duan et al., 2021).

The reason for the topic fragmentation is that the stick-breaking process (Sethuraman, 1994) used in existing models creates topics whose expected probability decays along the depth of the tree. Existing models alleviate the issue by restricting the tree structure, for example by truncating the depth to three levels. Isonuma et al. (2020) also introduced a topic-diversity regularizer and a heuristic rule to update topics, whereas Chen et al. (2021) truncated topics based on their corpus coverage.

To address this intrinsic issue of topic probabilities, we propose a scale-invariant hierarchical infinite topic model (ihLDA) and make three main contributions. First, the ihLDA adjusts the probability scale of the stick-breaking process at each level by considering the size of the parent topic to avoid fragmented topic structures. The expected topic probability of the ihLDA decays considerably slower than that of the existing models, thereby reflecting the diversity of topics in a corpus by using a flexible depth and width. The existing probabilistic and neural models that leverage the stick-breaking process can also benefit from our model.

Table 1 compares the top words of several topics from different topic models: the ihLDA estimates topics with general words in the shallower levels (L1 and L2), and topics with more specific words at a deeper level (L3). In contrast, nCRP and TSNTM will create topics with overlapping themes. The columns of nCRP and TSNTM show that most topics share the top words, even at the third level because of the issue described above.

Second, the ihLDA extends the tree-structured stick-breaking process (TSSB; Adams et al., 2010),

*Findings of the Association for Computational Linguistics: ACL 2023*, pages 11731–11746
July 9-14, 2023 ©2023 Association for Computational Linguistics

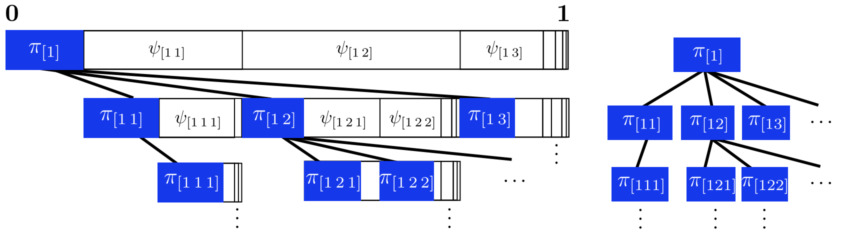


Figure 1: Overview of the tree-structured stick-breaking process (TSSB) in Adams et al. (2010). The blue intervals represent topics with probabilities proportional to their widths. A path of a topic is denoted in square brackets, and the solid lines show connections between topics. The right-hand tree recasts the topic structure. $\psi$'s are the horizontal probabilities of breaking the stick, and each $\pi$ denotes the probability of using a topic.

| **Proposed: ihLDA** | Probabilistic: nCRP (Blei et al., 2003a) | Neural: TSNTM (Isonuma et al., 2020) |
|---|---|---|
| L1: said year would also people | L1: said year one time would | L1: said show year also would |
| L2: said people mobile technology phone | L2: said year also would company | L2: said year game world time |
| L3: said software site user mail | L3: film show magic would child | L3: england first game ireland win |
| L2: said would government people law | L3: film indian star india actor | L3: said labour blair party election |
| L3: tax said government would budget | L3: film dvd effect extra man | L3: said would people law government |
| L3: labour election said party blair | L3: film harry potter dvd warner | L3: said would government election tax |
| L2: film said best award year | L2: best award film actor actress | L3: said would tax government election |
| L3: music band song year album | L2: film star story life singer | L3: said would tax government election |
| L3: game dvd film year sony | L2: film star movie actress also | L3: said would tax government election |

Table 1: Top words from the selected topics (BBC corpus). The ihLDA shows a clear topic hierarchy where children of the parent topics constitute the subtopics. nCRP and TSNTM can create topics with overlapping top words. The maximum number of levels is fixed at three (L3) for comparison.

a prior for a latent hierarchy that is also employed in recent neural models and various applications (Deshwar et al., 2015; Chien, 2016; Nassar et al., 2019). The ihLDA enables drawing an infinite topic tree for each document from a base infinite tree in a hierarchical Bayesian fashion.

Finally, we implement an efficient algorithm that can draw the topics and hierarchical structures from the tree-structured prior without enumerating all possible candidates.

We empirically show that the ihLDA performs better in topic quality using two measures and crowdsourced evaluation. Moreover, the number of estimated topics by the ihLDA is comparable to that by existing models, even when a tree is deeper than three levels.

## 2 Background: Tree-Structured Stick-Breaking Process

A tree-structured stick-breaking process (TSSB) (Adams et al., 2010) is a prior for constructing a topic tree of theoretically unbounded depth and width, comprising two types of stick-breaking processes (Sethuraman, 1994). Figure 1 illustrates a draw from the TSSB, where each blue interval represents a topic, whereas the square brackets denote the path to reach it.[1] Hierarchical topic models assign a latent topic to each word in a document.

As an equivalent representation of the Dirichlet process (see Appendix A for details), a stick-breaking process repeatedly breaks a stick of length 1, where each broken stick corresponds to a topic with the length equal to its probability. Appendix B provides a formal definition and illustration of the stick-breaking process.

Here, we introduce notation to formalize the TSSB. Topic $\boldsymbol{\epsilon}$ at the level $|\boldsymbol{\epsilon}|$ in a tree has its ancestors and children. Let $\boldsymbol{\kappa} \prec \boldsymbol{\epsilon}$ indicate that $\boldsymbol{\kappa}$ is an ancestor of $\boldsymbol{\epsilon}$: in Figure 1, topic $[1\;1\;1]$ has two ancestors, $\{\boldsymbol{\kappa} : \boldsymbol{\kappa} \prec [1\;1\;1]\} = \{[1], [1\;1]\}$. Specifically, we use a prime symbol $\boldsymbol{\prime}$ to denote the parent topic, i.e., $[1\;1\;1]' = [1\;1]$. The child topics of $\boldsymbol{\epsilon}$ are $\{\boldsymbol{\epsilon}k : k \in 1, 2, 3, \ldots\}$. For example, topic $[1\;2]$ in Figure 1 has children $[1\;2\;1], [1\;2\;2], \cdots$.

Given this setup, the probability assigned to a topic $\boldsymbol{\epsilon}$ under TSSB can be expressed as a product of stick-breaking processes:

$$\pi_{\boldsymbol{\epsilon}} = \nu_{\boldsymbol{\epsilon}} \prod_{\boldsymbol{\kappa} \prec \boldsymbol{\epsilon}} (1-\nu_{\boldsymbol{\kappa}}) \cdot \prod_{\boldsymbol{\kappa} \preceq \boldsymbol{\epsilon}} \phi_{\boldsymbol{\kappa}}\,, \tag{1}$$

where $\phi_{\boldsymbol{\epsilon}k} = \psi_{\boldsymbol{\epsilon}k} \prod_{j=1}^{k-1}(1-\psi_{\boldsymbol{\epsilon}j})$. The first term in Equation (1) is the probability of stopping at the topic $\boldsymbol{\epsilon}$ vertically. The next product terms refer to passing ancestors of $\boldsymbol{\epsilon}$ while horizontally stopping at $\boldsymbol{\epsilon}$ and its ancestors. These vertical and horizontal probabilities of stopping follow Beta distributions:

[1] This path notation is adapted from Isonuma et al. (2020) and is different from that in Adams et al. (2010).

$$\nu_\epsilon \sim \mathrm{Be}(1, \alpha_0)\,,\ \psi_\epsilon \sim \mathrm{Be}(1, \gamma_0)\,. \tag{2}$$

Appendix C presents an example of this process.

We also introduce a scaling factor $\lambda$ used in Adams et al. (2010) and set $\alpha_0$ at each level, $\alpha_\epsilon = \alpha_{|\epsilon|} \cdot \lambda^{|\epsilon|-1}$, $0 \leq \lambda \leq 1$, instead of $\alpha_0$ in Equation (2). This parametrization makes a word more likely to stop as $|\epsilon|$ becomes larger, i.e., deeper in the tree. Hereafter, we do not use subscript $\epsilon$ and denote $\alpha_\epsilon$ as $\alpha$ for simplicity.

## 3 Scale-Invariant TSSB

Although the TSSB constitutes a crucial building block of recent hierarchical topic models (Isonuma et al., 2020; Chen et al., 2021), the expected probability of each topic in the TSSB decays exponentially along the depth of the topic hierarchy. Figure 2(a) shows two topic trees drawn from the original TSSB: topics in the third and the fourth levels have extremely small probabilities compared to the topics in the higher levels, resulting in a topic fragmentation in the tree.

This property of the TSSB is attributed to the probability of a horizontal stop, $\psi_\epsilon$, having the same expectation regardless of the level. As shown in Appendix D, the expected probability of a horizontal break at the level $\ell$ is $\mathbb{E}[\phi|\ell] \approx 1/(2\gamma+1)^\ell$, where the level appears in the exponent of the denominator. The dotted line in Figure 3 depicts this exponential decay with $\ell$.

