CompareLDA: A Topic Model for Document Comparison

Maksim Tkachenko, Hady W. Lauw
School of Information Systems
Singapore Management University
maksim.tkatchenko@gmail.com, hadywlauw@smu.edu.sg

Abstract

A number of real-world applications require comparison of entities based on their textual representations. In this work, we develop a topic model supervised by pairwise comparisons of documents. Such a model seeks to yield topics that help to differentiate entities along some dimension of interest, which may vary from one application to another. While previous supervised topic models consider document labels in an independent and pointwise manner, our proposed Comparative Latent Dirichlet Allocation (*CompareLDA*) learns predictive topic distributions that comply with the pairwise comparison observations. To fit the model, we derive a maximum likelihood estimation method via augmented variational approximation algorithm. Evaluation on several public datasets underscores the strengths of *CompareLDA* in modelling document comparisons.

Introduction

Due to the abundance of text data, there is a need for exploratory analysis of a text corpus. Topic model is a class of probabilistic models that "reduce" an input corpus into a manageable number of "topics", where each topic congeals words that tend to co-occur with one another in documents, thus signifying some hidden semantics in the corpus. By identifying the topics that essentialize the corpus, and discerning which ones predominate in a specific document, a topic model is a crucial tool for sensemaking.

Increasingly, there are real-world scenarios where the purpose of analyzing a corpus is to compare entities based on their textual representations (documents). For example, analysts may seek to explore why one country could achieve a better healthcare (alternatively economic, educational, etc.) outcome than another based on certain documents such as country reports. Funding agencies or scientists may seek a better understanding of what may get a grant proposal funded over another based on proposal contents. Among the products browsed by consumers, some are purchased while others are not. Among the purchases, some satisfy customers more than others. Thus, delving into product descriptions or reviews could reveal insights on consumer preferences.

An unsupervised topic model, such as Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), is oriented towards capturing topics that could reflect the word co-occurrences in the corpus well. In doing so, its topics tend to capture general semantics. For instance, a topic model based on country reports may well discover topics aligned to geographical or linguistic commonalities, which however may or may not bear direct relevance to the question at hand (e.g., healthcare outcomes). Topics based on grant proposals may describe various scientific foci, though such topics may group competing proposals but may not be indicative of their likelihood of acceptance. In turn, product reviews may yield topics focused on brands or features, but such topics may coalesce opposing sentiment polarities as words with positive (e.g., "good") or negative (e.g., "bad") connotations tend to co-occur with similar words, e.g., "battery life".

Problem We postulate that introducing supervision that signals how one entity (document) compares to another into topic modeling would better align the topics discovered from a corpus to the comparison dimension of interest. Suppose that in addition to a corpus of documents, we are also given some pairwise comparisons among the documents. Each pairwise comparison indicates which of two documents is considered "higher" or "better" according to some desired dimension (e.g., a country healthier than another, a pair of accepted and rejected proposals, a product preferred to another). Constraining the topic model to "comply" with the pairwise comparison observations may yield topics that differentiate entities along the dimension of interest (e.g., why one product is preferred), rather than simply discovering commonality in words (e.g., products with similar features).

Such topic modeling supervised with pairwise comparisons between documents as we are proposing is indeed novel. Previous work on supervised topic modeling would expect a different supervision in the form of pointwise response variables, e.g., a numerical rating. A case in point is sLDA (Mcauliffe and Blei 2008). However, there are inherent advantages to modeling pairwise comparisons as opposed to pointwise ratings. The latter may not even be available in some scenarios. In the implicit feedback settings, comparisons are naturally relative, when it may be known that one entity is better, but not necessarily clear by how much in absolute terms. For instance, when a consumer browses but skips a product, and purchases another, the latter is probably preferred to the former. Even when pointwise ratings are available, fitting the absolute ratings may not al-

ways be appropriate. They may have been assigned by different human subjects (with varying biases and scales), rendering direct comparison across human subjects inequitable.

Proposed Approach and Contributions In a nutshell, our proposed model *CompareLDA* associates each topic with a distribution over words, and each document with a distribution over topics, as in a conventional topic model. In addition, a document topic proportion maps to a merit value that ranks documents. These entity merit values probabilistically determine the observed pairwise comparison outcomes. As a generative model, *CompareLDA* has generative capacity over unobserved pairwise comparisons. This enables the model to learn even with relatively few observed comparisons, as we will see in the experiments. It also generalizes to out-of-training documents whose topic distributions and entity merit values could be inferred accordingly.

In this work, we make the following contributions. *First*, we investigate the utility of jointly modeling topics and pairwise comparisons of documents within an integrated model, which is the first of its kind. We design the generative model for our proposed approach *CompareLDA*, and describe a learning algorithm to infer the parameters using variational inference and simulated annealing. *Second*, through comprehensive experiments on public datasets, we showcase the value of supervising topic model with pairwise comparisons, against baselines such as sLDA that learns from pointwise supervision induced from the same inputs, as well as LDA that is not informed by any comparison-based supervision.

