11-2021

Topic modeling for multi-aspect listwise comparison

Delvin Ce ZHANG
Singapore Management University, cezhang.2018@smu.edu.sg

Hady W. LAUW
Singapore Management University, hadywlauw@smu.edu.sg

Topic Modeling for Multi-Aspect Listwise Comparisons

Delvin Ce Zhang
Singapore Management University
Singapore
cezhang.2018@smu.edu.sg

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

ABSTRACT

As a well-established probabilistic method, topic models seek to uncover latent semantics from plain text. In addition to having textual content, we observe that documents are usually compared in listwise rankings based on their content. For instance, world-wide countries are compared in an international ranking in terms of electricity production based on their national reports. Such document comparisons constitute additional information that reveal documents' relative similarities. Incorporating them into topic modeling could yield comparative topics that help to differentiate and rank documents. Furthermore, based on different comparison criteria, the observed document comparisons usually cover multiple aspects, each expressing a distinct ranked list. For example, a country may be ranked higher in terms of electricity production, but fall behind others in terms of life expectancy or government budget. Each comparison criterion, or aspect, observes a distinct ranking. Considering such multiple aspects of comparisons based on different ranking criteria allows us to derive one set of topics that inform heterogeneous document similarities. We propose a generative topic model aimed at learning topics that are well aligned to multi-aspect listwise comparisons. Experiments on public datasets demonstrate the advantage of the proposed method in jointly modeling topics and ranked lists against baselines comprehensively.

CCS CONCEPTS

• **Information systems** → **Content ranking**; **Document topic models**; **Learning to rank**;

KEYWORDS

Generative Topic Model; Text Mining; Comparative Documents

ACM Reference Format:
Delvin Ce Zhang and Hady W. Lauw. 2021. Topic Modeling for Multi-Aspect Listwise Comparisons. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3459637.3482398

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482398

1 INTRODUCTION

Topic models provide a statistical framework for discovering latent "topics" that occur in a text corpus. Conventional topic models, such as LDA [3], learn topics in an unsupervised way based on word co-occurrences alone. When a cluster of words tend to co-occur in a significant subset of documents, these words likely convey a topic. Documents vary in the mixtures of topics they represent. In turn, a topic is characterized by a probability distribution over representative words. This fundamental modeling is elegant in its simplicity and generality, lending itself to use cases such as corpus exploration, dimensionality reduction, etc.

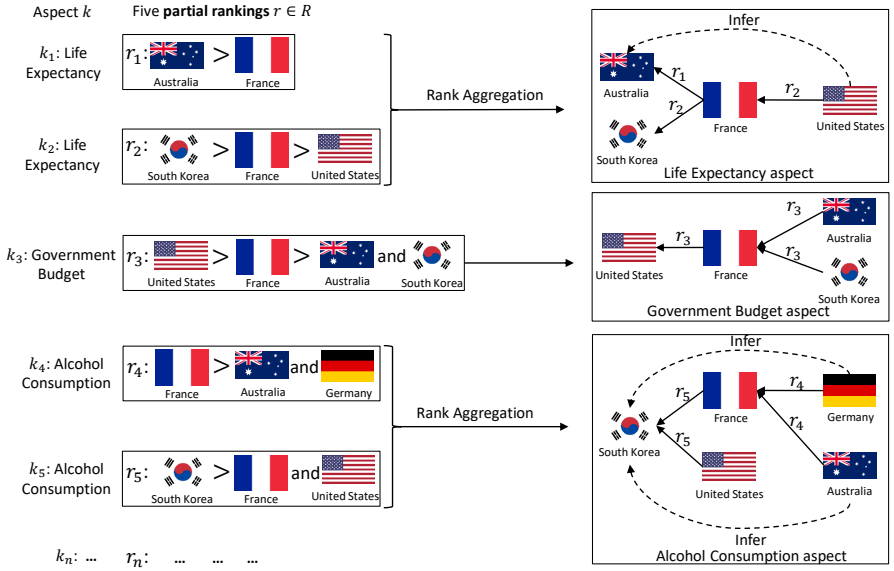
However, text documents – Web pages, academic papers, product reviews, etc. – are highly variable in their use of words. The presumption of self-sufficiency of word co-occurrences in discovering topics may not always bear out in practice. Assistively, supplementing the text, there may be further clues on the meanings contained in a document. They may come in various structures, such as categorical labels that serve as document classes [30], links between pairs of documents [10], user preferences [45], etc.

In this work, we seek to model a text corpus with the ancillary structure of *listwise partial rankings* involving subsets of documents. Consider a corpus involving entities of a domain. For a concrete instantiation, we allude to Wikipedia articles of countries in the world. A topic model would ostensibly discover topics that characterize a country vis-á-vis other countries, such as the nature of its economy, the outcomes of its healthcare policies, etc., based on the words used to describe these characterizations. While those may well manifest in words to some extent, we observe that there are informative structures, such as international rankings[1] that compare countries along *multiple aspects* such as life expectancy, government budget, alcohol consumption, etc.

Incorporating such rankings into topic modeling could yield comparative topics that help to differentiate and rank documents along certain aspects. Indicatively, two countries may share similar topics if they are ranked higher (resp. lower) than similar subset of other reference countries. For example, countries with significant agricultural sectors may be ranked similarly in terms of productions of various crops. Similarly, countries with larger industrial bases may be ranked similarly in terms of electricity consumption. For another instance, we may discover topics that help shine light on how a country's alcohol consumption is higher than others.

Given an aspect, its various listwise partial rankings could be aggregated into a more complete ranking for that aspect. Fig. 1 illustrates this concept: (a) shows five input listwise partial rankings r_1 to r_5; (b) shows how rankings that share the same aspect such as r_1 and r_2 and respectively r_4 and r_5 could each be aggregated

[1] https://en.wikipedia.org/wiki/List_of_international_rankings



(a) Illustration of **multi-aspect partial rankings** over a subset of documents.

(b) Illustration of **rank aggregation** and **transitive inference** within each aspect.

Figure 1: Illustration of (a) multi-aspect partial rankings; (b) rank aggregation *within* each aspect.

per aspect. One application of exploring such a corpus with comparisons is to help government officials compare at which level a country's alcohol consumption is ranked among others.

As another application, online users may purchase products and leave reviews. Some products satisfy users more than others. Making a comparison of product reviews can reveal insights into what factors of products can satisfy consumers more than others.

Problem. We formulate topic modeling for multi-aspect listwise comparison as follows. The inputs are two sets: one is a corpus of documents D, as in previous topic modeling; the other is a set of multi-aspect partial rankings R, as shown by Fig. 1(a). Each instance of partial ranking $r \in R$ comes from one aspect k, and is a listwise comparison over a subset of documents $D' \subseteq D$. We aim to design a topic model that can capture both sets of inputs to derive comparative topics that help to compare and differentiate documents along multiple aspects. We obtain two sets of outputs correspondingly: one is document-topic β and topic-word distribution θ for topic modeling; the other is aspect-specific ranking parameter u_k that determines the comparison outcome of documents along multiple aspects. We would use *ranking* and *comparison* interchangeably.

We assume only partial, instead of full, rankings over documents in D as it is the less restrictive assumption. There are often rankings observed only over a subset, e.g., Europe, ASEAN, etc. Moreover, we consider relative comparisons rather than assuming absolute labels. Usually, relative comparisons are simpler to obtain than absolute labels. For example, it is easier for government officials to compare several countries and figure out which one has higher alcohol consumption than to ask them to score each country. Even though absolute labels are available, considering relative comparisons is still meaningful. Comparing a set of documents yields comparative topics that help government officials understand why a country has higher alcohol consumption than others, while predicting only the individual consumption value of each country without cross comparison cannot discover such comparative topics. Though we are using countries as an example corpus, the formulation applies to many other corpora such as peer reviews of academic papers along dimensions such as impact, originality, etc.; reviews of companies as ranked according to values and benefits; reviews of hotels on aspects such as cleanliness, location, and service.