To avoid this exponential decay, we rescale $\gamma_0$ in Equation (2):

$$\psi_\epsilon \sim \mathrm{Be}(1, \phi_{\epsilon'}\gamma_0), \tag{3}$$

where we set $\phi_{\epsilon'} = 1$ when $\epsilon$ is the root topic. Hereafter, we denote $\gamma = \gamma_\epsilon = \phi_{\epsilon'}\gamma_0$ for simplicity.

The key idea in Equation (3) is to use the horizontal breaking proportion of a parent topic, $\phi_{\epsilon'}$, to draw a *relative* stick length for its child topic, $\psi_\epsilon$, creating a larger break if the stick to break is already short. As presented in Appendix D, this new parametrization yields the average stick length $\mathbb{E}[\phi|\ell] \approx 1/(2\gamma + 1/\mathbb{E}[\phi|\ell-1])$ for $\ell \geq 2$, which achieves an invariant partitioning scale by not decaying exponentially with $\ell$. The solid lines in Figure 3 depict the effect of this new parametrization.

Figure 2(b) shows our scale-invariant TSSB with the same hyperparameters as in (a). The probability of the topics is less likely to decrease at the deeper levels in (b).

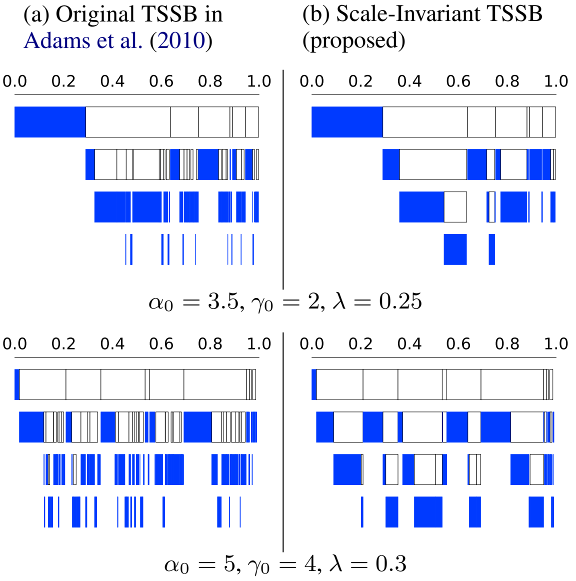


Figure 2: Original TSSB (left) and scale-invariant TSSB (right). Each interval is a latent topic with probability proportional to the width. Each row has the same hyperparameter values except for the proposed adjustment. The figure does not show the solid lines that indicate topic connections in Figure 1. The proposed method does not create topics with small probabilities.

## 4 Scale-Invariant Infinite Hierarchical Topic Model

We employ our scale-invariant TSSB to model the hierarchical latent topics in a document. Topic models consist of two types of distributions: document-topic distributions for topic composition and topic-word distributions for word emission. The ihLDA leverages the scale-invariant TSSB in Section 3 to construct the former and employs a hierarchical Pitman-Yor process (Teh, 2006) for the latter. Combining both distributions will embed the topics into an infinite tree, which we call the ihLDA, scale-invariant infinite hierarchical LDA.

### 4.1 Document-Topic Distribution

The topic composition of each document differs for each document, but the topics must be shared across all the documents. In this regard, we generalize the scale-invariant TSSB to a hierarchical tree-structured stick-breaking process (HTSSB). The HTSSB generates document-specific topic probabilities while making these topics shared by all documents.

Specifically, we hierarchically generate a child TSSB for a document from the base TSSB, as shown in Figure 4. It applies the hierarchical Dirichlet process (HDP; Teh et al., 2006) sepa-

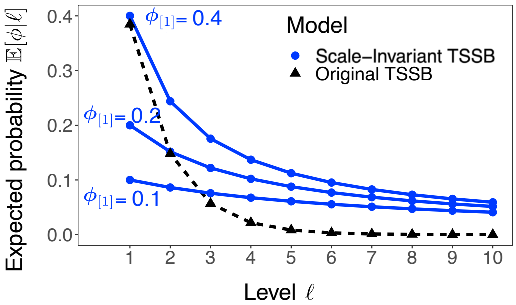


Figure 3: Expected probability of a horizontal stop at each level. Three different root topic probabilities are considered for the scale-invariant TSSB. The scale-invariant TSSB displays a slower decay than the original TSSB. The value of $\gamma_0$ is fixed at 0.8.

rately to the vertical and horizontal probabilities that constitute the TSSB in Equations (2) and (3). In this regard, the HTSSB is an infinite product of the HDPs in terms of its component probabilities. Appendix E provides a formal explanation of the HDP in a topic model context.

We use the tilde symbol ($\tilde{\ }$) to denote a corresponding topic in the base TSSB. When $\boldsymbol{\epsilon}$ is a topic (say, $\boldsymbol{\epsilon} = [1\ 1\ 4]$) in a child TSSB, $\tilde{\boldsymbol{\epsilon}}$ represents the same topic ([1 1 4]) in the base TSSB. We can determine the probabilities for vertical stopping at node $\boldsymbol{\epsilon}$ as follows, based on the theory of the HDP (Teh et al., 2006): $\nu_{\boldsymbol{\epsilon}} \sim \mathrm{Be}(a\tau_{\tilde{\boldsymbol{\epsilon}}}, a(1 - \sum_{\boldsymbol{\kappa} \preceq \tilde{\boldsymbol{\epsilon}}} \tau_{\boldsymbol{\kappa}}))$ where $\tau_{\boldsymbol{\epsilon}} = \nu_{\boldsymbol{\epsilon}} \prod_{\boldsymbol{\kappa} \prec \boldsymbol{\epsilon}} (1 - \nu_{\boldsymbol{\kappa}})$. Similarly, the probability for horizontal stopping at the $k$'th child of $\boldsymbol{\epsilon}$ is $\psi_{\boldsymbol{\epsilon}k} \sim \mathrm{Be}(b\phi_{\tilde{\boldsymbol{\epsilon}}k}, b(1 - \sum_{j=1}^{k} \phi_{\tilde{\boldsymbol{\epsilon}}j}))$. We can draw a topic tree, $\boldsymbol{\pi}$, for each document with these vertical and horizontal probabilities by using Equation (1). Note that the topic assignments in each document affect the base TSSB because each TSSB shares the same topics across the documents in the HTSSB.

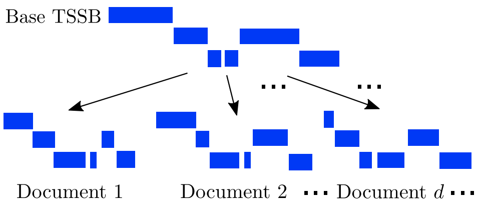


Figure 4: The HTSSB yields child TSSBs from the base TSSB. Although child TSSBs have the same topics as the base TSSB, the probabilities of the topics are different. Each child TSSB corresponds to the topic distribution for a single document. The horizontal probabilities in Figure 1 are omitted in this figure for simplicity.

### 4.2 Topic-Word Distributions

The hierarchical Pitman-Yor process (HPY; Teh, 2006) provides the semantic similarity between a parent topic and its children while increasing the specificity of the topics as the tree deepens. Let $H_{\boldsymbol{\epsilon}}$ be the probability distribution over words for topic $\boldsymbol{\epsilon}$. We use a Pitman-Yor process (Pitman and Yor, 1997; Goldwater et al., 2005) as a prior for $H_{\boldsymbol{\epsilon}}$: $H_{\boldsymbol{\epsilon}} \sim \mathrm{PY}(d_{|\boldsymbol{\epsilon}|}, \theta_{|\boldsymbol{\epsilon}|}, H_{\boldsymbol{\epsilon}'})$. We repeat this process until it reaches the root of the topic tree where we use $H_0$ as a prior: $H_{[1]} \sim \mathrm{PY}(d_0, \theta_0, H_0)$. If the size of the lexicon is $V$, we set $H_0 = 1/V$ for all words in the corpus. The tree structure of the topic-word distribution is the same as that of the base TSSB. Thus, all topics in a document have a corresponding topic-word distribution because a child TSSB is drawn from the base TSSB for each document.

### 4.3 Data Generation Process

We summarize the generation process of the documents in the ihLDA as follows: Let $\boldsymbol{\pi}^{(d)}$ specify a TSSB for a document $d$.