Related Work

Topic models (Hofmann 1999; Blei, Ng, and Jordan 2003) model associations between documents, topics, and words in a corpus. However, there may be auxiliary information that could reveal the core semantics in the corpus. Previous works model this as supervision to align the topics accordingly. The closest to our work is the pointwise sLDA (Mcauliffe and Blei 2008), which we use as a baseline to showcase the concept of pairwise comparison as opposed to pointwise supervision. There are yet others that pursue pointwise supervision, but explore other angles that are not directly comparable. (Lacoste-Julien, Sha, and Jordan 2009) introduces class-specific linear transformation to modify the topic distribution of a document, which would be applicable only to categorical labels but not continuous numerical responses. (Zhu, Ahmed, and Xing 2012) explores max-margin learning. (Ramage et al. 2009b; 2009a) associate a document with multiple labels (e.g., tags).

There also exist previous work that leverage document-pair supervision, such as two documents being similar or being linked in a network (Chang and Blei 2009; Mei et al. 2008; Erosheva, Fienberg, and Lafferty 2004; Yang, Boyd-Graber, and Resnik 2016; Chang, Boyd-Graber, and Blei 2009). In contrast, we are concerned with "merit-based ranking", i.e., one document considered "better" than another.

Other topic models focus on idiosyncratic notions of "comparison" different from ours. For instance, (Tkachenko and Lauw 2014) models comparison between two named entities within the same document (sentence), whereas we compare two different documents. In turn, (Zhai, Velivelli, and Yu 2004) compares two or more corpora, by finding shared topics and distinct topics between the corpora. Also focusing on corpora-level comparison, (Fang et al. 2012) seeks to identify contrasting opinions on specific topics.

Modeling pairwise comparison among documents is different from modeling sentiments, essentially pointwise categorical labels (Lin and He 2009). In some cases, a document has different sentiments represented by different sentences (Rahman and Wang 2016). It is also distinct from topic models that seek to capture personalized preferences (Wang and Blei 2011; McAuley and Leskovec 2013) or preferences in pairwise comparisons (Ding, Ishwar, and Saligrama 2015).

Model

In this section, we describe the development of our approach *CompareLDA*, as well as the methodology to fit the model parameters through variational inference.

Overview

CompareLDA is a supervised topic model with non-linear response. A response variable is associated with a pair of documents (each concerning an entity), and indicates the comparison result: which of the two entities is "better" or ranked higher than the other. The notion of comparison is latent, and may vary from application to application.

CompareLDA extends Latent Dirichlet Allocation (LDA). It has the same basis assumption regarding the association of topics and words, but also a significant distinction in its incorporation of pairwise comparisons (as we will see shortly). Each document is generated from a set of latent topics. A topic is an unknown distribution over the corpus vocabulary, which has to be inferred from the data. The documents in a corpus share the same set of topics, but mix them in different proportions. The topics are associated with the words and essentially defined as distributions over the vocabulary. Each word is a sample from only one topic distribution.

We are given a set of entities $D = \{d_i\}_{i=1}^{N}$. An entity is represented by its textual form, a document. The notation d_i is used to refer to either entity or document. Furthermore, we assume that an oracle takes a pair of entities at a time, d_i and d_j, 'glances' at their documents, and makes a comparison decision: which of the two entities is better according to some definition of merit, e.g., healthier, more likely to get funded, preferred by consumers. The decision would be based on the topics discussed in the text rather than on individual words. The oracle makes M pairwise comparison decisions, providing the training data. *CompareLDA* seeks to reproduce this judging process by learning the topics, and inferring these topics for unseen documents.

Definition

CompareLDA unfolds the process in the following way. Each entity is imbued with a latent merit value, inducing a pairwise comparison with another entity. Suppose m_i and m_j are the respective merit values for a pair of entities d_i and d_j. If $m_i > m_j$, then d_i is more likely, though not certainly, to come out the winner in a comparison with d_j.

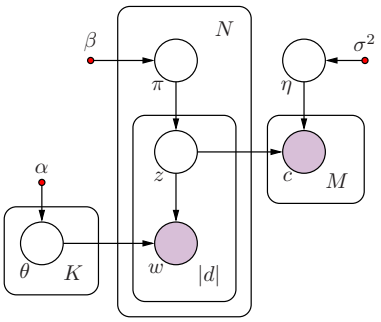


Figure 1: *CompareLDA* in plate notation.

To define the probability of winning in a comparison, we use the sigmoid function:

$$\mathrm{P}(d_i \succ d_j) = \sigma(m_i - m_j) = \frac{1}{1 + e^{-(m_i - m_j)}}. \quad (1)$$

The greater is m_i than m_j, the higher the probability that d_i would be favored by the oracle, as the probability in Eq. 1 tends towards 1. When the merit values are similar $m_i \approx m_j$, the probability reflects uncertainty in the outcome, i.e., $\mathrm{P}(d_i \succ d_j) \approx 0.5$.