Approach. We design a generative topic model MALIC, for **M**ulti-**A**spect **Li**stwise **C**omparisons. Broadly speaking, the principles underlying MALIC are two-pronged. First, from a *practical view*, we aim at a flexible model that takes as input a set of partial rankings of various lengths. For example, the comparison between France and Australia in Fig. 1(a) is pairwise, while three or more documents constitute a listwise comparison. Second, from a *summative view*, the observed partial rankings should support transitive inference, which discovers unobserved comparison outcomes by aggregation across multiple rankings of an aspect. Illustrated by Fig. 1(b), the aggregation of the first two rankings infers the comparison between Australia and US in life expectancy. This is achieved by the comparison graph for each aspect. Each edge points from lower-ranked document to the higher-ranked. An unobserved comparison is inferred if there is a directed path between them, and the edges on the path are not from the same partial ranking.

Contributions. We make the following contributions. *First*, we propose MALIC. By designing a topic-regression ranking probability component, MALIC flexibly accommodates different structures and lengths of partial rankings, and supports transitive inference. *Second*, we further extend it to a ranking mixture topic model to support multiple aspects. *Third*, through extensive experiments, we show that MALIC outperforms baselines on several tasks.

2 RELATED WORK

Unsupervised Topic Models. Previously, unsupervised topic models are generative models, e.g., PLSA [18], LDA [3], SAM [36]. More recent models are based on neural approaches. Auto-Encoder (AE) is a class of such neural methods. Its variants include CAE [37], DAE [41], VAE [21], Sparse AE [11, 27], etc. There are also methods that improve LDA by using similar inference method with VAE's, including ProdLDA [38], DVAE [7], ETM [12], etc. However, they learn topics based on word co-occurrences in an unsupervised way and cannot leverage listwise document comparisons. For the latter, one needs to pipeline their topics by a learning-to-rank method. In contrast, MALIC joints topic modeling and multi-aspect comparisons. Therefore, constraining the topic model to align with the observed multi-aspect comparisons may yield topics that help to compare documents along certain aspects, or comparison criteria.

Supervised Topic Models. In order to incorporate additional information within and across documents and derive more informative topics, extensions of LDA that model various metadata are proposed. *Pointwise* supervised topic models require a response variable for each document. sLDA [30] is designed with a regression component and supervised by numerical values. Other models, such as DiscLDA [22], modify topic distributions for categorical label supervision. LabeledLDA [34] is proposed for multi-label documents, while PLDA [35] is for partially labeled documents. MedLDA [52] integrates the max-margin concept into supervised topic models. As it relies on explicit response variable for each document, sLDA is not appropriate for the problem at hand. We will pipeline it to show the advantage of listwise over pointwise supervision. Models with categorical label as supervision are not comparable, since we cannot transform partial rankings into categorical labels.

Another class of supervised topic models is *pairwise* supervised models. The closest such work is CompareLDA [40] for single-aspect pairwise document comparisons. However, its single-aspect modeling and pairwise comparison format are not flexible enough to incorporate different structures and variable lengths of partial rankings, and cannot derive one set of topics for multiple aspects of comparisons. We will compare against it in the experiments to highlight the advantage of jointly modeling multiple aspects. Meanwhile, [10, 49, 50] model a pair of linked documents, but they consider a pair of documents as sharing similar topics, not comparatively "winner" or "loser".

Learning to Rank. We position our work as a joint topic modeling and learning-to-rank modeling [25]. Learning-to-rank, as its name reveals, represents a class of methods that learns a permutation, or ranking sequence, over a set of items. Bradley-Terry-Luce (BTL) model [5, 26] is proposed for *pairwise* item comparison, and CompareLDA is a topic model built on top of BTL. In this paper, we build on Plackett-Luce model [26, 32], which expresses a probability distribution of complete rankings in terms of item-specific utility scores, but does not support multiple aspects or the inference of a previously unseen item. There are several works aiming to design parameter estimation of Plackett-Luce model, including EM algorithm [39], Bayesian approach [16], Generalized Method of Moments [1], etc. Beyond a single ranking model, there are works exploring a mixture of Plackett-Luce models. [9] describes a Dirichlet process mixtures, [15] applies Plackett-Luce mixture to analyze Irish electorates, [51] proves the identifiability of Plackett-Luce mixture, etc. These models consider rankings alone without item features. Other models include Mallows model [28], SVM-Rank [19], RankNet [6], ListNet [8], etc. These learning-to-rank methods do not have topic component, while topic models discussed above are not appropriate for multi-aspect listwise comparisons.

Dealing with partial rankings for aggregation and inference is well studied in literature [1, 51], but not in the topic modeling context. Its real-word applications include meta-search [13, 43] that combines results from multiple search engines, and preference aggregation [44] that integrates preference of users. In this paper, we seek to joint topic modeling and partial document comparisons.

Others. There are also more distantly related existing works. *Aspect extraction* [17] seeks to extract entities (which may be a word) on which opinions have been expressed within a document and cluster those entities into different aspects. In contrast, we are dealing with documents as a whole, instead of entities within documents. Furthermore, the aspect we mention here belongs to partial rankings, not to entities. We also set apart from *preference learning* for personalized recommendation using topic models [29, 45], since we do not involve users here, and our task is not recommendations. *Comparative text mining* [31, 48] aims at discovering common topics across multiple collections of documents and collection-specific topics. We are different in that we are given only one collection of documents (e.g., country corpus), and the topics are not designed to capture any common or specific concepts. Modeling document comparisons is different from *sentiment analysis* [23]. Listwise rankings are expressed by a sequence of documents, while sentiment has two polarities, and does not have listwise rankings. Different sentences in a document may express different sentiments [33].

Table 1: Summary of notations.

Notation	Description
D	a corpus of documents
d	a specific document, $d \in D$
D'	a subset of documents, $D' \subseteq D$
R	a set of partial rankings of different structures and lengths
r^s	a specific partial ranking with structure s, $r^s \in R$
S	a set of partial ranking structures
s	a specific partial ranking structure, $s \in S$
v_d	positive utility score of document d
K	total number of aspects
k	the k^{th} aspect, $k = 1, 2, ..., K$
Z	total number of topics
β_d	document d's topic distribution, $\beta_d \in \mathbb{R}^Z$
β'_d	document d's unnormalized topic distribution, $\beta'_d \in \mathbb{R}^Z$
u_k	the k^{th} aspect's transformation parameter, $u_k \in \mathbb{R}^Z$
π	mixing coefficient, $\pi \in \mathbb{R}^K$, $\sum_{k=1}^K \pi_k = 1$
θ_z	word distribution of the z^{th} topic
N_d	total number of word occurrences in document d
$w_{d,n}$	the n^{th} word in document d, $n = 1, 2, ..., N_d$
W	vocabulary

3 MODEL ARCHITECTURE AND ANALYSIS

This section describes the proposed model MALIC, whose graphical model is shown by Fig. 2(b). Table 1 summarizes the notations.

Again, as input, we are given a corpus of documents D and a set of multi-aspect partial rankings R. Each specific partial ranking $r \in R$, of aspect k, is a listwise comparison over a subset of documents $D' \subseteq D$. As output, we derive document-topic distribution β, topic-word distribution θ, and rank parameter u_k, $k = 1, 2, ..., K$.