1. Draw a base TSSB $\tilde{\boldsymbol{\pi}}$.
2. Draw topic-word distributions $H_{\boldsymbol{\epsilon}}$ from the HPY for each topic in $\tilde{\boldsymbol{\pi}}$.
3. Draw a document-topic distribution for each document $d$, $\boldsymbol{\pi}^{(d)} \sim \mathrm{HTSSB}(\tilde{\boldsymbol{\pi}})$.
4. For each word position $i$ in a document $d$,
    - draw a topic, $z_{di} \sim \boldsymbol{\pi}^{(d)}$, and
    - draw a word, $w_{di} \sim H_{z_{di}}$.

## 5 Inference

### 5.1 Vertical and Horizontal Probabilities

The vertical and horizontal probabilities in Equations (2) and (3) determine the topic hierarchy. We employ the Chinese restaurant district process (CDP; Paisley and Carin, 2009, see Appendix F) representation of the Dirichlet process for each document-specific topic structure (child TSSB) and a shared topic structure (base TSSB).

We count the number of words that have stopped at a topic $\boldsymbol{\epsilon}$ as $n_0(\boldsymbol{\epsilon})$ for a vertical stop and $m_0(\boldsymbol{\epsilon})$ for a horizontal stop, along with the number of words that have passed $\boldsymbol{\epsilon}$ as $n_1(\boldsymbol{\epsilon})$ for a vertical pass and $m_1(\boldsymbol{\epsilon})$ for a horizontal pass. In addition, we define $n(\boldsymbol{\epsilon}) = n_0(\boldsymbol{\epsilon}) + n_1(\boldsymbol{\epsilon})$ and $m(\boldsymbol{\epsilon}) = m_0(\boldsymbol{\epsilon}) + m_1(\boldsymbol{\epsilon})$. After conditioning on the observed data and the rest of the probabilities, we can obtain the expectation of the posterior of the vertical and horizontal probabilities as: $\widehat{\nu}_{\boldsymbol{\epsilon}} = \mathbb{E}[\nu_{\boldsymbol{\epsilon}}\,|\,\mathrm{rest}] = (1 + n_0(\boldsymbol{\epsilon}))/(1 + \alpha + n(\boldsymbol{\epsilon}))$ and $\widehat{\psi}_{\boldsymbol{\epsilon}} = \mathbb{E}[\psi_{\boldsymbol{\epsilon}}\,|\,\mathrm{rest}] = (1 + m_0(\boldsymbol{\epsilon}))/(1 + \gamma + m(\boldsymbol{\epsilon}))$. Finally, using Equation (1), we can compute the expectation of the posterior $\pi_{\boldsymbol{\epsilon}}$ as $\mathbb{E}[\pi_{\boldsymbol{\epsilon}}|\text{rest}] = \widehat{\nu}_{\boldsymbol{\epsilon}} \prod_{\boldsymbol{\kappa} \prec \boldsymbol{\epsilon}} \left(1 - \widehat{\nu}_{\boldsymbol{\kappa}}\right) \cdot \prod_{\boldsymbol{\kappa} \preceq \boldsymbol{\epsilon}} \widehat{\phi}_{\boldsymbol{\epsilon}}$ where $\widehat{\phi}_{\boldsymbol{\epsilon}k} = \widehat{\psi}_{\boldsymbol{\epsilon}k} \prod_{j=1}^{k-1}(1 - \widehat{\psi}_{\boldsymbol{\epsilon}j})$.

Following the same idea, we apply a hierarchical CDP to the HTSSB. More specifically, when a word $w$ stops vertically at a topic $\boldsymbol{\epsilon}$ in a child TSSB for a document, we probabilistically update the counts of the corresponding topic $\tilde{\boldsymbol{\epsilon}}$ in the base TSSB with a probability proportional to $a\nu_{\tilde{\boldsymbol{\epsilon}}}/(a+n(\boldsymbol{\epsilon}))$. We update the count only in the child TSSB with a probability proportional to $n(\boldsymbol{\epsilon})/(a+n(\boldsymbol{\epsilon}))$. Horizontal probabilities have the same count update process: a probability proportional to $b\psi_{\tilde{\boldsymbol{\epsilon}}}/(b+m(\boldsymbol{\epsilon}))$ is used to update the base TSSB and $m(\boldsymbol{\epsilon})/(b+m(\boldsymbol{\epsilon}))$ for the child TSSB. The expectation of the posterior vertical and horizontal probabilities in a document $d$ is similar to that shown above,

$$\mathbb{E}\left[\nu_{\boldsymbol{\epsilon}}^{(d)}|\text{rest}\right] = \frac{a\tau_{\tilde{\boldsymbol{\epsilon}}} + n_0(\boldsymbol{\epsilon})}{a(1 - \sum_{\boldsymbol{\kappa} \prec \tilde{\boldsymbol{\epsilon}}} \tau_{\boldsymbol{\kappa}}) + n(\boldsymbol{\epsilon})},$$

$$\mathbb{E}\left[\psi_{\boldsymbol{\epsilon}k}^{(d)}|\text{rest}\right] = \frac{b\phi_{\tilde{\boldsymbol{\epsilon}}k} + m_0(\boldsymbol{\epsilon}k)}{b(1 - \sum_{j=1}^{k-1} \phi_{\tilde{\boldsymbol{\epsilon}}'j}) + m(\boldsymbol{\epsilon}k)}.$$

We employ slice sampling (Neal, 2003)[2] to estimate all hyperparameters in our model, that is, $\{\alpha_{|\boldsymbol{\epsilon}|}, \gamma_0, \lambda, a, b\}$.

### 5.2 Topic Assignments

The ihLDA has an infinite number of topics, and all possible topics in a tree cannot be enumerated. Our Gibbs sampling strategy implements a combination of retrospective sampling (Papaspiliopoulos and Roberts., 2008) and binary search, which follows the original approach used in the TSSB (Adams et al., 2010). The key observation is that each topic in a tree takes a certain share of a stick of length 1 (see Figure 1). Therefore, we draw a uniform random variable, $u \sim \text{Unif}\,[0, 1)$, to find a random topic that corresponds to $u$. Algorithm 1 outlines the Gibbs sampling process of topic assignment for each word. The function does not need to enumerate all the topics, because it only compares the new likelihood $q$ with the slice variable $\rho$. Algorithm 2 is a function for finding a topic that corresponds to a value in $[0, 1)$. This function rescales $u$ as it goes down the tree.

### 5.3 Other Parameters

Parameters in the HPY are also updated during the topic sampling in the HTSSB. Appendix B of Teh (2006) provides an inference strategy for sampling $\theta$ and $d$ used in the ihLDA.

**Algorithm 1** Gibbs sampling of a topic

```
function sample_assignment($\boldsymbol{\epsilon}$)
$a = 0$; $b = 1$; $\rho = \text{Unif}\,[0, 1) \cdot p(\boldsymbol{\epsilon})$
while True do
    $u = \text{Unif}\,[a, b)$
    $\boldsymbol{\epsilon}' = \text{find\_topic}(u, \boldsymbol{\epsilon}_{\text{root}})$
    $q = p(\boldsymbol{\epsilon}')$
    if $q > \rho$ then return $\boldsymbol{\epsilon}'$
    else
        if $\boldsymbol{\epsilon}' < \boldsymbol{\epsilon}$ then $b = u$ else $a = u$
    end if
end while
end function
```

**Algorithm 2** Finding a topic

```
function find_topic($u, \boldsymbol{\epsilon}$)
if $u < \nu_{\boldsymbol{\epsilon}}$ then
    return $\boldsymbol{\epsilon}$
else
    $u = (u - \nu_{\boldsymbol{\epsilon}})/(1 - \nu_{\boldsymbol{\epsilon}})$; $k = 1$
    while True do
        if $u < 1 - \prod_{j=1}^{k}(1 - \psi_{\boldsymbol{\epsilon}j})$ then break
        else
            $k \mathrel{+}= 1$
            Create $\boldsymbol{\epsilon}k$ if necessary
        end if
    end while
    $u = \dfrac{(u-1)(1-\psi_{\boldsymbol{\epsilon}k}) + \prod_{j=1}^{k}(1-\psi_{\boldsymbol{\epsilon}j})}{\psi_{\boldsymbol{\epsilon}k} \cdot \prod_{j=1}^{k}(1-\psi_{\boldsymbol{\epsilon}j})}$
    return find_topic($u, \boldsymbol{\epsilon}k$)
end if
end function
```

## 6 Experiments

### 6.1 Data

In our experiments, we used the *BBC News* corpus (Greene and Cunningham, 2006), the *20News* corpus (Lang, 1995), and the original *Wikipedia* corpus. The *BBC News* corpus contains 2,225 documents in five topic areas from the BBC news website, the *20News* corpus is a collection of 18,828 posts from 20 USENET newsgroups, and the *Wikipedia* corpus comprises 50,153 English articles randomly sampled from ten main categories[3] and their subcategories. We selected 80% of the data randomly for training.