Presumably, the oracle obtains the comparison information from the topics. For instance, preferred products or healthier countries may be associated with special qualities whose description manifests as topics. *CompareLDA* uses the empirical topic distributions of the texts, and transforms them into the entity merit values via a linear regression. Given a text d_i where each word w_{ij} is assigned to topic z_{ij}, we calculate its empirical topic distribution $\bar{z}_i$ as follows:

$$\bar{z}_i = \frac{1}{|d_i|} \sum_{j=1}^{|d_i|} z_{ij}. \quad (2)$$

For some regression parameters $\bar{\eta}$, we assume:

$$m_i = \bar{\eta} \cdot \bar{z}_i \quad (3)$$

Note that for such merit values as defined above, the bias term is effectively redundant, as it vanishes when comparison is concerned (Eq. 1).

The intuition behind regressing on topics is that some topics help to gain merit values (e.g., newly introduced product features), while others may decrease the merit values (e.g., discovered flaws). Considering the difference in the topic proportions of two products, $\bar{z}_i - \bar{z}_j$, we would be able to draw the conclusion on which entity is likely the winner.

Generative Process

Here we summarize the generative process of *CompareLDA*, whose plate notation is given in Figure 1.

1. We sample K topic distributions $\{\theta_i\}_{i=1}^K$ from Dirichlet distribution with α prior:

$$\theta_i \sim Dirichlet(\alpha).$$

2. We sample $\bar{\eta}$, the transformation weights from K-dimensional Gaussian with zero mean and σ^2 variance:

$$\bar{\eta} \sim \mathcal{N}(0, \sigma^2).$$

3. For each document:

(a) We sample its topic proportion $\{\pi_i\}_{i=1}^N$ from Dirichlet distribution with β prior:

$$\pi_i \sim Dirichlet(\beta).$$

(b) For each word w_{ij} in document d_i we sample its topic assignment variable z_{ij}:

$$z_{ij} \sim Categorical(\pi_i);$$

and based on the topic assignment the observed word:

$$w_{ij} \sim Categorical(\theta_{z_{ij}}).$$

(c) We calculate empirical topic proportion $\bar{z}_i$ and transform it to the entity merit value m_i:

$$m_i = \bar{\eta} \cdot \bar{z}_i = \bar{\eta} \cdot \left(\frac{1}{|d_i|} \sum_{i=1}^{|d_i|} z_{ij} \right).$$

4. For each pairwise comparison trial r_i, we sample the winner c_i:

$$c_i \sim Bernoulli(\sigma(m_{r_i[1]} - m_{r_i[2]})).$$

r_i is a pair of indices indicating which documents are compared in the trial, and c_i indicates the winner for the pair. If $c_i = 1$, then item $d_{r_i[1]}$ is the winner, otherwise $d_{r_i[2]}$ is.

The complete data likelihood for a set of entities/documents D and their corresponding pairwise comparison observations (R, C) is as follows:

$$\mathrm{P}(D, Z, \Theta, \Pi, R, C, \eta) = \mathrm{P}(\eta|\sigma^2) \prod_{i=1}^{K} \mathrm{P}(\theta_i|\alpha)$$
$$\times \prod_{i=1}^{N} \mathrm{P}(\pi_i|\beta) \times \prod_{i=1}^{N} \prod_{j=1}^{|d_i|} \mathrm{P}(z_{ij}|\pi_i) \mathrm{P}(w_{ij}|\theta_{z_{ij}})$$
$$\times \prod_{i=1}^{M} \mathrm{P}(c_i|m_{r_i[1]}, m_{r_i[2]}) \quad (4)$$

We consider the collapsed version of the likelihood by integrating out the multinomial parameters.

$$\mathrm{P}(D, Z, R, C, \eta) = \int_{\Theta} \int_{\Pi} \mathrm{P}(D, Z, \Theta, \Pi, R, \eta)$$
$$= \mathrm{P}(\bar{\eta}|\sigma^2) \times \prod_{i=1}^{N} \mathrm{P}(z_{i\cdot}|\beta) \times \mathrm{P}(W|Z, \alpha)$$
$$\times \prod_{i=1}^{M} \mathrm{P}(c_i|m_{r_i[1]}, m_{r_i[2]}), \quad (5)$$

where $\mathrm{P}(z_{i\cdot}|\beta)$ and $\mathrm{P}(W|Z, \alpha)$ are Dirichlet-multinomial distributions over topics and words.

Model Fitting

To find the maximizing latent parameters $\bar{\eta}$ and Z for the posteriori distribution, we use variational approximation algorithm to optimize the evidence lower bound $L(Z, R, C, \eta)$.

$$\begin{aligned}\log \mathrm{P}(D, Z, R, C, \eta) \geq L(Z, R, C, \eta) = \log \mathrm{P}(\eta|\sigma^2) \\ + \sum_{i=1}^{N} \langle \log \mathrm{P}(Z_i|\alpha) \rangle + \langle \log \mathrm{P}(W|Z, \beta) \rangle \\ + \sum_{i=1}^{M} \left\langle \log \mathrm{P}(c_i|m_{r_i[1]}, m_{r_i[2]}) \right\rangle + \mathrm{H}(q), \end{aligned} \tag{6}$$

where $\langle \cdot \rangle$ indicates expectation taken with respect to the variational distribution $q(Z)$, and $\mathrm{H}(\cdot)$ is the entropy operator. We treat $\bar{\eta}$ as a model parameter.

