**CONTRAVis: Contrastive and Visual Topic Modeling for Comparing Document Collections**

Tuan V. M. Le  
H. John Heinz III College  
Carnegie Mellon University  
tuanminhlv@yahoo.com

Leman Akoglu  
H. John Heinz III College  
Carnegie Mellon University  
lakoglu@andrew.cmu.edu

**ABSTRACT**

Given posts on ‘abortion’ and posts on ‘religion’ from a political forum, how can we find topics that are discriminative and those in common? In general, (1) how can we compare and contrast two or more different (‘labeled’) document collections? Moreover, (2) how can we visualize the data (in 2-d or 3-d) to best reflect the similarities and differences between the collections?

We introduce (to the best of our knowledge) the first contrastive and visual topic model, called CONTRAVis, that jointly addresses both problems: (1) contrastive topic modeling, and (2) contrastive visualization. That is, CONTRAVis learns not only latent topics but also embeddings for the documents, topics and labels for visualization. CONTRAVis exhibits three key properties by design. It is (i) **Contrastive**: It enables comparative analysis of different document corpora by extracting latent discriminative and common topics across labeled documents; (ii) **Visually-expressive**: Different from numerous existing models, it also produces a visualization for all of the documents, labels, and the extracted topics, where proximity in the coordinate space is reflective of proximity in semantic space; (iii) **Unified**: It extracts topics and visual coordinates simultaneously under a joint model;

Through extensive experiments on real-world datasets, we show CONTRAVis’s potential for providing visual contrastive analysis of multiple document collections. We show both qualitatively and quantitatively that CONTRAVis significantly outperforms both unsupervised and supervised state-of-the-art topic models in contrastive power, semantic coherence and visual effectiveness.

**CCS CONCEPTS**

- Computing methodologies → Topic modeling;  
- Mathematics of computing → Dimensionality reduction.

**KEYWORDS**

contrastive topic models; visualization; comparative text mining

**ACM Reference Format:**


This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

**1 INTRODUCTION**

Topic modeling is widely used for identifying latent themes (=topics) in a large collection (or corpus) of documents. Topics are simply groups of words that best represent the information in the corpus. As such, they aid in sense-making by characterizing the corpus at a high level, and can also be used for several downstream tasks like indexing, search and classification.

Vast majority of topic models are designed to extract topics from a single corpus. In many cases, on the other hand, one is interested in comparing and contrasting documents from multiple corpora. These corpora could be from different time periods, e.g., science articles 1960-80 vs. 1980-2000; spatial locations, e.g., news articles from ‘Asia’ vs. ‘USA’ vs. ‘Europe’; or different sub-populations, e.g., essays on neuroscience vs. psychology.
Our Goal. In this work, we focus on the latter area, broadly known as comparative text mining [28]. Specifically, we address the cross-collection modeling and visualization problems, motivated by the following questions. How can we find hidden topics from two or more document collections? What are the common and discriminative topics among these collections? Further, how can we visualize the documents and topics to easily explore, compare and contrast the collections? In this paper, we refer to the task of finding common and discriminative topics as contrastive topic modeling as these topics help to contrast document collections. In addition, the visualization produced by embedding document collections and their contrastive topics is referred to as contrastive visualization. Ideally, a contrastive visualization can help users to easily spot the discriminative and common topics. In this work, we are interested in scatterplot visualization where documents and topics are embedded in a 2-d or 3-d visualization space. §6 provides a brief review of other visualization forms for visual comparison.

Previous Approaches. Topic models like LDA [2] and its variants can be used to find hidden topics from document collections. However, since these models are unsupervised, the extracted topics may not be discriminative. For contrastive topic modeling, we can leverage the labels of the document collections to employ supervised topic models [9, 16, 20, 21, 30]. Although the learned topics may be discriminative, these models are not designed for visualization. One can perform a post-hoc embedding using dimensionality reduction methods such as t-SNE [14] for visualization. This two-phase approach, however, has two different objective functions (one for topic modeling and one for visualization) that are optimized separately, which may result in poor visualization. Recently, new models that follow a joint approach have been developed; tying together the two steps using a single objective function [7, 10–12]. This type of unified approach is pioneered by PLSV [7], which jointly models topics and visualization by locating documents and topics in the same visualization space. However, PLSV is unsupervised and therefore the extracted topics and visualization may not be discriminative for contrastive analysis. We compare to these previous approaches through extensive experiments in §4 and §5.