### 6.2 Experimental Setup

We compared the ihLDA against two probabilistic and two neural topic models, namely, the nested

[2] Specifically, we used the unbounded slice sampling (Mochihashi, 2020) to sample from $[0, \infty)$ effectively.

[3] Art, engineering, computer science, food, humanities, medicine, nature, social science, sports, and statistics

| Model | Max Lvl. | Tree Diversity (↑) |  |  | Topic Uniqueness (↑) |  |  | Average Overlap (↓) |  |  | # of Topics |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | BBC | 20News | Wiki | BBC | 20News | Wiki | BBC | 20News | Wiki | BBC | 20News | Wiki |
| ihLDA | 3 | 2.24 | **2.88** | **2.63** | **0.60** | **0.82** | **0.66** | 0.28 | 0.11 | 0.16 | 38 | 27 | 17 |
|  |  | (2.24) | (2.86) | (2.49) | (0.60) | (0.80) | (0.63) | (0.28) | (0.14) | (0.19) | (38) | (31) | (18) |
|  | $\geq 4$ | **2.53** | **2.88** | 2.50 | 0.55 | 0.76 | 0.65 | 0.26 | 0.12 | 0.15 | 85 | 67 | 73 |
|  |  | (2.54) | (2.80) | (2.51) | (0.49) | (0.51) | (0.63) | (0.30) | (0.38) | (0.16) | (134) | (203) | (101) |
| nCRP | 3 | 1.92 | 2.16 | – | 0.36 | 0.32 | – | **0.03** | 0.02 | – | 517 | 2108 | – |
| rCRP |  | 0.15 | – | – | 0.01 | – | – | 0.53 | – | – | 278 | – | – |
| TSNTM |  | 1.98 | 2.54 | 2.47 | 0.43 | 0.80 | 0.64 | 0.26 | 0.09 | 0.06 | 22 | 41 | 44 |
| nTSNTM |  | 2.11 | 2.57 | 2.34 | 0.46 | 0.68 | 0.60 | 0.09 | **0.01** | **0.02** | 68 | 81 | 111 |

Table 2: Evaluation on different corpora. The proposed ihLDA performs better than existing models in two measurements, Tree Diversity (higher is better) and Topic Uniqueness (higher is better). Average Overlap (lower is better) might have a limitation explained in the main text. Existing probabilistic models (nCRP and rCRP) create tiny topics, whereas neural models (nTSNTM and TSNTM) regularize the topics. For comparison, we truncate the topics that do not have at least 100 assigned words. The results without truncation are shown in parentheses for the ihLDA.

Chinese restaurant process (nCRP; Blei et al., 2003a), the recursive Chinese restaurant process (rCRP; Kim et al., 2012), the tree-structured neural topic model (TSNTM; Isonuma et al., 2020), and the nonparametric tree-structured neural topic model (nTSNTM; Chen et al., 2021). The publicly available replication codes for the rCRP, TSNTM, and nTSNTM were used along with a package for the nCRP (Lee, 2021). We used the default parameter values. As both neural models internally truncate the topics, the results were based on topics with at least 100 assigned words for a fair comparison. The maximum level of the ihLDA was six for *BBC News* and *20News* even when we made the model unbounded but we truncated the tree at four for *Wikipedia*. All the experiments were conducted on a cluster computer with a Python 3 environment (Intel Xeon CPU 2.2-2.3 GHz and 10 GB RAM). We do not report the results of nCRP on *Wikipedia* and rCRP on *20News* and *Wikipedia*, because they required more than two weeks to complete 10,000 iterations.

### 6.3 Numerical Evaluation

We employed two measures (TU and AO) from the existing literature and developed a new measure (TD) to compare the performance of the ihLDA with those of the existing approaches.

First, the topic uniqueness (TU) calculates the uniqueness of all topics (Nan et al., 2019; Masson and Montariol, 2020; Chen et al., 2021). A higher TU implies that the topics represent unique themes. Second, the average overlap (AO) measures the average repetition rate of the top $u$ words between the parent topic and its children (Chen et al., 2021). A lower AO indicates that less overlap occurs between the top words from a parent and those from its children. Although this measure was used in Chen et al. (2021), parent and child topics need some overlapping words to have semantic coherence; thus, a smaller AO does not always mean better interpretability. Appendix G provides formal definitions of these two measures.

Finally, the tree diversity (TD) is a new measure for assessing child topics as being unique, while considering the importance of the parent topics. Let $\mathcal{T}$ be a set of topics in the estimated tree, $\mathcal{C}(\boldsymbol{\epsilon})$ be a set of topics that are the children of a topic $\boldsymbol{\epsilon}$, $\mathcal{D}(\boldsymbol{\epsilon})$ be a set of topics that are descendants of a topic $\boldsymbol{\epsilon}$, and $\mathcal{V}_{\mathcal{N}}$ be a set of unique words that are used for the top $u$ words of a set of topics $\mathcal{N}$. We define TD as follows:

$$\mathrm{TD} = \sum_{\boldsymbol{\epsilon} \in \mathcal{T}} w_{\boldsymbol{\epsilon}} \frac{|\mathcal{V}_{\mathcal{C}(\boldsymbol{\epsilon})}|}{u|\mathcal{C}(\boldsymbol{\epsilon})|} \; ; \; w_{\boldsymbol{\epsilon}} = \frac{|\mathcal{D}(\boldsymbol{\epsilon})|}{\sum_{\boldsymbol{\kappa} \in \mathcal{T}} |\mathcal{D}(\boldsymbol{\kappa})|}.$$

The fraction in TD is the proportion of unique words among the top words of the children of $\boldsymbol{\epsilon}$. Then it takes the sum of the fraction weighted by the normalized importance of each topic, that is, the proportion of descendants of $\boldsymbol{\epsilon}$. A higher TD is better because it implies that the top words in child topics contain more unique words.

Table 2 summarizes the results and the estimated number of topics. All metrics are calculated with different numbers of top words ($u = 5, 10$, and 15), and we report their average. The ihLDA performs better than the existing models in terms of the TD and TU. Existing probabilistic models (nCRP and rCRP) create too many topics in comparison with the ihLDA, even though they truncate the topic tree at three levels. The ihLDA shows a reasonable number of topics even when it has a deeper tree without truncation, as shown in the parentheses. The two neural models, TSNTM and nTSNTM, find fewer topics than the ihLDA but have lower performance in the *BBC News* and *20 News* corpora and have a lot of redundancy as shown in Table 1. With the *Wikipedia* corpus, the ihLDA estimates 17 topics when the depth is fixed at three, which is a reasonable number given that the *Wikipedia* corpus is sampled from ten categories and their subcategories (see footnote 3).

### 6.4 Crowdsourced Evaluation

We devised three human evaluation tasks to assess both the interpretability and the hierarchical structure of the topics. An interpretable hierarchical topic model should show similarity between parent and child topics, while each child topic is coherent and distinctive from others.

The first task, *Word Intrusion*, is a slight alteration from the methods in Chang et al. (2009) and Ying et al. (2022). To measure the coherence of the estimated topics, crowdsourced workers observed four different word sets (each word set consists of four words). Three word sets were randomly selected from the top words of one topic, whereas the other set (the *intruder*) was randomly selected from those of a different topic that did not share the parent topic with the three word sets. The "correct" answer means that a worker identified the intruder word set.

The second task, *Vertical*, is an original task to measure the hierarchical structure. The workers observed four items and categorized them into two groups. We represented the items and groups with four words randomly chosen from the top words, where each item was a child topic of one of the groups. The "correct" answer means that a worker categorized a child topic into its parent topic.

The third task, *Horizontal*, is also an original task to measure horizontal distinctiveness. The workers grouped four items represented by four words randomly selected from the top words of topics that had the same parent topic. The same topic could appear in multiple items. The "correct" answer means that a worker categorized items from the same topic into the same group. If a model estimates overlapping topics, a worker cannot provide the correct answer.

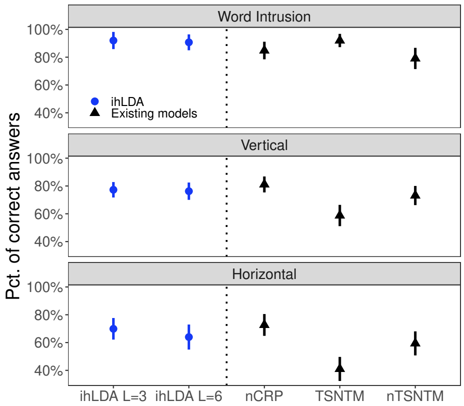


Figure 5: Crowdsourced validation. The performance of the ihLDA (the maximum levels are three and six) is at least statistically indistinguishable from the best existing model and better than the worst existing model.