We factorize the variational distribution into independent factors, one for each z_{ij}. In most of the cases, assuming the fully factorized distribution is enough to obtain the tractable closed-form update equations, where each factorized distribution will have the same form as their conjugate priors. However, due to non-linear interaction term between the entities, the update formulas are intractable. We additionally assume that each $q(z_{ij})$ is an indicator probability distribution, which places all the probability mass on the one topic $q(z_{ij}) = q(z_{ij}|v_{ij}) = \mathbb{I}_{[z_{ij}=v_{ij}]}$. Essentially each v_{ij} represents empirical topic assignment for word w_{ij}.

$$q(Z|V) = \prod_{i=1}^{N} \prod_{j=1}^{|d_i|} q(z_{ij}|v_{ij}) = \prod_{i=1}^{N} \prod_{j=1}^{|d_i|} \mathbb{I}_{[z_{ij}=v_{ij}]} \tag{7}$$

Note that under this assumption, expectation operator does not change any $f(V)$: $\langle f(V) \rangle = f(V)$.

We use coordinate-ascent variational approximation, and maximize the evidence lower bound with respect to $\bar{\eta}$ and V, optimizing each parameter in turn. Further we assume that the document comparisons are configured in such a way that for any i, $c_i = 1$, to simplify the description of equations.

Optimizing η: With V fixed, we want to optimize the following objective:

$$\begin{aligned} f(\bar{\eta}) &= \log \mathrm{P}(\bar{\eta}|\sigma^2) + \sum_{i=1}^{M} \left\langle \log \mathrm{P}(c_i|m_{r_i[1]}, m_{r_i[2]}) \right\rangle \\ &= -\sum_{i=1}^{K} \frac{\eta_i^2}{2\sigma^2} - \sum_{i=1}^{M} \log\left(1 + e^{-\bar{\eta} \cdot \left(\bar{v}_{r_i[1]} - \bar{v}_{r_i[2]}\right)}\right) \end{aligned} \tag{8}$$

where $\bar{v}_i = \left(\sum_{j=1}^{|d_i|} v_{ij}\right) / |d_i|$. We develop a basic gradient ascent algorithm, taking derivative of $f(\bar{\eta})$ with respect to η_j:

$$f'_{\eta_j}(\bar{\eta}) = -\frac{\eta_j}{\sigma^2} - \sum_{i=1}^{M} \frac{(\bar{v}_{r_i[1]})_j - (\bar{v}_{r_i[2]})_j}{1 + e^{-\bar{\eta} \cdot \left(\bar{v}_{r_i[1]} - \bar{v}_{r_i[2]}\right)}}. \tag{9}$$

Optimizing V: With $\bar{\eta}$ fixed, we seek the empirical topic assignments that maximize (6). As exhaustive search for the optimal solution has exponential complexity and, therefore, is infeasible for any reasonable datasets, we exploit the probabilistic nature of the model and develop Metropolis-Hastings (Hastings 1970) procedure for approximate optimization. Metropolis-Hastings is a method for obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult. In case of *CompareLDA*, the procedure changes one word-topic assignment v_j as a time, eventually approximating the probability distribution of word-topic assignments for the whole corpus. We work with probability distribution induced by the empirical lower bound. Here we compute the difference between two word-topic assignments, that are different only in one assignment, current assignment $v_{ij} = a$ and evaluated assignment $v_{ij} = b$ (chosen at random).

$$\begin{aligned} E_{a \to b}^{v_{ij}} = \log\left(\beta + n_i(a) - 1\right) - \log\left(\beta + n_i(b)\right) \\ + \log\left(\alpha + n(a, w_{ij}) - 1\right) - \log\left(\alpha + n(b, w_{ij})\right) \\ - \log\left(\alpha|X| + n(a, \cdot)\right) - 1) + \log\left(\alpha|X| + n(b, \cdot)\right) \\ + \sum_{k=1}^{M} \log\left(1 + e^{\bar{\eta} \cdot \left(\bar{v}_{r_k[1]} - \bar{v}_{r_k[2]}\right) + \delta_i \left(\mathbb{I}_{[r_k[1]=i]} - \mathbb{I}_{[r_k[2]=i]}\right)}\right) \\ - \sum_{k=1}^{M} \log\left(1 + e^{\bar{\eta} \cdot \left(\bar{v}_{r_k[1]} - \bar{v}_{r_k[2]}\right)}\right) \end{aligned} \tag{10}$$

where X is vocabulary, $n_i(z)$ is document-topic count, $n(z, w)$ is term-topic count, $n(z, \cdot) = \sum_{w \in X} n(z, w)$, and $\delta_i = (\eta_b - \eta_a) / |d_i|$. The acceptance probability $\gamma = \exp\left(-E_{a \to b}^{v_{ij}}\right)$ then indicates how probable the evaluated sample is with respect to the current assignment, according to the approximated sample. If we attempt to move to an assignment which is more probable than the current one w.r.t the evidence lower bound, we always accept the move. If the move is taken towards the less probable assignment, it will be accepted with γ probability.