Generative Process. As an overview, the MALIC model is described by the following generative process.

(1) For each topic $z = 1, 2, ..., Z$, draw its distribution over words $\theta_z \sim Dirichlet(\alpha_1)$
(2) For each document $d \in D$:
  (a) Draw topic distribution $\beta'_d \sim \mathcal{N}(0, \sigma_1^{-1}\mathrm{I})$ and transform to $\beta_d = softmax(\beta'_d)$
  (b) For each word $w_{d,n}$ where $n = 1, 2, ..., N_d$:
    (i) Draw a topic $z \sim Categorical(\beta_d)$
    (ii) Draw a word $w_{d,n} \sim Categorical(\theta_z)$
(3) Draw an aspect distribution $\pi \sim Dirichlet(\alpha_2)$
(4) For each aspect $k = 1, 2, ..., K$:
  (a) Draw Z-dimensional parameter $u_k \sim \mathcal{N}(0, \sigma_2^{-1}\mathrm{I})$
(5) For each partial ranking $r \in R$:
  (a) Draw an aspect $k \sim Categorical(\pi)$
  (b) Draw a partial ranking $r \sim \mathrm{PL}(r|u_k, \{\beta_d\}_{d \in r})$

The first two steps concern content generation. Differently from LDA [3], Step 2 uses a softmax $\beta_{d,z} = \frac{\exp(\beta'_{d,z})}{\sum_{z'=1}^{Z} \exp(\beta'_{d,z'})}$ for normalization, instead of Dirichlet. As we will see at parameter learning, the optimization of β'_d does not have a closed form and requires gradient descent. Drawing from a Dirichlet would add a sum-to-one constraint, leading to a more complex optimization.

The next three steps concern the ranking modeling. The key is how topic and ranking modeling interact via topic distributions β_d. This is borne out in the graphical model in Fig. 2(b), topic model on the left and ranking model on the right, linked by β'_d.

3.1 Topic-Regression Ranking

First, we illustrate how a document's topic distribution β'_d helps determine its rank position (Step 5b). We build on a ranking probability model based on Plackett-Luce [26, 32]. A ranking r is a comparison over items. Plackett-Luce (PL) defines a probability distribution over all possible rankings of the same set of items D. It associates each item $d \in D$ with a positive utility score $v_d > 0$. The higher the value of v_d, the more likely d would be ranked on top. Formally, the probability of a specific ranking $r = [d_{i_1} > d_{i_2} > ... > d_{i_{|D|}}]$ is

$$\mathrm{PL}(r|V) = \prod_{l=1}^{|D|} p_l(r|V) \quad \text{where} \quad p_l(r|V) = \frac{v_{i_l}}{\sum_{q=l}^{|D|} v_{i_q}}. \tag{1}$$

$V = \{v_{i_l}\}_{l=1}^{|D|}$ are learnable parameters. We interpret Eq. 1 as follows. We rank items from the first position to the last one in sequence. At the beginning, all items are candidates, thus the probability of any item ranked at the top-1 place is $p_1(r|V)$. The higher the utility score, the more probable the corresponding item is selected. Having ranked the first one, we select the second-place item from the remaining $|D|-1$ candidates by $p_2(r|V)$. We repeat this process until the rank list r is finished and yield $\mathrm{PL}(r|V)$ as a joint probability.

A limitation of the basic PL is that it assumes all items are observed beforehand. It cannot infer utility score for unseen items, as learnable parameters V are item-specific. One solution is to bring learnable parameters to the feature space. To rank documents, we express utility score v_d by a topic regression function using d's topic distribution $\beta_d \in \mathbb{R}^Z$ and model-specific parameter $u \in \mathbb{R}^Z$ by $v_d = \exp(u^T \beta_d)$, Z is the number of topics. We use exponential to ensure a positive value. Two documents with similar topics would present similar score v_d, thus are ranked at similar positions.

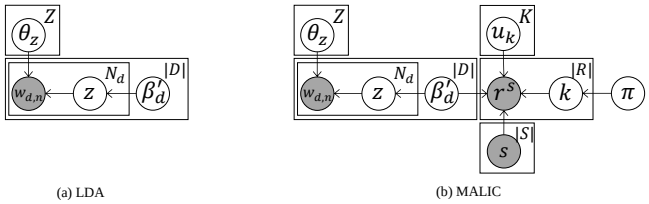


Figure 2: Graphical model of (a) LDA and (b) MALIC.

3.2 Partial Ranking Structures

In the exposition so far, presumably r at Eq. 1 is a full ranking over the entire set D, from the first item to the last. However, as mentioned in the Introduction section, here we assume *partial rankings of various lengths and structures* over a subset of documents. From a practical view, we thus propose partial rankings over a subset of items $D' \subseteq D$. As illustrated by Fig. 1(a), different partial rankings present different structures. The first two r_1 and r_2 are strict comparisons with different lengths, while the next three r_3, r_4, and r_5 allow unranked documents. Based on the intuition of ranking process of Eq. 1, by appropriate redefinition, we model distributions for partial ranking structures below. We associate each partial ranking r^s with structure $s \in S$. For clarity and illustration purpose, we propose three common structures (others can be similarly defined).

(1) L-way Partial Ranking. We rank a subset of items D' where $|D'| = L$. $r^{L-way} = [d_{i_1} > d_{i_2} > ... > d_{i_L}]$.

$$\mathrm{PL}(r^{L-way}|V) = \prod_{l=1}^{L} \frac{v_{i_l}}{\sum_{q=l}^{L} v_{i_q}}. \tag{2}$$

(2) Top-L Partial Ranking. We rank top-L items within a subset D' where $|D'| > L$, other items are unranked. $r = [d_{i_1} > d_{i_2} > ... > d_{i_L} > D' \backslash \{d_{i_1}, ..., d_{i_L}\}]$.

$$\mathrm{PL}(r^{Top-L}|V) = \prod_{l=1}^{L} \frac{v_{i_l}}{\sum_{q=l}^{|D'|} v_{i_q}}. \tag{3}$$

(3) Choice-L Partial Ranking. We select the best item within a subset D' where $|D'| = L > 1$, and other items are unranked. $r = [d_{i_1} > D' \backslash \{d_{i_1}\}]$.

$$\mathrm{PL}(r^{Choice-L}|V) = \frac{v_{i_1}}{\sum_{q=1}^{L} v_{i_q}}. \tag{4}$$

In a nutshell, L-way is a strict comparison over a subset D', corresponding to r_1 and r_2 at Fig. 1(a) with 2-way and 3-way, respectively; top-L allows unranked or no-difference items, and r_3 is top-2 with length of 4 documents; choice-L is one-step selection, which is the format of r_4 and r_5 with choice-3. These structures do not limit

comparisons into any strict format. Pairwise is a special case of 2-way or choice-2. Step 5b draws a partial ranking using Eq. 2–4.

Definitions above obey transitive closure shown at Fig. 1(b). Given two partial rankings r_1 and r_2 at Fig. 1(a) as an example, the maximization of log-likelihood of r_1, i.e., $\log \frac{d_A}{d_A+d_F} = \log \frac{1}{1+d_F/d_A}$, pushes Australia's utility score d_A higher than France's d_F. Meanwhile, the log-likelihood of r_2 is the sum of two terms, $\log \frac{d_S}{d_S+d_F+d_U}$ and $\log \frac{d_F}{d_F+d_U}$. r_2's maximization pushes d_S higher than d_F and d_U (the first term), and d_F higher than d_U (the second term). As reflected by the utility scores, $d_A > d_F > d_U$, transitive closure is preserved. More generally, the definition of Plackett-Luce at Eq. 1 ranks items from the first to the last in sequence. Items with higher utility scores are more likely to be ranked higher than others. Since transitivity is reflected by utility scores, Plackett-Luce model as well as proposed partial rankings Eq. 2–4 preserve such properties.