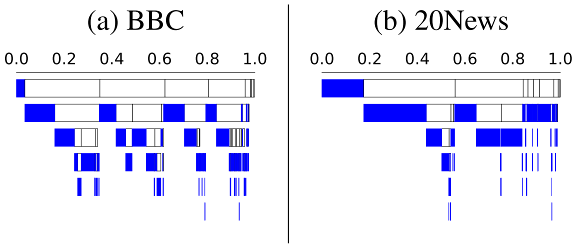


Figure 6: The estimated global base tree prior $\widetilde{\pi}$ in HTSSB. Topics do not decay exponentially along the depth of the tree.

We used the outputs from the *BBC News* corpus because news articles are accessible and familiar to crowdsourced workers from Amazon Mechanical Turk (Ying et al., 2022). We dropped workers who failed to pass our quality check questions and those who spent too little (bottom 10%) or too much (top 10%) time[4]. Appendix H describes more details of crowdsourced experiments.

Figure 5 illustrates the proportion of correct answers, weighted to represent each level equally. The performance of the proposed model is at least statistically indistinguishable from the best existing model and better than the worst existing model in all tasks. nCRP exhibits competitive performance, but this is because it creates numerous specific topics even for a small corpus as shown in Table 2.

### 6.5 Estimated Tree Structure

Figure 6 displays $\widetilde{\pi}$, the estimated global tree prior in the ihLDA. Both (a) and (b) show that topics do not decay significantly even if the maximum level is six.

Figure 7 presents the top five words for some topics that facilitate comparison between models. The ihLDA estimates topics with general words in the first and the second levels, and topics become more specific at lower levels. Figure 8 supports this topic specificity: proper nouns have higher

[4] The total number of observations was 535 (Word Intrusion), 900 (Vertical), and 620 (Horizontal).

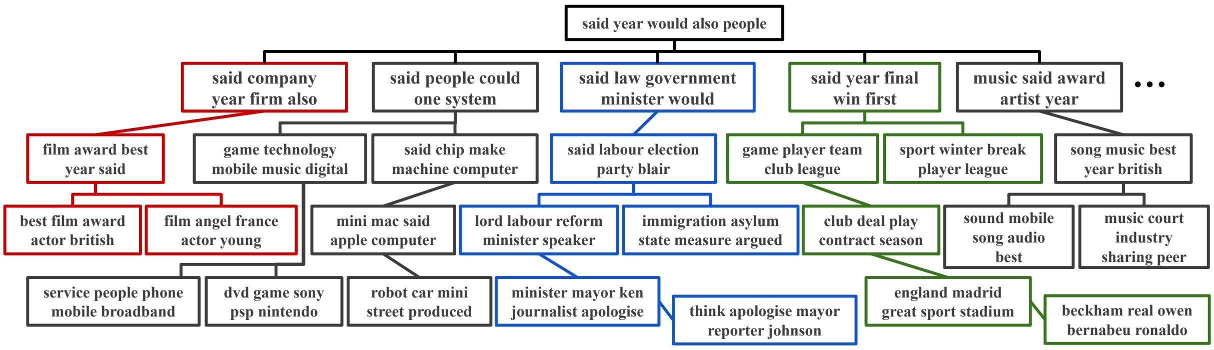

said year would also people
said company year firm also
said people could one system
said law government minister would
said year final win first
music said award artist year
...
film award best year said
game technology mobile music digital
said chip make machine computer
said labour election party blair
game player team club league
sport winter break player league
song music best year british
best film award actor british
film angel france actor young
mini mac said apple computer
lord labour reform minister speaker
immigration asylum state measure argued
club deal play contract season
sound mobile song audio best
music court industry sharing peer
service people phone mobile broadband
dvd game sony psp nintendo
robot car mini street produced
minister mayor ken journalist apologise
think apologise mayor reporter johnson
england madrid great sport stadium
beckham real owen bernabeu ronaldo

Figure 7: Top words from the selected topics (BBC corpus). The maximum level is six. Figure 8 explains the colored branches.

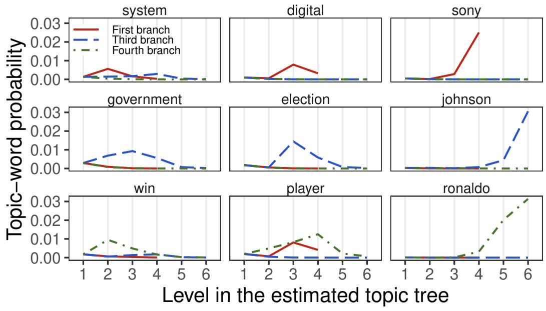


Figure 8: Selected topic-word probabilities from the first (red), third (blue), and fourth (green) branches from the left in Figure 7. Proper nouns have higher probabilities at the bottom of the tree, while general terms appear more frequently at the top levels. Each branch represents a different theme.

probabilities at the bottom of the tree, whereas general terms appear more frequently at the top levels.

## 7 Related Work

The existing models have heuristically addressed the issue of topic fragmentation by truncating topics at a certain threshold and truncating the tree structure to a small number of levels. Adams et al. (2010) constructed a hierarchical document model with the TSSB[5] and truncated the topics with less than 50 assigned documents when the corpus had 1740 documents. The nTSNTM (Chen et al., 2021) sequentially selected the topics until the sum of probabilities in the corpus exceeded 95%. Existing probabilistic approaches (Blei et al., 2003a, 2010; Kim et al., 2012; Paisley et al., 2015) only consider three levels. Isonuma et al. (2020) introduced neural architectures but fixed the number of levels to three with an initial number of three branches for both the second and third levels.

Another advantage of the ihLDA is that it employs a hierarchical Bayesian extension of the TSSB to draw a child TSSB from the base TSSB (see Figure 4). Unlike some probabilistic models that restrict a document-topic distribution to a single or multiple topic-path on a tree (Blei et al., 2003a, 2010; Paisley et al., 2015), ihLDA does not limit topics that can appear in a document.

Hierarchical topic models have a wide range of extensions (Mao et al., 2012; Yang and Hsu, 2016; Shin and Moon, 2017; Xu et al., 2018; Zou et al., 2019; Isonuma et al., 2021), and the ihLDA is orthogonal to them and useful for these extensions.

## 8 Conclusion

Existing hierarchical topic models yield topics with exponentially smaller probabilities. To address this intrinsic issue, we propose the ihLDA, a nonparametric Bayesian model that learns a latent topic hierarchy with arbitrary depth and width. Our model adjusts topic creation to achieve the expected topic probability without dependence on its depth, which can also improve other models that use the stick-breaking process. As a topic model, the ihLDA is a hierarchical extension of the TSSB and draws topic assignments efficiently without enumerating all possible candidates. Our experiments on standard document datasets confirm that the ihLDA outperforms the existing methods, including the latest neural models, and extracts meaningful topic structures with better hierarchical diversity and uniqueness.

[5] The topic model in Adams et al. (2010) differs from our setting. In their experiments, each node has a unique topic distribution.

## Limitations

Although the ihLDA shows better performance than existing models in multiple experiments, there are three limitations that we did not fully address in this paper.

First, the Gibbs sampling is slower than other approaches such as autoencoding variational Bayes (Kingma and Welling, 2014), which limits data scalability. We can incorporate the literature on distributed algorithms for topic modeling (Newman et al., 2009; Yu et al., 2015; Karras et al., 2022) and variational inference (Wang and Blei, 2009; Wang et al., 2011; Bryant and Sudderth, 2012; Hughes et al., 2015) in future research.

Second, crowdsourced evaluation limits a corpus choice because we should not expect workers to have any prior knowledge (Ying et al., 2022). Our crowdsourced evaluation only used *BBC News*, the most accessible documents among the three corpora. Future research can thoroughly validate the performance of the crowdsourced workers and trained coders. Existing literature (Buhrmester et al., 2016; Kees et al., 2017) found that MTurk had a comparable quality against traditional survey panels, but they did not use MTurk for evaluating outputs from a machine learning model.

Third, an estimated hierarchical structure does not necessarily match the semantic hierarchy human readers expect. This mismatch is not surprising because unsupervised models do not directly incorporate information about a tree structure. Existing papers improved the interpretability of flat topic models by providing topic-specific sets of keywords (Jagarlamudi et al., 2012; Harandizadeh et al., 2022) and labels (Mcauliffe and Blei, 2007; Ramage et al., 2009), which is a future direction for a hierarchical topic model.

## Acknowledgements

We thank Adam Breuer, Dean Knox, Tomoya Sasaki, Yuki Shiraito, Soichiro Yamauchi, and members of the Imai Research Group at Harvard University for helpful discussions and comments on this project. We would also like to acknowledge anonymous reviewers for their constructive feedback.

## References

Ryan P. Adams, Zoubin Ghahramani, and Michael I. Jordan. 2010. Tree-structured stick breaking for hierarchical data. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, pages 19–27.