For the purpose of optimization, the Metropolis-Hastings algorithm can be converted to simulated annealing procedure, where acceptance probability γ is reduced over time, to prevent the moves towards less probable states. We use $\gamma^{\frac{1}{T}}$ as the probability of accepting a new assignment, where T, a temperature, approaches 0 as the iteration count increases.

Experiments

Our experimental objective is to validate the efficacy of *CompareLDA* in deriving topics that are well-aligned to document comparisons. First, we investigate the utility of modeling pairwise comparisons as supervision on topic models, vis-à-vis a baseline with pointwise supervision. Thereafter, we move to additional experiments and discussions, which shed light upon various aspects of the model.

Datasets

For experiments, we rely on public text corpora, whereby not only it is meaningful to attach the notion of comparisons to entities within a corpus, but the comparisons also define

a part of the core semantics of the corpus. We identify three such corpora that yield five experimental datasets as follows.

Wikipedia The first is a set of three datasets constructed from Wikipedia[1] pages with country infobox and category. The corpus contains 467 entities (countries and associations, e.g., BRICS, NATO). The page content is the document. As supervision, we induce three sets of pairwise comparisons from Wikipedia's lists of countries: by alcohol consumption (AC), by cigarette consumption (CC), and by life expectancy (LE). Each list results in a different number of pairwise comparisons: 17,955 for AC, 16,290 for CC, and 16,653 for LE. Coupled with the text corpus, each set of pairwise comparisons constitute a dataset. Our intention is to study if *CompareLDA* could derive different topics from the same corpus, but with different pairwise comparisons.

Product Reviews The second dataset is from Amazon as described in (McAuley, Pandey, and Leskovec 2015; McAuley et al. 2015). Here, an entity is a review from the Electronics category. We assume that the reviews mention various features and qualities to illustrate the product's intrinsic merit. As supervision, we induce pairwise comparisons based on the number of stars indicated by the reviews. The "positive" reviews (5 and 4 stars) are compared to the "negative" reviews (2 and 1 stars), i.e., positive "win" over negative. We sample 10,000 reviews at random to assemble the corpus. For this dataset, out of all the induced comparisons, we randomly sample 0.25% to simulate a realistic scenario of where only partial comparisons have been observed. The dataset contains 43,881 pairwise comparisons.

Movie Reviews The third dataset contains movie reviews (Pang and Lee 2005). We used the 4-star scale as described in (Pang and Lee 2005) to induce pairwise comparisons, i.e., a review with more stars "wins". The intuition here is to discover the topics that are aligned with what makes a good movie. As before for the Product Reviews corpus, we retain only 0.25% of comparisons as supervision. The corpus contains 5,006 documents along with 21,965 comparisons.

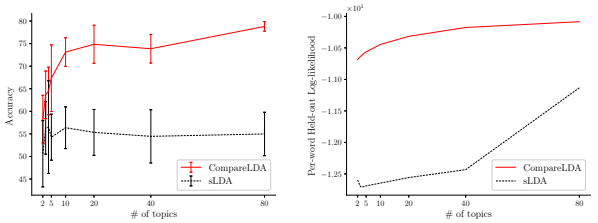
Each dataset is split into training and testing folds in 80:20 proportion respectively. Conservatively, comparisons that cross folds are ignored during training and evaluation. The corpora undergo the same preprocessing steps, i.e., removing short documents, punctuation, stop-words; the tokens converted to their lemmas. For each dataset, we retain top 5000-term vocabulary selected by tf-idf.

Evaluation

To jointly model topics and pairwise comparisons, a method should be adept at both assigning topics to words and assessing the ranking among documents. We explore these respective dimensions of evaluation.

Ranking To assess ranking quality, we report the ranking accuracy. For two entities d_i and d_j, we define a function f, where $f(d_i, d_j)$ returns 1 if d_i is preferred over d_j in comparison, 0 when the preference between d_i and d_j is not assumed, and -1 when d_j is preferred over d_i. Given a set of entities $D = \{d_i\}_{i=1}^{M}$ and reference comparison function f (ground-truth) and its approximation g (prediction),

[1]We used the Wikipedia dump dated 30 July 2018.



(a) Accuracy (b) Log-likelihood

Figure 2: Wikipedia dataset ranked by alcohol consumption

we define the ranking accuracy (or accuracy) as follows:

$$A = \frac{\sum_{i=1}^{M}\sum_{j=1}^{M} \mathbb{I}_{[f(d_i,d_j)=1]}\mathbb{I}_{[g(d_i,d_j)=1]}}{\sum_{i=1}^{M}\sum_{j=1}^{M} \mathbb{I}_{[f(d_i,d_j)=1]}} \tag{11}$$

When the approximation and reference functions are identical, then approximation is good and $A = 1$. In case of complete disagreement, $g(d_i, d_j) \neq 1$ for every d_i and d_j such that $f(d_i, d_j) = 1$, then $A = 0$. The ranking accuracy is closely related to Kendall's Tau. While Kendall's Tau is suitable for totally ordered sets, the proposed metric consider only items for which relative comparison make sense, and thus it is more appropriate in our study.