Our Model. We introduce a supervised joint technique called ContraVis that addresses both contrastive topic modeling and contrastive visualization. It simultaneously learns (a) latent topics among multiple corpora for sensemaking and (b) 2-d or 3-d embeddings for documents, identified topics, and labels (where each corpus is represented with a unique label) for visual analysis. The topics capture both the common themes across the corpora as well as discriminative themes specific to each corpus/label. Moreover, the embeddings in the visual coordinate system reflect the semantic proximities between (i) the documents and topics and (ii) the topics and labels. To the best of our knowledge, ContraVis is the first supervised joint technique for simultaneous contrastive topic modeling and visual embedding.

Example. To demonstrate ContraVis’s novel aspects, we show Figure 1 on an example real-world dataset from NYTNews (see Section 4.1) where we contrast three labeled corpora: sports, arts and style. In (b) and (c), we use unsupervised and supervised topic models, PLSV [7] and sLDA [16] respectively, to extract topics and topic probability distributions of the labeled documents. For sLDA, we also perform a post-hoc embedding of the documents to 2-d based on their topic representations using t-SNE [14]. In contrast, we show the output of our ContraVis in (a). Besides extracting topics, ContraVis readily produces a visual embedding (i.e., no post-hoc embedding is necessary). Moreover, it embeds not only the documents but also the labels (triangles) and the topics (circles). This provides a holistic view of the data. In (a), one can see the top few representative words or wordclouds per topic, like other models. What is more: (i) one can also see how documents relate to different topics and visually infer each document’s probability distribution over topics, and (2) how topics relate to different labels (corpora) which facilitates identifying the common and discriminative ones.

To summarize, our model exhibits the following key properties, which constitute the main contributions of our work.

• **Contrastive power:** We propose a new topic model called ContraVis for the comparative analysis of multiple document corpora. It quantifies its identified topics in terms of their relevance to different corpora, revealing common and discriminative ones.

• **Visual-expressiveness:** ContraVis produces embeddings, to visualize (all of) documents, labels, and topics in 2- or 3-d. Proximity in embedding space is reflective of proximity in semantic space, which helps easily infer (i) most dominant topics per document and (ii) most relevant/discriminative topics per label.

• **Unified nature:** ContraVis estimates the topics and the embedding coordinates simultaneously within a single, joint model. Thus, it is a more holistic model for comparative text analysis than existing models that require post-hoc embeddings.

We evaluate the semantic, contrastive, and visualization quality of ContraVis both qualitatively and quantitatively on diverse real-world datasets. We show the superiority of ContraVis and its potential as a method for visual contrastive analysis of multiple document collections, from diverse domains and from different sub-populations, space, or time.

Reproducibility: We share the source code for ContraVis and all of our public-domain datasets at https://github.com/tuanlvvm/ContraVis.

2 PROBLEM DEFINITIONS

In this paper we aim to build a model to compare and contrast documents across different corpora, e.g., three collections of posts respectively on ‘gun control’, ‘taxes’, and ‘military’. Documents belonging to a specific corpus can be labeled by its name. As such, the input is a collection of single-label documents $D = \{(d_1, s_1), \ldots, (d_N, s_N)\}$, where $s_n \in L$ is the label of document $d_n$ and $L$ is the set of unique labels from a finite domain. For output, our model seeks to produce a comparison that is interpretable; by inferring common topics and discriminative ones for each label. In addition, for an effective exploratory and contrastive analysis, we propose to use an visualization where documents, topics and labels are embedded in the same space such that the visual proximities between them reflect their semantic similarities and differences. Given these objectives, we give our problems formally.