Theorem 1. *Based on the definitions of partial ranking above, we draw the following observations about their relationships.*

(1) *For the same set of items D' where $|D'| = L$, Top-1 ranking is equivalent to Choice-L ranking.*
(2) *For the same set of items D', if $|D'| = L+1$, Top-L ranking is equivalent to $(L+1)$-way ranking.*
(3) *L-way ranking is the recursive process of Choice-L ranking for all items in D'.*

Proof. (1) $\mathrm{PL}(r^{Top-1}|V) = \frac{v_{i_l}}{\sum_{q=l}^{|D'|} v_{i_q}} = \mathrm{PL}(r^{Choice-L}|V)$.

(2) $\mathrm{PL}(r^{Top-L}|V) = \prod_{l=1}^{L} \frac{v_{i_l}}{\sum_{q=l}^{|D'|} v_{i_q}} = \prod_{l=1}^{L} \frac{v_{i_l}}{\sum_{q=l}^{L+1} v_{i_q}} \times \frac{v_{i_{L+1}}}{v_{i_{L+1}}} = \prod_{l=1}^{L+1} \frac{v_{i_l}}{\sum_{q=l}^{L+1} v_{i_q}} = \mathrm{PL}(r^{(L+1)-way}|V)$.

(3) $\mathrm{PL}(r^{L-way}|V) = \prod_{l=1}^{L} \frac{v_{i_l}}{\sum_{q=l}^{L} v_{i_q}} = \prod_{l=1}^{L} \mathrm{PL}(r^{Choice-L}|\{v_{i_q}\}_{q=l}^{L})$. □

3.3 Multi-Aspect Mixture Model

Having introduced listwise partial rankings (Step 5b), we now turn to the concept of multiple aspects (Step 5a). Each PL regression model represents a distinct ranked list by utility scores of a particular aspect. To accommodate multiple aspects jointly, we propose a mixture with K aspects as a distribution for partial rankings.

$$\mathrm{PL}^{(K)}(r^s|\pi, V_1, ..., V_K) = \sum_{k=1}^{K} \pi_k \mathrm{PL}(r^s|V_k) \tag{5}$$

where $\sum_{k=1}^{K} \pi_k = 1$. $V_k = \{v_{k,d}\}_{d \in r^s} = \{\exp(u_k^T \beta_d)\}_{d \in r^s}$ is topic regression for every aspect k and every document d. When $K \geq 2$, two aspects with similar ground-truth rankings tend to present similar parameters u_k. During the generative process, one set of topics of a document is shared across multiple aspects, thus different aspects interact with each other through this shared parameter. Even though one document does not contain certain content of one aspect for comparison, it may present relevant content of another related aspect. Through their similar aspect-specific parameters u_k, these two aspects can complement useful information of that document for each other. Therefore, given observed partial rankings, different aspects collaboratively extract useful information for learning, and both ranking modeling and topic modeling can be improved. When $K = 1$, Eq. 5 degenerates to Eq. 2–4.

Summary. Having elaborated all three key designs in ranking modeling component, here we summarize them as a complete model. Step 3 draws the mixture of aspects π; Step 4 draws parameters for aspect-specific ranking u_k. Step 5 generates observed multi-aspect partial rankings. As topic model (Steps 1-2) and ranking model (Steps 3-5) interact by topic distribution β'_d, the generation of text forwards influences ranking modeling, the optimization of rankings backwards enhances topic model in a comparative way.

Unsupervised vs. Supervised. The generative process outlined above assumes a fully *unsupervised* setting where we do not observe the ground-truth aspects of partial rankings. We model such uncertainty by probabilistically drawing an aspect k at Step 5a. Our model also accommodates a *supervised* setting where we observe the ground-truth aspects k of a proportion of partial rankings. Thus Step 5a can be replaced with setting k to the ground-truth aspect k_{r^s}. We will test both settings in experiments.

4 PARAMETER LEARNING

Optimization is conducted by maximum a posteriori (MAP) with EM algorithm [2]. Log-likelihood is

$$\begin{aligned}\mathcal{L}(\Psi|D, R) =& \lambda \sum_{d \in D} \sum_{w_{d,n}} \log \sum_{z=1}^{Z} p(w_{d,n}|z, \theta_z) p(z|\beta_d) \\ &+ \sum_{s \in S} \sum_{r^s \in R} \log \sum_{k=1}^{K} \pi_k \mathrm{PL}(r^s|V_k).\end{aligned} \tag{6}$$

where λ is a hyperparameter, balancing the relative importance of two modeling components. Parameters to be inferred are $\Psi = \{\pi, u_k, \theta_z, \beta'_d\}$ for $k = 1, ..., K$, $z = 1, ..., Z$, and $d \in D$. Note that we optimize β'_d instead of β_d. The conditional expectation of the complete-data log-likelihood with priors is

$$\begin{aligned}Q(\Psi|\hat{\Psi}) =& \lambda \sum_{d \in D} \sum_{w_{d,n}} \sum_{z=1}^{Z} \gamma(z_{d,w_{d,n}}) \log p(w_{d,n}|z, \theta_z) p(z|\beta_d) \\ &+ \sum_{s \in S} \sum_{r^s \in R} \sum_{k=1}^{K} \gamma(k_{r^s}) \log \pi_k \mathrm{PL}(r^s|V_k) + \log p(\pi|\alpha_2) \\ &+ \sum_{k=1}^{K} p(u_k|\sigma_2) + \sum_{z=1}^{Z} p(\theta_z|\alpha_1) + \sum_{d \in D} p(\beta'_d|\sigma_1).\end{aligned} \tag{7}$$

where $\hat{\Psi}$ is current estimation.

E step. Optimization is conducted by repeating E step and M step until log-likelihood convergence. We first present E step.

$$\gamma(z_{d,w_{d,n}}) = p(z|d, w_{d,n}) = \frac{p(w_{d,n}|z, \theta_z) p(z|\beta_d)}{\sum_{z'=1}^{Z} p(w_{d,n}|z', \theta_{z'}) p(z'|\beta_d)}. \tag{8}$$

$$\gamma(k_{r^s}) = p(k|r^s) = \frac{\sum_{c=0,1} \pi_k \delta_k^c \mathrm{PL}^c(r^s|V_c)}{\sum_{k'=1}^{K} \sum_{c=0,1} \pi_{k'} \delta_{k'}^c \mathrm{PL}^c(r^s|V_c)}. \tag{9}$$

M step. After evaluating posterior probabilities at E step, we now update parameters $\Psi = \{\pi, u_k, \theta_z, \beta'_d\}$ at M step. π and θ_z have closed-form solution, and are updated by Eq. 10–11. To update

others, we use gradient-based numerical optimization method, such as Quasi-Newton method [24]. Gradients are evaluated at Eq. 12–13.