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2010. The nested Chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):1–30.

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. 2003a. Hierarchical topic models and the nested Chinese restaurant process. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pages 17–24.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003b. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Michael Bryant and Erik Sudderth. 2012. Truly nonparametric online variational inference for hierarchical dirichlet processes. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2016. Amazon's mechanical turk: A new source of inexpensive, yet high-quality data?

Jonathan Chang and David M Blei. 2010. Hierarchical relational models for document networks. *The Annals of Applied Statistics*, 4:124–150.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models.

Jianfei Chen, Kaiwei Li, Jun Zhu, and Wenguang Chen. 2016. WarpLDA: A cache efficient O(1) algorithm for latent Dirichlet allocation. *Proceedings of the VLDB Endowment*, 9.

Ziye Chen, Cheng Ding, Zusheng Zhang, Yanghui Rao, and Haoran Xie. 2021. Tree-structured topic modeling with nonparametric neural variational inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2343–2353.

Jen-Tzung Chien. 2016. Hierarchical theme and topic modeling. *IEEE Transactions on Neural Networks and Learning Systems*, 27(3):565–578.

Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. 2015. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome biology*, 16(1):1–20.

Zhibin Duan, Dongsheng Wang, Bo Chen, Chaojie Wang, Wenchao Chen, Yewen Li, Jie Ren, and Mingyuan Zhou. 2021. Sawtooth factorial topic embeddings guided gamma belief network. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2903–2913.

Sharon Goldwater, Mark Johnson, and Thomas Griffiths. 2005. Interpolating between types and tokens by estimating power-law generators. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, volume 18.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd International Conference on Machine learning*, pages 377–384. ACM Press.

Bahareh Harandizadeh, J Hunter Priniski, and Fred Morstatter. 2022. Keyword assisted embedded topic model. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 372–380.

Michael Hughes, Dae Il Kim, and Erik Sudderth. 2015. Reliable and Scalable Variational Inference for the Hierarchical Dirichlet Process. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 370–378. PMLR.

Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2020. Tree-structured neural topic model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 800–806.

Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2021. Unsupervised abstractive opinion summarization by generating sentences with tree-structured topic guidance. *Transactions of the Association for Computational Linguistics*, 9:945–961.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213.

Christos Karras, Aristeidis Karras, Dimitrios Tsolis, Konstantinos C Giotopoulos, and Spyros Sioutas. 2022. Distributed gibbs sampling and lda modelling for large scale big data management on pyspark. In *2022 7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, pages 1–8. IEEE.

Jeremy Kees, Christopher Berry, Scot Burton, and Kim Sheehan. 2017. An analysis of data quality: Professional panels, student subject pools, and amazon's mechanical turk. *Journal of Advertising*, 46(1):141–155.

Joon Hee Kim, Dongwoo Kim, Suin Kim, and Alice Oh. 2012. Modeling topic hierarchies with the recursive Chinese restaurant process. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 783–792. ACM Press.

Diederik P Kingma and Max Welling. 2014. Stochastic gradient vb and the variational auto-encoder. In *Proceedings of the 2nd International Conference on Learning Representations*, volume 19, page 121.

Ken Lang. 1995. NewsWeeder: Learning to filter Netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339.

Minchul Lee. 2021. tomotopy.

Aaron Q Li, Amr Ahmed, Sujith Ravi, and Alexander J Smola. 2014. Reducing the sampling complexity of topic models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 891–900.

Xian-Ling Mao, Zhao-Yan Ming, Tat-Seng Chua, Si Li, Hongfei Yan, and Xiaoming Li. 2012. SSHLDA: A semi-supervised hierarchical topic model. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 800–809.

Corentin Masson and Syrielle Montariol. 2020. Detecting omissions of risk factors in company annual reports. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 15–21.

Jon Mcauliffe and David Blei. 2007. Supervised topic models. *Advances in neural information processing systems*, 20.

Daichi Mochihashi. 2020. Unbounded slice sampling. *ISM Research Memorandum No. 1209.*

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.*

Josue Nassar, Scott W. Linderman, Monica Bugallo, and Il Memming Park. 2019. Tree-structured recurrent switching linear dynamical systems for multi-scale modeling. In *International Conference on Learning Representations.*

Radford M Neal. 2003. Slice sampling. *The Annals of Statistics*, 31:705–767.

David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2009. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10(8).

John Paisley and Lawrence Carin. 2009. Hidden markov models with stick-breaking priors. *IEEE Transactions on Signal Processing*, 57(10):3905–3917.

John Paisley, Chong Wang, David M. Blei, and Michael I. Jordan. 2015. Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):256–270.

Omiros Papaspiliopoulos and Gareth O. Roberts. 2008. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186.

Jim Pitman and Marc Yor. 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 248–256.

Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. 2016. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111:988–1003.

Jayaram Sethuraman. 1994. A constructive definition of dirichlet priors. *Statistica Sinica*, 4(2):639–650.

Su Jin Shin and Il Chul Moon. 2017. Guided HTM: Hierarchical topic model with dirichlet forest priors. *IEEE Transactions on Knowledge and Data Engineering*, 29(2):330–343.

Yee Whye Teh. 2006. A Bayesian interpretation of interpolated Kneser-Ney. Technical report, National University of Singapore School of Computing.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Chong Wang and David Blei. 2009. Variational inference for the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, volume 22.

Chong Wang, John Paisley, and David M. Blei. 2011. Online variational inference for the hierarchical dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 752–760. PMLR.

Yueshen Xu, Jianwei Yin, Jianbin Huang, and Yuyu Yin. 2018. Hierarchical topic modeling with automatic knowledge mining. *Expert Systems with Applications*, 103:106–117.

Ming Yang and William H. Hsu. 2016. HDPauthor: A new hybrid author-topic model using latent Dirichlet allocation and hierarchical Dirichlet processes. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 619–624.

Luwei Ying, Jacob M Montgomery, and Brandon M Stewart. 2022. Topics, concepts, and measurement: A crowdsourced procedure for validating topics as measures. *Political Analysis*, 30(4):570–589.

Hsiang-Fu Yu, Cho-Jui Hsieh, Hyokun Yun, S V N Vishwanathan, and Inderjit S Dhillon. 2015. A scalable asynchronous distributed algorithm for topic modeling. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1340–1350.

Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric P. Xing, Tie Yan Liu, and Wei Ying Ma. 2015. LightLDA: Big topic models on modest computer clusters. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1351–1361.

Xi Zou, Yuelong Zhu, Jun Feng, Jiamin Lu, and Xiaodong Li. 2019. A novel hierarchical topic model for horizontal topic expansion with observed label information. *IEEE Access*, 7:184242–184253.

# Appendix

## A Dirichlet Process

The Dirichlet process (DP) is the foundation of nonparametric Bayesian models. As the ihLDA is an infinite mixture model (i.e., an infinite number of topics can exist), we draw the topics using the DP.

Formally, $G$ is a DP with a base distribution $G_0$ and a concentration parameter $c$:

$$G \sim \mathrm{DP}(c, G_0). \tag{4}$$

The DP has three representations: the stick-breaking process, the Chinese restaurant process, and the Chinese restaurant district process.

## B Stick-Breaking Process

Formally, the stick-breaking representation of a DP with a base distribution $G_0$ and a concentration parameter $c$, $G \sim \mathrm{DP}(c, G_0)$, is

$$\begin{aligned} &G = \sum_{k=1}^{\infty} \delta_{\eta_k} \pi_k, \ \pi_k = v_k \prod_{j=1}^{k-1} (1 - v_j), \\ &v_k \sim \mathrm{Be}(1, c), \ \eta_k \sim G_0, \end{aligned}$$

where $\eta_k$ takes a distinct value to indicate a single category. Figure 9 depicts this process.

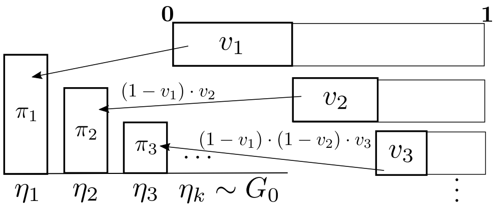


Figure 9: Stick-breaking process. This process sequentially breaks a stick of length 1 proportional to $v_k$. The length of a broken piece represents the probability of the category.