Given a collection of single-label documents $D$ containing words from a finite vocabulary $V$ and labels from set $L$, number of topics $K$, and visualization dimension $d$;

**Problem 1 (Contrastive Topic Modeling).** Find

(i) $K$ latent topics, and word probability distributions of topics,
Table 1: Notation used in text.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>number of documents, (</td>
</tr>
<tr>
<td>N</td>
<td>vocabulary, (</td>
</tr>
<tr>
<td>d</td>
<td>a specific document</td>
</tr>
<tr>
<td>M_d</td>
<td>number of words in document (d)</td>
</tr>
<tr>
<td>(L)</td>
<td>set of unique labels, (</td>
</tr>
<tr>
<td>(s_n)</td>
<td>single label of document (d_n)</td>
</tr>
<tr>
<td>c</td>
<td>common label, which is applied to all documents</td>
</tr>
<tr>
<td>(\Lambda_n)</td>
<td>(\Lambda_n = {s_n, c})</td>
</tr>
</tbody>
</table>

Input Corpus

| K       | total number of topics |
| z       | a specific latent topic |
| \(\beta_z\) | word distribution of topic \(z\) |
| \(\phi_l\) | latent coordinate of topic \(z\) in visualization space |
| \(\mu_l\) | latent coordinate of label \(l\) in visualization space |
| \(x_l\) | collection of \(x_l\), \(z\) for all topics |
| \(\lambda\) | \(\lambda = \{\phi_l|\phi_l\}\) for all topics |
| \(\phi_l\) | collection of \(\phi_l\), \(l\) for all topics |
| \(\mu_l\) | collection of \(\mu_l\), \(l\) for all topics |

Visualization

(ii) topic distributions of documents, and (iii) topic and word distributions of labels.

**Problem 2 (Contrastive Document Visualization).** Find \(d\)-dimensional visualization coordinates for (i) \(N\) documents, (ii) \(K\) topics, and (iii) \(L\) labels, such that the spatial proximities between documents, topics, and labels are reflective of (or proportional to) the topic-document and topic-label probability distributions above.

In our formulation, we introduce a new label \(c\) as a common label, which is applied to all documents. Hereafter, we denote \(\mathcal{L} = \{\mathcal{L}, c\}\) and \(\Lambda_n = \{s_n, c\}\) as the set of labels for each \(d_n\). We use label \(c\) to capture the common topics among documents. Note that we infer \(K\) latent topics overall and find the probability distribution of these topics over different labels. This provides a "soft assignment" of topics to labels, from which we can deduce the relevance of each topic to each label, as well as mutual and common topics across labels. This reduces the parameter complexity of our model; as compared to finding \(K\) mutual topics per label \(l \in \mathcal{L}\) and \(K\) common topics, which would require \(L + 1\) hyperparameters to choose, we only set \(K\) (details in §3.2.2).

3 PROPOSED CONTRAVIS MODEL

We introduce CONTRAVIS, a new, visually expressive contrastive topic model that jointly addresses Problem 1 and Problem 2. Specifically, CONTRAVIS introduces an innovative generative process that ties together the probability distributions with the visualization coordinates, as we describe next.

3.1 Generative Process

Besides inferring topic-word and document-topic probability distributions, our additional objective is to learn latent visualization coordinates \(x_l\) of each document \(d_n\) and \(\phi_l\) for each topic. In addition, each label \(l\) is also assumed to have a latent coordinate \(\mu_l\) in the same visualization space. Let \(\mu = \{\mu_l\}_{l=1}^{L}\). The label distribution of \(d_n\) is expressed by its Euclidean distances to labels as follows:

\[
P(l|\mu, \mu) = \frac{\exp(-\frac{1}{2}||x_n - \mu||^2)}{\sum_{l'=1}^{L} \exp(-\frac{1}{2}||x_n - \mu_l'||^2)}
\]

According to Eq. (1), \(P(l|\mu, \mu)\) is high when \(x_n\) is close to \(\mu_l\) since we observe the label \(s_n\) of document \(d_n\). \(d_n\) should be placed close to \(\mu_{s_n}\) in the visualization space so that \(P(s_n|\mu, \mu)\) is high. Note that each document also has a probability \(P(c|x_n, \mu)\) of belonging to the common label \(c\). By introducing the common label, we can model the overlap among labels while still aiming to separate them for extracting discriminative topics.