$$\pi_k = \frac{\sum_{s\in S}\sum_{r^s\in R}\gamma(k_{r^s}) + \alpha_2}{\sum_{k'=1}^{K}\sum_{s\in S}\sum_{r^s\in R}\gamma(k'_{r^s}) + K\alpha_2}. \tag{10}$$

$$\theta_{z,w} = \frac{\lambda\sum_{d\in D}\sum_{w_{d,n}}\mathbb{I}(w_{d,n}=w)\gamma(z_{d,w_{d,n}}) + \alpha_1}{\lambda\sum_{w\in W}\sum_{d\in D}\sum_{w_{d,n}}\mathbb{I}(w_{d,n}=w)\gamma(z_{d,w_{d,n}}) + |W|\alpha_1}. \tag{11}$$

$$\frac{\partial Q}{\partial u_{k,z}} = \sum_{s\in S}\sum_{r^s\in R} p(k|r^s)\frac{\partial \log \mathrm{PL}(r^s|V_k)}{\partial u_{k,z}} - \sigma_2 u_{k,z}. \tag{12}$$

$$\frac{\partial Q}{\partial \beta'_{d,z}} = \lambda\sum_{w_{d,n}}(\gamma(z_{d,w_{d,n}}) - \beta_{d,z}) + \sum_{s\in S}\sum_{r^s\in R}\mathbb{I}(d\in r^s)\sum_{k=1}^{K}\gamma(k_{r^s})\sum_{c=0,1}\gamma(c_{r^s,k})\frac{\partial \log \mathrm{PL}^c(r^s|V_c)}{\partial \beta'_{d,z}} - \sigma_1\beta'_{d,z}. \tag{13}$$

Complexity. For E step, we have $O(Z|D|\sum_{d\in D}N_d + |R|K + L_{\max}^2)$ for one iteration. $L_{\max}$ is the maximum length of partial rankings. Since L-way is the most complex one among the listed three, we assume all rankings are L-way as the worst case. For M step, we have $O(|W||D|(\sum_{d\in D}N_d+Z)+Z|D|K|R|_{\max}L_{\max}^2)$ for one iteration. $|W|$ is the size of vocabulary, and $|R|_{\max}$ is the maximum number of rankings containing the same document.

Brief Comment on Running Time. Since MALIC is the first topic model for multi-aspect listwise comparisons, our focus is on effectiveness, not efficiency. We just briefly report running time. The training takes less than 1h on small datasets, less than 10 hours on large dataset. Experiments were conducted on a machine with Intel Xeon E5-2650v4 2.20 GHz CPU and 256GB RAM.

Testing. We can use parameters for prediction at testing.

i) **Ranking Prediction**. Given an unseen test document $d' \in D_{test}$, we infer its topic distribution and utility score by

$$\beta_{d'} = p(z|d') = \frac{p(z,d')}{\sum_{z'=1}^{Z}p(z',d')} = \frac{\sum_{d\in D_{train}}\prod_{w_{d',n}}p(w_{d',n}|z)p(z|d)p(d)}{\sum_{z'=1}^{Z}\sum_{d\in D_{train}}\prod_{w_{d',n}}p(w_{d',n}|z')p(z'|d)p(d)} \tag{14}$$

$$v_{k,d'} = \exp(u_k^T\beta_{d'}) \quad k = 1,...,K. \tag{15}$$

$p(d) = \frac{1}{|D_{train}|}$. The position of d' in aspect k is determined by $v_{k,d'}$ in a descending order, greater value ranks higher.

ii) **Aspect Assignment**. Given an unseen partial ranking r^s, we predict its aspect by posterior probability Eq. 9.

5 EXPERIMENTS

The main objective is to evaluate the quality of topics derived by our model from multi-aspect listwise partial rankings.

Datasets. We rely on four public datasets as listed in Table 2. In addition Wikipedia's Country dataset, our model is applicable to review datasets that compare product reviews to reveal which factors/attributes better satisfy consumer preferences, such as prices, quality, and functionalities. Note that here we consider relative comparisons of *a set of* documents, instead of rating prediction.

- **Country.** Each document is a Wikipedia page of a country. We include *12 aspects* from Wikipedia's lists of international rankings, i.e., *life expectancy*, *net exports*, *alcohol consumption*, *wealth per adult*, *cigarette consumption*, *natural disaster risk*, *vehicles per capita*, *GDP*, *electricity production*, *obesity rate*, *irrigated land area*, and *government budget*.
- **Paper Review.** Each document is a paper review at a CS conference [20]. Each review judges *6 aspects* of a submission, including *meaningful comparison*, *originality*, *impact*, *substance*, *appropriateness*, and *clarity*. All 6 aspects are originally given integer ratings from 1 to 5, which we transform into partial listwise rankings.
- **Company Review.** Each review evaluates *5 aspects* of a company, they are *culture values*, *career opportunities*, *company benefit*, *work balance*, and *senior management*.
- **Hotel Review.** Each hotel review has *7 aspects*, including *check-in at front desk*, *value*, *cleanliness*, *location*, *service*, *business service*, and *rooms* [46, 47].

Table 2: Dataset statistics.

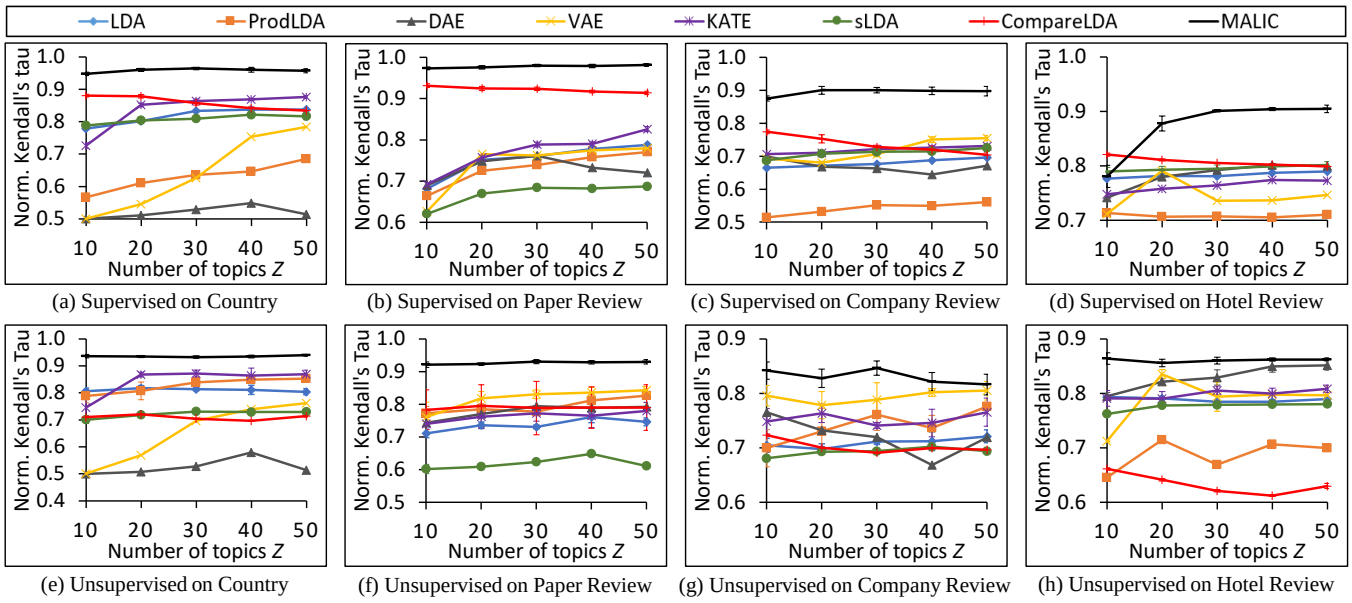
Name	#Documents	Vocabulary	#Aspects	#Partial Rankings	PCC Among Aspects
Country	312	2,920	12	5,400	0.38
Paper Review	1,104	2,761	6	2,700	0.20
Company Review	1,870	3,006	5	2,250	0.52
Hotel Review	53,507	5,017	7	110,250	0.62

After removing short documents, stop words, punctuations, we keep the most frequent 3,000 words for the first three datasets, and 5,000 for the large dataset. Table 2 presents the summary statistics. Pearson Correlation Coefficient (PCC) is the average correlation of rankings across all pairs of aspects. Higher means more correlated.