Topics Topic models are commonly evaluated by estimating probability of held-out documents. The intuition is that a better model will give rise to the likelihood of held-out documents D. $L = \frac{\log \mathrm{P}(D|\mathcal{M})}{\sum_{d\in D}|d|}$ is per-word log-likelihood for an LDA model with parameters $\mathcal{M}$. To approximate L marginalized over all possible topic assignments, we use Chib-style estimator (Wallach et al. 2009).

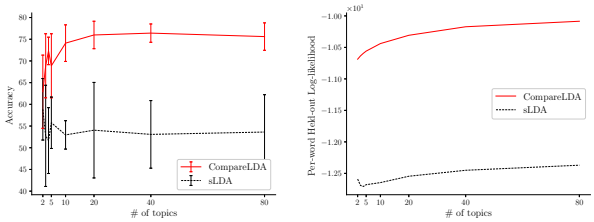
Comparison to Baseline

As our proposed *CompareLDA* is the first topic model supervised by pairwise comparisons, our main baseline is the previous topic model supervised by pointwise response variables. Among such models (see Related Work), sLDA[2] bases the prediction on empirical topic assignments, which makes the former an ideal baseline to *CompareLDA* that also uses empirical topic assignments. sLDA predicts merit values of documents directly via regression. When supervision is supplied in terms of pairwise comparisons, this model is not immediately applicable. Instead, it requires preprocessing to convert the pairwise comparisons into pointwise merit values for each document, which are then supplied to sLDA. For conversion, we employ the Bradley-Terry-Luce (BTL) model (Bradley and Terry 1952; Luce 2012) due to its similarity to the comparison component of *CompareLDA*.

We evaluate both models by varying the number of topics (default is 80). The experiments are repeated 10 times with different random initializations. Figures 2 to 6 show the results for the five datasets for both accuracy and likelihood.

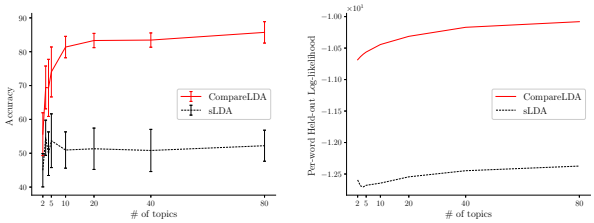
CompareLDA consistently outperforms sLDA on each dataset with respect to both evaluation dimensions. For

[2]We used the following implementation: https://github.com/vietansegan/segan/



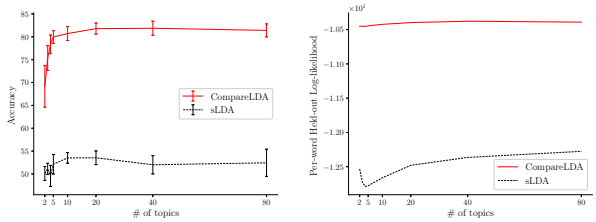
(a) Accuracy (b) Log-likelihood

Figure 3: Wikipedia dataset ranked by cigarette consumption



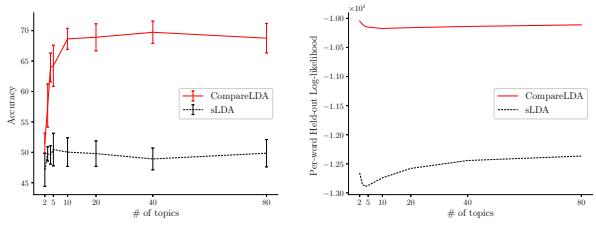
(a) Accuracy (b) Log-likelihood

Figure 4: Wikipedia dataset ranked by life expectancy



(a) Accuracy (b) Log-likelihood

Figure 5: Product reviews



(a) Accuracy (b) Log-likelihood

Figure 6: Movie reviews

instance, Figure 2(a) shows the ranking accuracy for the Wikipedia dataset ranked by alcohol consumption. In general *CompareLDA* achieves better results as the number of topics increases. The accuracy gap over sLDA increases significantly, when the number of topics hits 10 and beyond. Bars denote the standard deviation. The deviation tends to reduce as the number of topics increase. This reflects well on *CompareLDA*, suggesting that supervision in the form of pairwise comparisons helps to uncover the ranking structure. *CompareLDA* demonstrates better alignment of topics with rankings, reaching higher than 75% accuracy. sLDA shows lower performance, hovering around 55%, which is close to random; this suggests that the regression objective does not fit the problem, when pairwise supervision is concerned.

In turn, Figure 2(b) shows that *CompareLDA* reaches significantly higher log-likelihood than sLDA as well. The log-likelihood plots show significant outperformance by *CompareLDA* over sLDA even when the number of topics is small. It seems that the regression objective interferes with the objective to infer "good" predictable topics.