For each document, we introduce a label-dependent topic distribution \(P(z|\mu, \mu, \Phi)\), which is determined by:

\[
P(z|\mu, \mu, \Phi) = \frac{\exp(-\frac{1}{2}||\mu - \phi_z||^2)\exp(-\frac{1}{2}||x_n - \phi_z||^2)}{\sum_{z'=1}^{K} \exp(-\frac{1}{2}||\mu - \phi_{z'}||^2)\exp(-\frac{1}{2}||x_n - \phi_{z'}||^2)}
\]

We can see that for a label \(l\), \(x_n\) has high \(P(z|\mu, \mu, \Phi)\) when it is close to topic \(\phi_z\) and \(\phi_z\) is close to label \(\mu_l\). Therefore, as \(x_n\) is to be placed close to its observed label \(s_n\), topics that best describe \(d_n\) and are discriminative for \(s_n\) are also to be placed close to \(s_n\)

The steps of the generative process of CONTRAVIS are as follows:

1. For each label \(l \in \mathcal{L}\) (including common label c):
   (a) Draw \(l\) ’s coordinate: \(\mu_l \sim \text{Normal}(0, \sigma_0^{-1}I)\)
   (b) Draw \(z\)’s word distribution: \(\beta_z \sim \text{Dirichlet}(\lambda)\)
   (c) Draw \(z\)’s coordinate: \(\phi_z \sim \text{Normal}(0, \sigma_z^{-1}I)\)

2. For each topic \(l = 1, \ldots, K\):
   (a) Draw \(z\)’s word distribution: \(\beta_z \sim \text{Dirichlet}(\lambda)\)
   (b) Draw \(z\)’s coordinate: \(\phi_z \sim \text{Normal}(0, \sigma_z^{-1}I)\)
   (c) Draw a topic: \(z \sim \text{Mult}(\{P(z|\mu, \mu, \Phi)\}_{z=1}^{K})\)
   (d) Draw a word: \(w_{nm} \sim \text{Mult}(\beta_z)\)

3. Set \(\Lambda_n = \{\text{unique labels of } w_{nm}\}\)

Above, \(I\) is the \(d \times d\) identity matrix. \(\sigma_0, \varphi, \text{ and } \gamma\) are hyperparameters that control the variance of the Normal distributions. Similar to LDA [2], CONTRAVIS associates each topic with a word probability distribution \(\beta_z\) where \(P(\beta_z) = \frac{1}{\Gamma((\lambda+1)W)} \prod_{w=1}^{W} \beta_{z,w}^\lambda\) and \(\lambda\) is the Dirichlet prior. Coordinates of topics and labels are drawn from spherical Normal distributions with mean at zero. For each document, we assume that its coordinate is drawn from Normal distributions with mean at \(\mu_{s_n}\). Note that this does not ensure that its generated label will be \(s_n\). For each word \(w_{nm}\) in document \(d_n\), first we draw a label \(l\) and draw a topic \(z\) from \(P(z|\mu, \mu, \Phi)\), \(w_{nm}\) is then drawn from multinomial \(\beta_z\). Finally, labels of \(d_n\) are unique labels of \(d_n\)’s words. Figure 2 illustrates the corresponding graphical model for the generative process of our model.

3.2 CONTRAVIS Inference

The parameters of CONTRAVIS include the word distributions of topics \(\beta = \{\beta_z\}_{z=1}^{K}\) document coordinates \(x = \{x_n\}_{n=1}^{N}\) topic coordinates \(\Phi = \{\phi_z\}_{z=1}^{K}\) and label coordinates \(\mu = \{\mu_l\}_{l=1}^{L}\) (including common label \(c\)). Let \(\Omega = \{\beta, x, \Phi, \mu\}\). In addition to the model parameters, the label and topic assignment of words in the documents are unobserved latent variables (See (3bi) and (3bii) of the generative process in §3.1). Therefore, we use the EM algorithm.

\[
\text{EM} = \{(\Omega, \mu, \mu, \Phi)\} = \{(\beta, x, \Phi, \mu)\}
\]
to infer $\Omega$ from $D$ based on maximum a posteriori estimation. Data likelihood can be written as follows (hyperparameters omitted for simplicity).