For each aspect, we consider 9 lengths of partial rankings, of 3 structures, i.e., {3, 4, 5}-way, top-{2, 3, 4}, and choice-{5, 10, 15}. We vary the length of rankings in Sec. 5.3. For each length, we randomly sample 50 partial rankings for the first three datasets, and 1,750 for Hotel Review dataset, since its corpus size is relatively 35 times larger than previous three. The full breakdown of these sampled rankings results in 80:20 for the number of pairwise comparisons to documents (this ratio is consistent with previous work). We investigate the effect of different number of sampled rankings in Section 5.3. Our model can incorporate more structures of partial rankings, but for clarity, here we enumerate these three.

Baselines. We compare against three categories of topic models.

(1) *Rank-agnostic.* Topic models including generative models such as LDA and ProdLDA, and neural Auto-Encoders such as DAE, VAE, KATE only use word co-occurrences without document labels for training. As rank-agnostic baselines cannot incorporate document comparisons, we pipeline their topics by a comparable learning-to-rank model PLRMM [39], whose aspect-awareness helps to outperform other learning-to-rank methods, e.g., RankNet [6], ListNet [8], etc. By comparing against these pipelined modeling, we highlight the advantage of jointly modeling topics and ranking.
There are supervised and unsupervised settings introduced in the discussion of Sec. 3. For supervised setting, we observe the ground-truth aspect of partial rankings during training. Unsupervised setting does not observe any aspect beforehand, and the model needs to infer the clustering of partial rankings. PLRMM can accommodate both settings.



(a) Supervised on Country (b) Supervised on Paper Review (c) Supervised on Company Review (d) Supervised on Hotel Review

(e) Unsupervised on Country (f) Unsupervised on Paper Review (g) Unsupervised on Company Review (h) Unsupervised on Hotel Review

Figure 3: Supervised and unsupervised partial ranking prediction with different number of topics.

(2) *Pointwise.* Following [40], we consider sLDA as the pointwise baseline. It requires PLRMM's preprocessing (either supervised or unsupervised) to convert listwise comparisons into pointwise scores before feeding to sLDA as supervision for each aspect. By comparison, we showcase the utility of listwise vs. pointwise supervision.

(3) *Pairwise.* Since MALIC incorporates multi-aspect listwise comparisons, our main baseline is the recent CompareLDA for single-aspect pairwise comparisons. Pairwise model CompareLDA treats single-aspect pairwise comparisons as input, thus we fully break down all partial rankings into pairwise, and apply CompareLDA to each aspect. By comparison, we show the importance of jointly learning multiple aspects vs. single aspect. In the unsupervised setting, where the aspects are not known, we need to use PLRMM to cluster partial rankings first, then apply CompareLDA to each cluster/aspect.

We choose hyperparameters based on validation set. We randomly split 80% documents for training, among them 10% are for validation. For DAE, Gaussian noise with 0.25 std.dev. generates the best results. For KATE, we set number of non-sparse neurons to 4, 6, 8, 10, 12 when $Z = 10, 20, 30, 40, 50$, respectively. For other baselines, we use their default hyperparameters. For MALIC, we set $\alpha_1 = \alpha_2 = 0.01$, $\sigma_1 = \sigma_2 = 0.01$, $\lambda = 0.01$. Each result is obtained by 5 independent runs, we report both average and standard deviation.

5.1 Ranking Evaluation

Since our model comprises two components, topic modeling and ranking modeling, we conduct experiments to evaluate each of them. We focus on ranking quality here and evaluate topic modeling next.

Partial Ranking Prediction. We expect a model to generalize well to unseen documents and accurately infer their comparisons. Following [40], we randomly hide 20% documents and their associated partial rankings. We only observe 80% documents and partial rankings among them for training. During testing, we infer utility scores of held-out documents by Eq. 14–15 and compare their inferred partial rankings against the previously hidden ground-truths.

Fig. 3 presents the results when varying number of topics. We use normalized Kendall's tau [14] (from 0 to 1) as metric (higher is better). We report std.dev. of MALIC and best-performing baselines. Some models perform stably, thus their error bars are not visible. CompareLDA performs well on supervised setting, but the results deteriorate in unsupervised setting. Compared to supervised setting where we explicitly apply CompareLDA to each aspect, this disjoint process increments the error from two separate components, thereby influencing the results. MALIC outperforms CompareLDA, due to modeling multiple aspects jointly. KATE also presents decent results, but is still worse than our model, demonstrating the importance of incorporating rankings for learning. Most models increase results before 30 topics, after which some keep flat while others deteriorate, we fix 30 topics for following experiments.

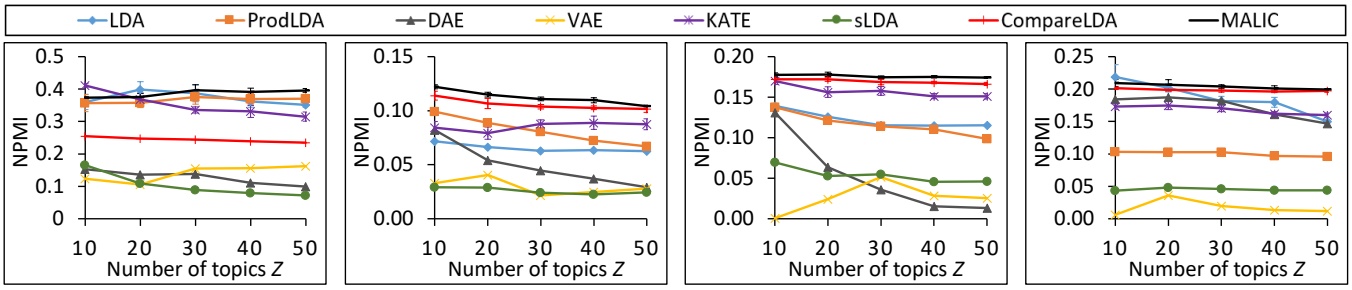
Rank Aggregation. As mentioned in Section 1, a good model should well aggregate partial rankings for transitive inference. For evaluation, we input all documents and partial rankings from multiple aspects. The goal is to test how well we aggregate observed rankings for transitivity, while partial ranking prediction above aims to test the generalization ability. Table 3 presents the normalized Kendall's tau at $Z = 30$ on both supervised and unsupervised settings. MALIC consistently outperforms baselines. Among baselines, CompareLDA and sLDA perform better, verifying the advantage of using rankings to learn aspect-oriented topics compared to unsupervised baselines. MALIC still presents better results than CompareLDA and sLDA, which demonstrates that the transitivity can be captured by the proposed ranking mixture model.

Table 3: Supervised and unsupervised experiments on rank aggregation at $Z = 30$ (results are in percentage).