## C An Example of TSSB

In a hierarchical topic model, we consider both the vertical and horizontal placements of each word in a tree. To assign a topic for each word in the corpus, we first determine whether the word uses the root topic [1] (i.e., the word stops at [1]) or not (i.e., the word passes [1]) according to a vertical probability $\nu_{[1]} \sim \mathrm{Be}(1, \alpha_0)$. The probability of stopping at [1] is thus $\pi_{[1]} = \nu_{[1]}$. If the word passes [1], it goes down to the next level. Each solid line in Figure 1 connects a parent topic to its children in the next level. At the second level, horizontal stopping probabilities, $\psi_{[1\,1]}, \psi_{[1\,2]}, \psi_{[1\,3]}, \cdots$, determine the child topic to descend. Subsequently, the vertical probabilities $\nu_{[1\,1]}, \nu_{[1\,2]}, \nu_{[1\,3]}, \cdots$, decide whether the word stops or proceeds further down the tree. We repeat this process until the word stops both vertically and horizontally. For example, the probability of using the topic [1 2 2] can be computed as follows:

$$\begin{aligned}\pi_{[1\,2\,2]} = &\left(1-\nu_{[1]}\right)\\ &\times \left(1-\nu_{[1\,2]}\right)\cdot\psi_{[1\,2]}\cdot\left(1-\psi_{[1\,1]}\right)\\ &\times \nu_{[1\,2\,2]}\cdot\psi_{[1\,2\,2]}\cdot\left(1-\psi_{[1\,2\,1]}\right).\end{aligned}$$

## D The Expected Probability of Topics in Hierarchical Stick-Breaking Process

We consider the following stick-breaking process:

$$\phi_k = v_k \prod_{j=1}^{k-1}(1-v_j),\ v_k \sim \mathrm{Be}(1,\gamma).$$

The expectation of parameter $v_k$ is $\mathbb{E}[v_k] = 1/(1+\gamma)$; hence, the expected probability of the $k$th broken stick is,

$$\mathbb{E}[\phi_k] = \frac{1}{1+\gamma}\left(\frac{\gamma}{1+\gamma}\right)^{k-1} = \frac{1}{\gamma}\left(\frac{\gamma}{1+\gamma}\right)^{k}.$$

Next, we consider the expected probability of a topic in the stick-breaking process,

$$\begin{aligned}\mathbb{E}[\phi] &= \sum_{k=1}^{\infty}\mathbb{E}[\phi_k]\cdot\phi_k\\ &\approx \sum_{k=1}^{\infty}\mathbb{E}[\phi_k]^2\\ &= \sum_{k=1}^{\infty}\left(\frac{\gamma}{\gamma+1}\cdot\frac{1}{\gamma}\right)^2\\ &= \frac{1}{\gamma^2}\sum_{k=1}^{\infty}\left(\frac{1}{(1+\frac{1}{\gamma})^2}\right)^k\\ &= \frac{1}{\gamma^2}\cdot\frac{1}{(1+\frac{1}{\gamma})^2-1}\\ &= \frac{1}{2\gamma+1},\end{aligned}$$

where the expectation of $\phi_k$ is used for the approximation. Using the standard stick-breaking process, the expected probability of the topic at the $\ell$th level is

$$\mathbb{E}[\phi\,|\,\ell] = \mathbb{E}[\phi\,|\,\ell-1]\cdot\mathbb{E}[\phi] \approx \frac{1}{(2\gamma+1)^{\ell}},$$

where the first equality means that the $(\ell-1)$th level stick is broken at the $\ell$th level. Because of the modification described in Section 3, the expectation becomes

$$\begin{aligned}\mathbb{E}[\phi\,|\,\ell] &= \mathbb{E}[\phi\,|\,\ell-1]\cdot\mathbb{E}[\phi]\\ &\approx \mathbb{E}[\phi\,|\,\ell-1]\cdot\frac{1}{2(\gamma\cdot\mathbb{E}[\phi\,|\,\ell-1])+1}\\ &= \frac{1}{2\gamma+1/\mathbb{E}[\phi\,|\,\ell-1]}\quad\text{for } \ell\geq 2\,.\end{aligned}$$

If $\ell = 1$ (the root level), then $\mathbb{E}[\phi\,|\,\ell=1] = 1/(2\gamma+1)$. The expected probability of the topic does not become exponentially smaller even when proceeding down the tree.

## E Hierarchical Dirichlet Process

Suppose that the global distribution of topics $G$ is distributed as a DP with the concentration parameter $c$: $G \sim \mathrm{DP}(c, G_0)$. The actual distribution over the topics in the $d$th document, $G_d$, follows another DP, $G_d \sim \mathrm{DP}(c_0, G)$; hence, the distribution of $G_d$ varies around $G$. Given $G_d$, we can draw a topic assignment for each word in the $d$th document.

## F Chinese Restaurant District Process

As shown in Figure 10, the CDP uses the counts of $n$ words to determine a category $z$ for the next

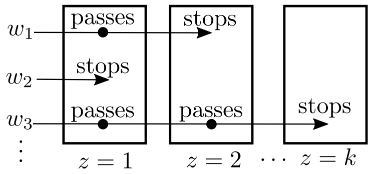
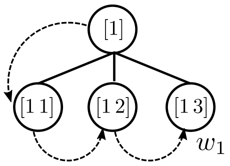


Figure 10: Chinese restaurant district process (CDP). Each word either passes or stops at a category. The CDP creates a new category if a word does not stop at any existing category.



| Topic | $n_0$ | $n_1$ | $m_0$ | $m_1$ |
|---|---|---|---|---|
| [1] | 0 | 1 | 1 | 0 |
| [1 1] | 0 | 0 | 0 | 1 |
| [1 2] | 0 | 0 | 0 | 1 |
| [1 3] | 1 | 0 | 1 | 0 |

Figure 11: Updating counts according to the CDP. The first word $w_1$ passes [1], [1 1], and [1 2], and then stops at [1 3].

word:

$$p(z_{n+1} \geq k \,|\, \mathbf{z}_{1:n}) = \frac{1 + \sum_{j=k+1}^{\infty} n_j}{1 + \alpha + \sum_{j=k}^{\infty} n_j},$$

where $\alpha$ is the concentration parameter corresponding to Equation (2). In the CDP terminology, a word using the $k$th category is referred to as "stopping at $k$" and that using the $j$th ($j > k$) category is referred to as "passing $k$". Each word passes through categories until it stops; hence, we keep track of the number of data points that stopped and passed at each category.

To compute the vertical and horizontal probabilities $\nu_{\boldsymbol{\epsilon}}$ and $\psi_{\boldsymbol{\epsilon}}$, we count the number of words that have stopped at a topic $\boldsymbol{\epsilon}$ as $n_0(\boldsymbol{\epsilon})$ for a vertical stop and $m_0(\boldsymbol{\epsilon})$ for a horizontal stop, as well as the number of words that have passed $\boldsymbol{\epsilon}$ as $n_1(\boldsymbol{\epsilon})$ for a vertical pass and $m_1(\boldsymbol{\epsilon})$ for a horizontal pass.

Suppose that the first word stops at [1 3] in Figure 11. For this to occur, the word passes the root topic, [1], goes down to the next level, and passes two child topics, [1 1] and [1 2]. Hence, $n_0([1]) = 0$ and $n_1([1]) = 1$ when passing the root topic, and $m_0([1\,1]) = m_0([1\,2]) = 0$, $m_1([1\,1]) = m_1([1\,2]) = 1$, $m_0([1\ 3]) = 1$, and $m_1([1\ 3]) = 0$ when horizontally stopping at the third child of the root topic. As the word vertically stops at the topic [1 3], the vertical count becomes $n_0([1\ 3]) = 1$ and $n_1([1\ 3]) = 0$. If the word vertically passes the topic [1 3] and further goes down the topic tree, then $n_0([1\ 3]) = 0$ and $n_1([1\ 3]) = 1$. In addition, we define $n(\boldsymbol{\epsilon}) = n_0(\boldsymbol{\epsilon}) + n_1(\boldsymbol{\epsilon})$ and $m(\boldsymbol{\epsilon}) = m_0(\boldsymbol{\epsilon}) + m_1(\boldsymbol{\epsilon})$.