$$P(D|\Omega) = \sum_{L,W} \sum_{Z_{W}} P(D, L_{W}, Z_{W}|\Omega) = \prod_{n=1}^{N} P(\Lambda_n) P(d_n|x_n, \mu, \Phi, \beta)$$

$$= \prod_{n=1}^{N} \left[ P(\Lambda_n) \sum_{m=1}^{M} \sum_{l_{n}=1}^{L_{n}} \sum_{z_{n}=1}^{K} P(l|x_n, \mu) P(z|l, x_n, \Phi) P(w_{nm}|b_{n}) \right]$$

where $L_{W}, Z_{W}$ are respectively label assignments and topic assignments to all the words $W$ in $D$, and

$$P(\Lambda_n) = \sum_{m=1}^{M} \sum_{l_{n}=1}^{L_{n}} P(l|x_n, \mu)$$

The EM algorithm seeks to find the MLE of the marginal likelihood with priors under the current estimate of the parameters $\hat{\Omega}$ is

$$C(\Omega|\hat{\Omega}) = E_{(L_{W}, Z_{W})|D, \hat{\Omega}}[\log P(D, L_{W}, Z_{W}|\Omega)]$$

$$= \sum_{n=1}^{N} \sum_{m=1}^{M} \log \sum_{l=1}^{L} P(l|x_n, \mu)$$

$$+ \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{l=1}^{L} \sum_{z=1}^{K} P(l, z, x_n, \mu) \log \left[ P(l|x_n, \mu) P(z|l, x_n, \Phi) P(w_{nm}|b_{n}) \right]$$

$$+ \sum_{n=1}^{N} \log P(x_n) + \sum_{l=1}^{L} \log P(\phi_{l}) + \sum_{l=1}^{L} \log P(\beta_{l}) + \sum_{l=1}^{L} \log P(\mu_{l})$$

where

$$P(x_n) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} \| x_n - \mu_{n} \|^2 \right)$$

$$P(\phi_{l}) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} \| \phi_{l} \|^2 \right)$$

$$P(\mu_{l}) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} \| \mu_{l} \|^2 \right)$$

We set hyper-parameters $\varphi = 0.1N$ as used in PLSV [7], $\gamma = L^2$ and $\sigma_0 = 0.1N$ which work well in practice. The conditional distribution of label $l$ and topic $z$ given the $n$th word in the $n$th document under current estimate of the parameters is given as

$$P(l, z|n, m, \hat{\Omega}) = \frac{P(l|x_n, \mu) P(z|l, x_n, \Phi) P(w_{nm}|b_{n})}{\sum_{l'_{n}=1}^{L_{n}} \sum_{z'_{n}=1}^{K} P(l'|x_n, \mu) P(z'|l', x_n, \Phi) P(w_{nm}|b_{n})}$$

M step: We update the entries of each $b_{n}$ as follows:

$$\hat{b}_{n} = \frac{\sum_{n=1}^{N} \sum_{m=1}^{M} w_{nm} \mathbb{1}(w_{nm} = w) P(z|n, m, \hat{\Omega}) + \lambda W}{\sum_{n=1}^{N} \sum_{W_{n}=1}^{W} \sum_{m=1}^{M} \mathbb{1}(w_{nm} = w) P(z|n, m, \hat{\Omega}) + \lambda W}$$

where $\mathbb{1}(X) = 1$ if $X$ is true and 0 otherwise. As for the other parameters, $x_n, \phi_{z}, \mu_{l}$'s, we do not have closed form solutions. We update them using the quasi-Newton method L-BFGS [17] with the following gradients of $C(\Omega|\hat{\Omega})$ w.r.t $x_n, \phi_{z}, \mu_{l}$.

$$\frac{\partial C(\Omega|\hat{\Omega})}{\partial x_n} = \sum_{m=1}^{M} \sum_{l=1}^{L} \left( P(l|x_n, \mu) P(z|l, x_n, \Phi) - P(l, z|n, m, \hat{\Omega}) \right) (2x_n - \mu_{l})$$

$$\frac{\partial C(\Omega|\hat{\Omega})}{\partial \phi_{z}} = \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{l=1}^{L} \left( P(l|x_n, \mu) P(z|l, x_n, \Phi) - P(l, z|n, m, \hat{\Omega}) \right) (2\phi_{z} - \mu_{l})$$

$$\frac{\partial C(\Omega|\hat{\Omega})}{\partial \mu_{l}} = \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{l=1}^{L} \left( P(l|x_n, \mu) P(z|l, x_n, \Phi) - P(l, z|n, m, \hat{\Omega}) \right) (2\mu_{l} - \mu_{l})$$