Model	Supervised				Unsupervised			
	Country	Paper Review	Company Review	Hotel Review	Country	Paper Review	Company Review	Hotel Review
LDA	78.20±0.01	58.93±0.00	66.95±0.00	73.53±0.00	66.69±1.20	53.82±0.51	63.92±0.66	72.48±0.27
ProdLDA	60.49±0.06	56.09±0.04	54.77±0.04	50.29±0.33	51.63±0.29	51.63±1.00	51.16±0.15	50.07±0.04
DAE	57.73±0.01	59.78±0.05	66.48±0.01	74.20±0.01	52.05±1.71	55.43±1.55	61.68±1.60	66.26±1.43
VAE	61.20±0.00	61.44±0.05	70.62±0.01	69.06±0.00	52.81±3.58	53.16±0.02	60.80±1.19	64.69±3.13
KATE	81.84±0.01	61.83±0.02	72.07±0.00	72.30±0.00	62.94±1.30	58.06±1.20	68.10±0.58	69.39±0.23
sLDA	75.92±0.33	64.12±0.68	70.27±0.38	75.61±0.13	68.62±0.31	54.16±0.65	69.28±0.31	74.43±0.08
CompareLDA	82.70±0.21	80.87±0.84	74.03±0.44	72.19±0.08	58.54±1.19	55.56±1.74	58.73±0.66	58.52±0.09
MALIC	**95.26±0.08**	**84.63±0.45**	**86.26±0.41**	**83.20±0.25**	**73.11±0.09**	**68.22±0.65**	**75.11±0.45**	**79.82±1.07**

Table 4: Supervised and unsupervised experiments on aspect assignment at $Z = 30$ (results are in percentage).

Model	Supervised				Unsupervised			
	Country	Paper Review	Company Review	Hotel Review	Country	Paper Review	Company Review	Hotel Review
LDA	30.30±0.37	53.08±1.98	28.84±1.65	23.25±0.13	15.26±0.12	21.92±0.21	2.60±0.67	2.57±0.10
ProdLDA	17.05±0.86	49.85±2.26	20.89±1.28	14.35±0.11	5.17±0.46	18.96±2.07	1.41±0.30	0.05±0.00
DAE	10.04±1.45	50.41±2.14	26.13±1.31	23.78±0.13	8.25±1.21	20.11±2.12	2.19±0.43	2.85±0.33
VAE	11.45±1.00	44.55±0.42	26.29±2.59	18.93±0.18	5.76±0.75	17.25±1.32	2.56±0.50	2.58±0.11
KATE	32.41±0.67	51.70±1.24	31.16±1.31	23.33±0.04	21.09±1.24	21.28±1.40	3.69±1.06	2.49±0.03
sLDA	30.56±0.71	57.54±1.40	28.76±1.60	24.52±0.26	16.77±0.63	27.39±1.85	2.85±0.77	3.08±0.09
CompareLDA	21.07±1.17	60.65±1.71	29.55±1.77	24.07±0.21	10.73±0.67	31.30±1.93	5.04±1.37	5.44±0.11
MALIC	**52.83±1.87**	**85.85±1.80**	**52.89±3.75**	**35.72±0.93**	**36.09±1.50**	**68.39±2.87**	**18.80±4.23**	**9.24±0.72**



(a) Country (b) Paper Review (c) Company Review (d) Hotel Review

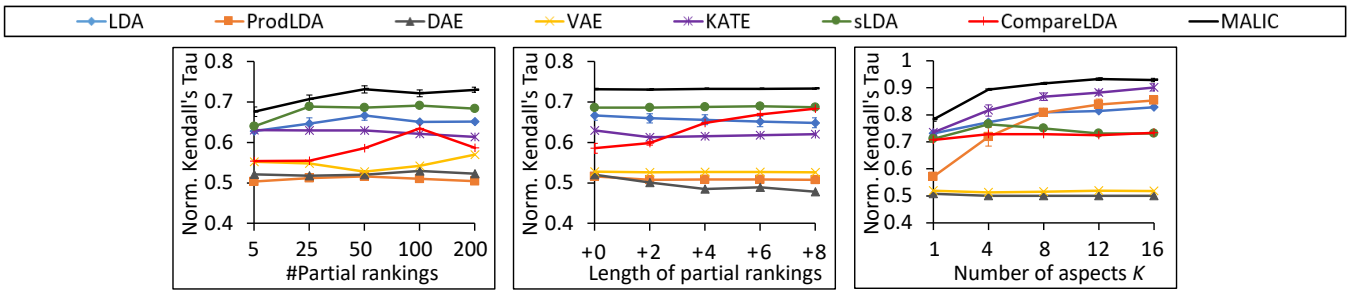
Figure 4: Topic coherence, Normalized Pairwise Mutual Information (NPMI), with different number of topics.

Aspect Assignment. Given a previously unseen partial ranking, we could predict its aspect. We split dataset the same way as for partial ranking prediction. After convergence, we use posterior probability Eq. 9 to predict aspects. Since supervised setting observes aspects of training rankings, and aspects of testing rankings are hidden, this task becomes partial ranking classification. We report classification accuracy over testing rankings as metric. On the contrary, unsupervised setting does not observe aspects of any rankings, including both training and testing set, this task becomes partial ranking clustering. We use Normalized Mutual Information (NMI) [42] for clustering evaluation. Table 4 presents the results at $Z = 30$. KATE, sLDA, and CompareLDA tend to outperform other baselines. Results, especially NMI, on Company Review and Hotel Review are lower than the other two datasets for all models, since their rankings of different aspects are highly correlated (see PCC at Table 2), a partial ranking is likely observed by more than one aspect, making aspect assignment more difficult. But overall, MALIC still assigns correct aspects to partial rankings more accurately than baselines on both settings, due to its multi-aspect modeling.

5.2 Topic Analysis

To see if MALIC's gain in ranking quality is at the expense of topic modeling, we evaluate topic coherence and perplexity. Since we do not observe significant difference between supervised and unsupervised settings, for clarity, we report unsupervised results.

Topic Coherence. Topic-word distribution $\theta \in \mathbb{R}^{Z \times W}$ indicates the keywords of each topic. Each row of θ_z corresponds to one topic, and its keywords are those with highest probability on that row. We select top-10 words of each topic, and use Normalized PMI (NPMI) [4] for evaluation of word pair associations. Fig. 4 presents the results when varying topics. MALIC performs better than baselines most of the time, indicating that modeling multi-aspect rankings does not hurt, and tends to improve the topic modeling. Since multi-aspect rankings provide additional information on document



(a) Different number of partial rankings (b) Different length of partial rankings (c) Different number of aspects

Figure 5: Model analysis on Country dataset.

Table 5: Top-5 keywords of 5 randomly selected aspects.

Aspect	Keywords of aspect's most related topic
natural disaster risk	saudi, puerto, arabia, oil, rico
alcohol consumption	country, lithuania, hungary, european, national
GDP	world, states, china, european, largest
wealth per adult	european, netherlands, union, luxembourg, austria
life expectancy	world, south, korea, country, china

Table 6: Perplexity of test documents, smaller is better.

Model	Country	Paper Review	Company Review	Hotel Review
LDA	9.630±0.115	8.189±0.071	7.583±0.066	7.591±0.014
ProdLDA	7.904±0.002	7.244±0.003	7.133±0.003	8.341±0.000
DAE	14.567±0.322	18.545±0.041	19.654±0.089	16.134±0.070
VAE	58.955±0.712	20.905±0.144	21.917±0.180	19.320±0.018
KATE	15.744±1.236	11.166±0.012	13.436±0.074	46.720±0.019
sLDA	8.237±0.010	7.883±0.019	7.926±0.011	8.314±0.007
CompareLDA	13.114±0.003	14.161±0.006	12.665±0.008	9.637±0.002
MALIC	**7.576±0.002**	**7.225±0.001**	**7.125±0.001**	**7.258±0.000**

relative similarities, modeling them enhances topic modeling quality. CompareLDA generally extracts more coherent topics than other baselines, verifying the effectiveness of modeling rankings.