Using pass and stop counts from the CDP, we can obtain the posterior distribution of the vertical and horizontal probabilities, $\nu_{\boldsymbol{\epsilon}}$ and $\psi_{\boldsymbol{\epsilon}}$, because the construction of $\pi_{\boldsymbol{\epsilon}}$ is the result of choosing "stop" or "pass" on the way to reach $\boldsymbol{\epsilon}$:

$$\nu_{\boldsymbol{\epsilon}} \,|\, \text{rest} \sim \text{Be}(1 + n_0(\boldsymbol{\epsilon}), \alpha + n_1(\boldsymbol{\epsilon})) \tag{5}$$

$$\psi_{\boldsymbol{\epsilon}} \,|\, \text{rest} \sim \text{Be}(1 + m_0(\boldsymbol{\epsilon}), \gamma + m_1(\boldsymbol{\epsilon})) \tag{6}$$

Note that each probability is conditioned on the observed data and rest of the probabilities. By taking the expectations of Equations (5) and (6), we obtain

$$\begin{aligned} \widehat{\nu}_{\boldsymbol{\epsilon}} &= \mathbb{E}[\nu_{\boldsymbol{\epsilon}} \,|\, \text{rest}] = \frac{1 + n_0(\boldsymbol{\epsilon})}{1 + \alpha + n(\boldsymbol{\epsilon})}, \\ \widehat{\psi}_{\boldsymbol{\epsilon}} &= \mathbb{E}[\psi_{\boldsymbol{\epsilon}} \,|\, \text{rest}] = \frac{1 + m_0(\boldsymbol{\epsilon})}{1 + \gamma + m(\boldsymbol{\epsilon})}. \end{aligned} \tag{7}$$

## G Details of Evaluation Measures

### G.1 Topic Uniqueness

Topic uniqueness (TU) calculates the uniqueness of all topics (Nan et al., 2019; Masson and Montariol, 2020; Chen et al., 2021). Let $\mathcal{T}$ be a set of topics in the estimated tree. We define TU as follows:

$$\text{TU} = \frac{1}{|\mathcal{T}|} \sum_{\boldsymbol{\epsilon} \in \mathcal{T}} \left( \frac{1}{u} \sum_{u'=1}^{u} \frac{1}{n(u', \boldsymbol{\epsilon})} \right),$$

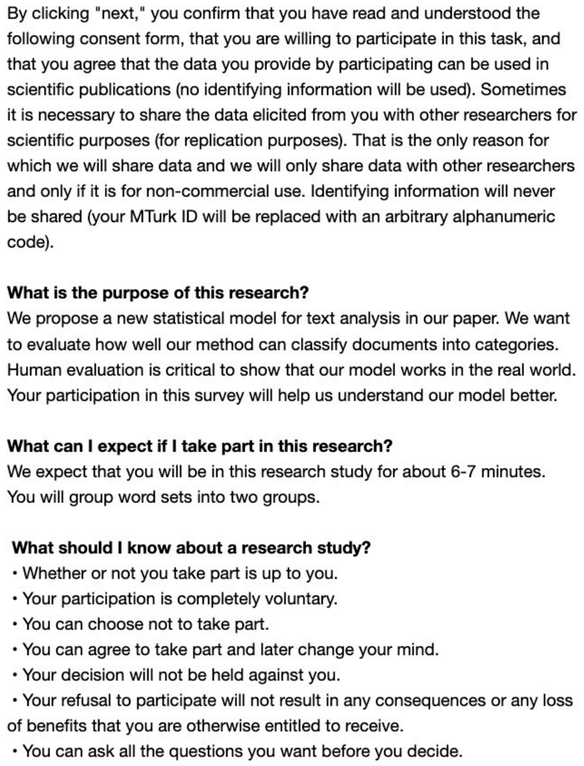
where $n(u', \boldsymbol{\epsilon})$ is the total number of times that the $u'$th top word in topic $\boldsymbol{\epsilon}$ appears in the top $u$ words across all topics. A higher TU implies that the topics represent unique themes.

### G.2 Average Overlap

Average overlap (AO) measures the average repetition rate of the top $u$ words between the parent topic and its children (Chen et al., 2021),

$$\text{AO} = \frac{1}{|\mathcal{T}|} \sum_{\boldsymbol{\epsilon} \in \mathcal{T}} \frac{|\mathcal{V}_{\boldsymbol{\epsilon}} \cap \mathcal{V}_{\boldsymbol{\epsilon}'}|}{u},$$

where $\mathcal{V}_{\boldsymbol{\epsilon}}$ is a set of unique words that appear in the top $u$ words of a node $\boldsymbol{\epsilon}$. A lower AO indicates that less overlap occurs between the top words from a parent and those from its children. Although this measure was used in Chen et al. (2021), parent and child topics need some overlapping words to have semantic coherence; thus, less overlap does not necessarily mean better interpretability.



By clicking "next," you confirm that you have read and understood the following consent form, that you are willing to participate in this task, and that you agree that the data you provide by participating can be used in scientific publications (no identifying information will be used). Sometimes it is necessary to share the data elicited from you with other researchers for scientific purposes (for replication purposes). That is the only reason for which we will share data and we will only share data with other researchers and only if it is for non-commercial use. Identifying information will never be shared (your MTurk ID will be replaced with an arbitrary alphanumeric code).

**What is the purpose of this research?**
We propose a new statistical model for text analysis in our paper. We want to evaluate how well our method can classify documents into categories. Human evaluation is critical to show that our model works in the real world. Your participation in this survey will help us understand our model better.

**What can I expect if I take part in this research?**
We expect that you will be in this research study for about 6-7 minutes. You will group word sets into two groups.

**What should I know about a research study?**
- Whether or not you take part is up to you.
- Your participation is completely voluntary.
- You can choose not to take part.
- You can agree to take part and later change your mind.
- Your decision will not be held against you.
- Your refusal to participate will not result in any consequences or any loss of benefits that you are otherwise entitled to receive.
- You can ask all the questions you want before you decide.

Figure 12: The consent form used in the crowdsourced evaluation. We did not include the contact information in this screenshot. The Institutional Review Board reviewed this consent form.

# H Additional Information for the Crowdsourced Evaluation

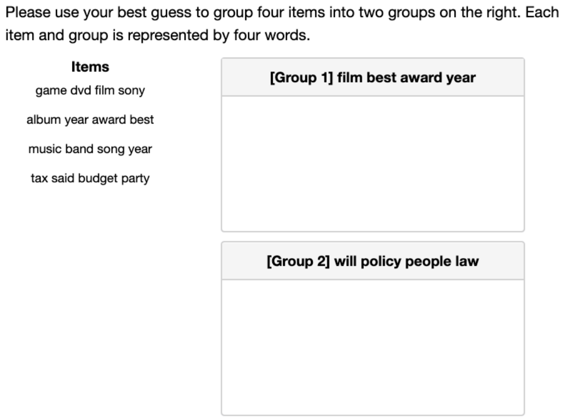
## H.1 Design

We recruited participants via Amazon Mechanical Truk and used Qualtrics to prepare our evaluation tasks. Once participants agreed on the consent form (Figure 12), they read the instruction before conducting five tasks (Figure 13). We compensated the participants through payment ($0.5 to $0.55 per participant). The amount of compensation is determined to match the federal minimum wage in the United States.

## H.2 Quality Check

The last task was the same as the one the participants saw in the instruction. We did not include the participants who failed to answer this quality check question, because we expected that careful crowdsourced workers could answer the question explained in the instruction. Additionally, dropped those who spent too little (bottom 10%) or too much (top 10%) time to complete the tasks.

The total number of observations was 535 (Word Intrusion), 900 (Vertical), and 620 (Horizontal).



Please use your best guess to group four items into two groups on the right. Each item and group is represented by four words.

**Items**
game dvd film sony
album year award best
music band song year
tax said budget party

**[Group 1] film best award year**

**[Group 2] will policy people law**

Figure 13: An example task.

## H.3 An Ethics Review

An institutional review board of an author's institution reviewed our experimental design.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Yes, we discussed the limitations of our work in Section 9 (after the conclusion section). The limitations are particularly related to the scope of our claims.*

□ A2. Did you discuss any potential risks of your work?
*Not applicable. Our paper makes contributions to the machine learning theory and does not meet any of the criteria mentioned in the checklist.*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*Section 1 is the introduction. The abstract comes before the introduction.*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☒ Did you use or create scientific artifacts?

*Left blank.*

□ B1. Did you cite the creators of artifacts you used?
*Not applicable. Left blank.*

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*Not applicable. Left blank.*

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*Not applicable. Left blank.*

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*Not applicable. Left blank.*

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Not applicable. Left blank.*

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Not applicable. Left blank.*

### C ☑ Did you run computational experiments?

*Section 6 compares our model against existing models.*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Since our model is not a neural model, our theory section (Section 4) enumerates all parameters; hence, the number of parameters is evident from the paper. Also, we did not use GPU at all. Section 6.2 describes the cluster computer (only CPU) we used.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Section 6.2 explains our experimental setup.*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*Section 6 describes our results. Figure 5 illustrates the result with error bars. Table 2 reports the means of evaluation metrics for three different numbers of top words. Section 6.3 explains how we took the means.*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*We used the existing implementations of machine learning models for evaluation in Section 6. We listed the packages we used in Section 6.2 and described the parameter settings (we used the default parameter values).*

**D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*We used crowdworkers to evaluate models in Section 6.4.*

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*Appendix H has a screenshot of the consent form and an example task. Our experiment did not show offensive content.*

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*We mentioned the crowdsourcing platform (Amazon Mechanical Turk) in Section 6.4 and Appendix H. We mentioned the compensation in Appendix H.*

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*We included a screenshot of the consent form in Appendix H.*

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Sections 6.4 and H.3*

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*We could not retrieve such information from Amazon Mechanical Turk.*