Our estimated model parameters, $\Omega = (\beta, \chi, \Phi, \mu)$, respectively capture topic word distributions (Problem 1 (i)) as well as document, topic, and label coordinates (Problem 2). Using $\Omega$ we can also compute the topic probability distribution of a document $d_n$ (Problem 1 (ii)) and topic and word distributions of a label $l$ (Problem 1 (iii)), respectively as follows.

$$P(z|d_n, \Omega) = \sum_{l=1}^{L} P(l|x_n) P(z|l, x_n, \Phi)$$

$$P(z|l, \Omega) = \frac{\sum_{n=1}^{N} P(d_n) P(l|x_n) P(z|l, x_n, \Phi)}{\sum_{n=1}^{N} P(d_n) P(l|x_n)}$$

$$P(w|l, \Omega) = \frac{\sum_{n=1}^{N} \sum_{m=1}^{M} P(d_n) P(w_{nm}) P(z|l, x_n, \Phi) P(w_{nm}|b_{n})}{\sum_{n=1}^{N} \sum_{m=1}^{M} P(d_n) P(w_{nm})}$$

3.2.1 Quantifying Common and Discriminative Topics

Based on $p(z|l)$ as in Eq. (15), we define cnm_score $= p(z|c)$ and disc_score $= \frac{p(z|l)}{\max_{l'}p(z|l')}$ to identify common and discriminative topics, respectively. Visually, common topics are those close to the common label and discriminative topics of a label are those near to that label but far from the other labels in the embedding space.

3.2.2 Choosing $K$.

In practice, setting $K$ is challenging for the contrastive task for guessing the total number of relevant topics across corpora is nontrivial. Topic models are typically evaluated by either measuring performance on some external task, such as document classification, or by estimating the probability of unseen held-out documents. On average, a better model produces a higher probability of held-out documents.

In our case, we aim to find $K$ that gives rise to a model that explains (or represents) the original data $D$ the best. To this end,
we fix topic word distributions as well as the topic and label coordinates, and pick $K$ (among a set of alternatives) that maximizes the following marginal likelihood of the input documents.

$$P(D | \mu, \Phi, \beta) = \frac{1}{N} \sum_{n=1}^{N} P(\Lambda_n) P(d_n | \Lambda_n, \mu, \Phi, \beta) P(\Lambda_n) \, dx_n.$$  \hfill (16)

where $P(d_n | x_n, \mu, \Phi, \beta)$ is as given in Eq. (3). Here we treat input documents as though they are unseen and integrate out the unknown variables (their coordinates) to compute their likelihood. The integral in Eq. (16) is not tractable, as such, we use importance sampling \cite{27} to estimate the marginal likelihood, where we sample $S=(500)$ samples of $x_n$ from $P(x_n)$. The marginal likelihood is then approximated as:

$$P(D | \mu, \Phi, \beta) \approx \frac{1}{S} \sum_{s=1}^{S} P(\Lambda_n) P(d_n | x_n^{(s)}, \mu, \Phi, \beta).$$  \hfill (17)

3.2.3 Computational Complexity. In ContraVis inference algorithm, E-step is performed in $O(NLKW)$ where we evaluate Eq. \hfill (9) for all words of each document. For M-step, the most expensive task is to perform the L-BFGS algorithm. As we know, for each iteration, L-BFGS has a computational cost of $O(gp)$ plus the cost to evaluate $C(\Omega \dot{\Omega})$ in Eq. (5) and its gradients \cite{17}. Here, $g$ is the number of steps stored in the memory and $p = (N + L + K)d$ is the number of variables. The evaluations of $C(\Omega \dot{\Omega})$ and its gradients can be performed in $O(NLKW)$. Since modest values of $g$ are often between 3 and 20 \cite{17} and the number of iterations in L-BFGS is often small, M-step can be performed at a cost of $O(NLKW)$. Therefore, ContraVis inference has an overall computational cost of $O(NLKW)$ per iteration, which is linear in the number of documents, labels, topics and vocabulary size.