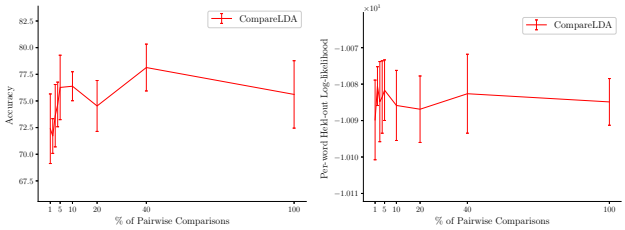
The other Wikipedia datasets ranked by cigarette consumption (Figure 3) and life expectancy (Figure 4) show similar trends, evidence that *CompareLDA* could derive different topic models from the same corpus by fitting different supervisions. For the review datasets (Figures 5 and 6), the outperformance is more vivid and starts with a few topics.

Amount of Supervision

We study the amount of supervision, as the number of comparisons for the fully ordered set of N elements is quadratic, $\mathcal{O}(N^2)$, i.e., harder to obtain than independently labeling each document. Figure 7 shows the performance when the amount of supervision gradually increases from 1% to 100% on the Wikipedia dataset ranked by cigarette consumption for 80 topics (other rankings and topic counts show similar results). Figure 7(a) shows that initially ranking accuracy grows fast as the amount of supervision increases. When 5% of supervision is supplied, rankings accuracy remains stable. Figure 7(b) shows that log-likelihood remains stable regardless of the amount of supervision. These results indicate that *CompareLDA* does not require fully ordered set to fit the model and, therefore, a small subset of comparisons may be used to achieve high ranking performance and topic quality.

Joint vs. Pipeline Models

One may improbably surmise that LDA may naturally align with document comparisons anyway, even without super-



(a) Accuracy (b) Log-likelihood

Figure 7: Varying number of pairwise comparisons. Wikipedia dataset ranked by cigarette consumption. The other rankings show similar behavior.

Data	*CompareLDA*	LDA+BTL-R
Wikipedia (AC)	$\mathbf{78.8} \pm 0.8$	75.4 ± 1.6
Wikipedia (CC)	75.6 ± 2.3	74.9 ± 1.8
Wikipedia (LE)	85.7 ± 2.3	84.0 ± 1.8
Product Reviews	$\mathbf{81.4} \pm 1.0$	76.7 ± 1.0
Movie Reviews	$\mathbf{68.8} \pm 1.7$	63.8 ± 1.3

Table 1: Accuracy, AC - Alcohol Consumption, CC - Cigarette Consumption, LE - Life Expectancy.

Data	*CompareLDA*	LDA(+BTL-R)
Wikipedia (AC)	$-10.084 \pm .007$	$-10.084 \pm .010$
Wikipedia (CC)	$-10.085 \pm .005$	$-10.084 \pm .010$
Wikipedia (LE)	$-10.081 \pm .010$	$-10.084 \pm .010$
Product Reviews	$-10.391 \pm .003$	$-10.389 \pm .003$
Movie Reviews	$-10.111 \pm .004$	$-10.111 \pm .002$

Table 2: Log-likelihood, AC - Alcohol Consumption, CC - Cigarette Consumption, LE - Life Expectancy

vision. To debunk this, we consider a decoupled form of *CompareLDA*, which first discovers topics with LDA, and then solves the comparison problem using the empirical topic assignments. To tackle the pairwise comparison, we introduce Bradley-Terry-Luce regression (BTL-R), which is similar to *CompareLDA*'s comparison modeling but done as a separate step. We refer to this pipeline as LDA+BTL-R.

Table 1 shows the ranking accuracy (and 95% confidence intervals). Bold typeface indicates statistically significant difference. *CompareLDA* shows better results than its pipeline equivalent on every dataset, with significant improvement on Wikipedia ranked by alcohol consumption and the review datasets. In turn, for the held-out log-likelihood, one may expect some decrease in performance due to additional objective to satisfy the comparison supervision, whereas LDA (BTL-R part does not influence topic inference in this case) cares only about getting the topics right. Gratifyingly, Table 2 shows that in fact there is no significant difference between the topics derived by *CompareLDA* and LDA, supporting that *CompareLDA* could align topics to comparisons well without hurting the likelihood.

sLDA Supervision

As mentioned earlier, sLDA requires pointwise supervision. When the input is pairwise, we need a preprocessing step. In a scenario where some form of pointwise response exists, we could alternatively use that directly, e.g., the rank position in the list for the Wikipedia dataset. We look into whether the two forms of supervision affect the results much. sLDA* is supervised with the ranked list, whereas sLDA is supervised with comparisons. Table 3 shows that there is no significant difference for ranking accuracy between the two. The BTL transformation matters when we explore held-out log-likelihood on Wikipedia ranked by alcohol consumption (see Figure 4), where it helps to achieve significantly better performance. However, the differences for the other Wikipedia rankings are not significant. In any case, the form of sLDA supervision would not affect the earlier conclusions on the relative comparisons with *CompareLDA*.