To develop an intuitive sense of what the learned topics capture, we randomly select 5 aspects and present the most positive topic in their parameter u_k. Table 5 shows Country dataset. Alcohol consumption aspect tends to discuss European countries (Lithuania, Hungary). GDP reveals US, EU, and China. Modeling multiple aspects can help gather countries that are similar across aspects.

Perplexity. A topic model should generalize well to test documents. For evaluation, we split the dataset the same as for partial ranking prediction. We evaluate perplexity [3], $\exp\{-\frac{\log p(D_{test})}{\sum_{d' \in D_{test}} N_{d'}}\}$, of the held-out 20% documents. Since perplexity is exponential and varies much w.r.t. its power, we instead report the power $-\frac{\log p(D_{test})}{\sum_{d' \in D_{test}} N_{d'}}$ for clarity (smaller is better). Table 6 reveals that MALIC provides high likelihood to test documents, which we attribute to modeling multi-aspect rankings, since they bring additional information on relative document similarities.

5.3 Model Analysis

To better explore the sensitivity of models on different scenarios, we now conduct several analyses.

Different Numbers of Partial Rankings. In above experiments, we fix 50 and 1,750 partial rankings for each length and each aspect. To test the effect of partial ranking densities, we vary the number of partial rankings from sparse to dense. Fig. 5(a) shows unsupervised rank aggregation. Since each aspect has 9 lengths of partial rankings, of three structures, horizontal axis represents different number of partial rankings for each length and each aspect. MALIC performs well even when a small set of rankings is available. When the number of partial rankings increases, most models improve their results, since they obtain more useful information for training. After 100 rankings, most models reach the limit of performance and start to keep flat. The full breakdown of 200 rankings already results in more pairwise comparisons than what the ground-truth rank list contains. With such a dense observation on comparisons, MALIC still outperforms others, highlighting its ability to model document comparisons better than baselines.

Different Lengths of Partial Rankings. In above experiments, we select $\{3, 4, 5\}$-way, top-$\{2, 3, 4\}$, and choice-$\{5, 10, 15\}$ as the 9 lengths of partial rankings. To explore how models perform w.r.t. different lengths, we vary the length from short to long. Results on unsupervised rank aggregation are at Fig. 5(b). +2 at horizontal axis means we increase the length of each ranking by 2 documents, i.e., $\{3+2, 4+2, 5+2\}$-way, top-$\{2+2, 3+2, 4+2\}$, and choice-$\{5+2, 10+2, 15+2\}$. Longer rankings provide more information on document similarities, boosting results for sLDA and CompareLDA, while others keep flat, since they reach the limit of performance.

Different Numbers of Aspects. Modeling multiple aspects of rankings is one key design of our model. To look into how MALIC benefits from multiple-aspect modeling, we vary the number of aspects for partial ranking prediction. Fig. 5(c) summarizes the results. Similarly, when the number of aspects increases, both MALIC and baselines observe an improved performance, since these models indeed leverage multiple aspects to learn heterogeneous rankings. Compared to the extreme case of one aspect where all rankings are considered coming from the same aspect, MALIC benefits from modeling multiple aspects and provides a better result.

6 CONCLUSION

We propose a topic model for multi-aspect listwise comparisons. By designing topic-regression ranking mixture, MALIC incorporates multiple structures of partial rankings from different aspects. Experiments demonstrate the effectiveness of MALIC.

REFERENCES

[1] Hossein Azari Soufiani, William Ziwei Chen, David C Parkes, and Lirong Xia. Generalized method-of-moments for rank aggregation. In *Advances in Neural Information Processing Systems*. Neural Information Processing Systems Foundation, Inc., 2013.

[2] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[4] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40, 2009.

[5] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[6] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

[7] Sophie Burkhardt and Stefan Kramer. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *Journal of Machine Learning Research*, 20(131):1–27, 2019.

[8] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136, 2007.

[9] Francois Caron, Yee Whye Teh, Thomas Brendan Murphy, et al. Bayesian nonparametric plackett–luce models for the analysis of preferences for college degree programmes. *Annals of Applied Statistics*, 8(2):1145–1181, 2014.

[10] Jonathan Chang and David Blei. Relational topic models for document networks. In *Artificial Intelligence and Statistics*, pages 81–88, 2009.

[11] Yu Chen and Mohammed J Zaki. Kate: K-competitive autoencoder for text. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 85–94, 2017.

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

[13] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622, 2001.

[14] Ronald Fagin, Ravi Kumar, and Dakshinamurthi Sivakumar. Comparing top k lists. *SIAM Journal on discrete mathematics*, 17(1):134–160, 2003.

[15] Isobel Claire Gormley and Thomas Brendan Murphy. Exploring voting blocs within the irish electorate: A mixture modeling approach. *Journal of the American Statistical Association*, 103(483):1014–1027, 2008.

[16] John Guiver and Edward Snelson. Bayesian inference for plackett-luce ranking models. In *proceedings of the 26th annual international conference on machine learning*, pages 377–384, 2009.

[17] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397, 2017.

[18] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

[19] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, 2002.

[20] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635*, 2018.

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[22] Simon Lacoste-Julien, Fei Sha, and Michael Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. *Advances in neural information processing systems*, 21:897–904, 2008.

[23] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384, 2009.

[24] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[25] Tie-Yan Liu. *Learning to rank for information retrieval*. Springer Science & Business Media, 2011.

[26] R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2012.

[27] Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

[28] Colin L Mallows. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130, 1957.

[29] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172, 2013.

[30] Jon Mcauliffe and David Blei. Supervised topic models. *Advances in neural information processing systems*, 20:121–128, 2007.

[31] Michael Paul. Cross-collection topic models: Automatically comparing and contrasting text. *Urbana*, 51:61801, 2009.

[32] Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

[33] Md Mustafizur Rahman and Hongning Wang. Hidden topic sentiment model. In *Proceedings of the 25th International Conference on World Wide Web*, pages 155–165, 2016.

[34] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 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, 2009.

[35] Daniel Ramage, Christopher D Manning, and Susan Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–465, 2011.

[36] Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J Mooney. Spherical topic models. In *ICML*, 2010.

[37] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Icml*, 2011.

[38] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017.

[39] Maksim Tkachenko and Hady W Lauw. Plackett-luce regression mixture model for heterogeneous rankings. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 237–246, 2016.

[40] Maksim Tkachenko and Hady W Lauw. Comparelda: A topic model for document comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7112–7119, 2019.

[41] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[42] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010.

[43] Maksims N Volkovs, Hugo Larochelle, and Richard S Zemel. Learning to rank by aggregating expert preferences. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 843–851, 2012.

[44] Maksims N Volkovs and Richard S Zemel. A flexible generative model for preference aggregation. In *Proceedings of the 21st international conference on World Wide Web*, pages 479–488, 2012.

[45] Chong Wang and David M Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456, 2011.

[46] Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 783–792, 2010.

[47] Hongning Wang, Yue Lu, and ChengXiang Zhai. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 618–626, 2011.

[48] ChengXiang Zhai, Atulya Velivelli, and Bei Yu. A cross-collection mixture model for comparative text mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 743–748, 2004.

[49] Ce Zhang and Hady W Lauw. Topic modeling on document networks with adjacent-encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6737–6745, 2020.

[50] Delvin Ce Zhang and Hady W Lauw. Semi-supervised semantic visualization for networked documents. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2021.

[51] Zhibing Zhao, Peter Piech, and Lirong Xia. Learning mixtures of plackett-luce models. In *International Conference on Machine Learning*, pages 2906–2914. PMLR, 2016.

[52] Jun Zhu, Amr Ahmed, and Eric P Xing. Medlda: maximum margin supervised topic models. *the Journal of machine Learning research*, 13(1):2237–2278, 2012.