Data	sLDA	sLDA*
Wikipedia (AC)	55.0 ± 3.4	55.6 ± 4.2
Wikipedia (CC)	53.6 ± 6.2	52.6 ± 5.2
Wikipedia (LE)	52.2 ± 3.3	50.8 ± 3.5

Table 3: Accuracy, AC - Alcohol Consumption, CC - Cigarette Consumption, LE - Life Expectancy

Data	sLDA	sLDA*
Wikipedia (AC)	$\mathbf{-11.132} \pm .008$	$-12.367 \pm .007$
Wikipedia (CC)	$-12.370 \pm .006$	$-12.368 \pm .011$
Wikipedia (LE)	$-12.374 \pm .009$	$-12.373 \pm .016$

Table 4: Log-likelihood, AC - Alcohol Consumption, CC - Cigarette Consumption, LE - Life Expectancy

Topics

To get a sense of the semantics reflected by the topics, we show 5 topics associated with top positive and top negative $\bar{\eta}$ parameters. For Product Reviews (see Table 5), the $\bar{\eta}$-positive topics tend to associate with words of positive connotations, e.g., great, well, good, recommend, love, etc. $\bar{\eta}$-negative topics tend to talk about issues, problems, money, returns, and warranty.

Top $\bar{\eta}$-positive Topics	Top $\bar{\eta}$-negative Topics
work great well phone use also everything since need easy good set recommend issue clear	product would one back new month work buy get warranty worked return issue problem year
picture great tv price good quality love amazing feature best get really recommend got better	one would tried time review product work bought got money new first returned different try
color great little look came still perfect get easy could love easily really would want	device work adapter even connection get computer use unit time product well network car ca
one fan great two really also work air new room put purchased right got connector	year bought still first week working month warranty another one since would two completely last
one use bought price year well good model work still frame know used wanted made	one thing like money buy get even worth really could got make review cheap going

Table 5: *CompareLDA* topics for Product Reviews

Conclusion

We describe *CompareLDA*, a topic model for document comparison. It is novel in its incorporation of pairwise com-

parison to align the topics learnt to the comparison dimension of interest. Experiments show that it helps to uncover more conducive topics for assessing the relative merits between entities than baseline with pointwise supervision. Moreover, it learns well even with partial supervision assuaging the need for many comparison labels.

Acknowledgments

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its NRF Fellowship Programme (Award No. NRF-NRFF2016-07).

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *JMLR* 3(Jan).

Bradley, R. A., and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*.

Chang, J., and Blei, D. M. 2009. Relational topic models for document networks. In *AIStats*, volume 9, 81–88.

Chang, J.; Boyd-Graber, J.; and Blei, D. M. 2009. Connections between the lines: augmenting social networks with text. In *SIGKDD*. ACM.

Ding, W.; Ishwar, P.; and Saligrama, V. 2015. A Topic Modeling Approach to Ranking. In Lebanon, G., and Vishwanathan, S. V. N., eds., *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, 214–222. PMLR.

Erosheva, E.; Fienberg, S.; and Lafferty, J. 2004. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences* 101(suppl 1).

Fang, Y.; Si, L.; Somasundaram, N.; and Yu, Z. 2012. Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, 63–72. ACM.

Hastings, W. K. 1970. Monte carlo sampling methods using markov chains and their applications. *Biometrika*.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *SIGIR*. ACM.

Lacoste-Julien, S.; Sha, F.; and Jordan, M. I. 2009. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *NIPS*.

Lin, C., and He, Y. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and Knowledge Management*, 375–384. ACM.

Luce, R. D. 2012. *Individual choice behavior: A theoretical analysis*. Courier Dover Publications.

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

McAuley, J.; Targett, C.; Shi, Q.; and van den Hengel, A. 2015. Image-based recommendations on styles and substitutes. In *SIGIR*, 43–52.

McAuley, J.; Pandey, R.; and Leskovec, J. 2015. Inferring networks of substitutable and complementary products. In *KDD*, 785–794.

Mcauliffe, J. D., and Blei, D. M. 2008. Supervised topic models. In *NIPS*.

Mei, Q.; Cai, D.; Zhang, D.; and Zhai, C. 2008. Topic modeling with network regularization. In *WWW*, 101–110. ACM.

Pang, B., and Lee, L. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, 115–124. Association for Computational Linguistics.

Rahman, M. M., and Wang, H. 2016. Hidden topic sentiment model. In *Proceedings of the 25th International Conference on World Wide Web*, 155–165. International World Wide Web Conferences Steering Committee.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009a. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*. Association for Computational Linguistics.

Ramage, D.; Heymann, P.; Manning, C. D.; and Garcia-Molina, H. 2009b. Clustering the tagged web. In *WSDM*. ACM.

Tkachenko, M., and Lauw, H. W. 2014. Generative modeling of entity comparisons in text. In *CIKM*. ACM.

Wallach, H. M.; Murray, I.; Salakhutdinov, R.; and Mimno, D. 2009. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, 1105–1112. ACM.

Wang, C., and Blei, D. M. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 448–456. ACM.

Yang, W.; Boyd-Graber, J.; and Resnik, P. 2016. A discriminative topic model using document network structure. In *ACL*.

Zhai, C.; Velivelli, A.; and Yu, B. 2004. A cross-collection mixture model for comparative text mining. In *SIGKDD*. ACM.

Zhu, J.; Ahmed, A.; and Xing, E. P. 2012. MedLDA: maximum margin supervised topic models. *JMLR* 13(Aug).