4 QUANTITATIVE EVALUATION

We evaluate ContraVis both quantitatively (in this section) and qualitatively (in §5, through many case studies). Before detailing our experiment results, we describe the datasets we used for evaluation, which come from three different domains: news articles, political posts, and employee peer reviews.

4.1 Dataset Description

NYTnews consists of all the news articles published at the New York Times from January 1987 to June 2007.\footnote{Publicly available at \url{https://catalog.ldc.upenn.edu/ldc2008t19}} Articles are labeled by online sections where they are published to. We omit articles published in more than one online section and select the following $L=9$ major departments: sports, business, arts, style, travel, technology, real-estate, magazine, and health. We also use time information to contrast articles, like technology news from 1990’s vs. 2000’s.

PolForum contains posts from a political forum, which is organized into various threads of subject areas for discussion.\footnote{Publicly available at \url{https://github.com/tuanlvm/ContraVis/tree/master/data}} Posts are labeled by the thread under which they appear. In total there are $L=8$ different threads; in order of frequency: abortion, religion, taxes, economy, military, gun-control, environment, and race.

PeerReview contains peer-reviews among employees of a ride-sharing company (proprietary dataset). We group all the reviews each employee received into a document and label documents by their department. The company has $L=4$ major departments: operations, engineering, people-analytics, and finance.

Table 2 shows the size of each dataset after preprocessing. In our experiments, for each dataset, we sample 250 documents from each label uniformly at random and construct 10 such collections with different samples. We report results averaged across 10 collections.

In this section, we evaluate ContraVis on three different quantitative tasks, and compare it to existing topic models rigorously. The tasks capture both semantic and visualization quality of the competing methods. We present quantitative performance on all datasets; NYTnews with 9, PolForum with 8, and PeerReview with 4 labels.

Before detailing our empirical results, we describe the baseline methods that we compared to.

4.2 Comparison Methods

Table 3 lists comparative methods and their properties. We compare ContraVis to both unsupervised and supervised topic models. ContraVis makes use of document labels for topic modeling and visualization, as such it is in the supervised category.

(1) (Supervised) sLDA\hfill [16] (s for supervised)

(2) (Supervised) DiscLDA\hfill [4] (Disc for discriminative)

(3) (Unsupervised): LDA\hfill [2]

The above baselines do not produce an explicit visual embedding of the documents (nor of the topics or labels), unlike ContraVis. For our quantitative tasks that evaluate visualization quality, we pipeline these methods with dimensionality reduction. Specifically, we obtain the documents’ topic representation and use t-SNE \hfill [14] to embed them into 2-d. (See for e.g., Fig. 8b)

(4) (Unsupervised & Joint (Topics+Vis)): PLSV\hfill [7]

(5) (Supervised & Joint (Topics+Vis)): ContraVis [this paper].

4.3 Task 1: Discrimination in 2-d

With this work, we set out to not only perform topic modeling but also visualization that enables contrastive analysis— which sets our work apart. As such, the first quantitative task involves contrastive power in the visualization space. Intuitively, a good contrastive visualization should well separate the documents with common topics from those belonging to discriminative topics. Moreover, it should be easy for users to spot these documents in the visualization.

Table 2: Dataset statistics. Descriptions given in §4.1.

<table>
<thead>
<tr>
<th></th>
<th>#labels $L$</th>
<th>#documents $N$</th>
<th># words $W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYTnews</td>
<td>9</td>
<td>515252</td>
<td>8666</td>
</tr>
<tr>
<td>PolForum</td>
<td>8</td>
<td>18732</td>
<td>6774</td>
</tr>
<tr>
<td>PeerReview</td>
<td>4</td>
<td>5336</td>
<td>4032</td>
</tr>
</tbody>
</table>

Table 3: Comparison of methods by properties.

<table>
<thead>
<tr>
<th>Properties vs. Methods</th>
<th>LDA</th>
<th>sLDA</th>
<th>DiscLDA</th>
<th>PLSV</th>
<th>ContraVis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrastive (Supervised)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Visually-expressive</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Unified/jointly modeled</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

\footnote{We use author implementation at \url{https://github.com/blei-lab/class-slda}. This is a variant of sLDA with a categorical response \cite{5}.}

\footnote{We obtained the implementation from the author Prof. Simon Lacoste-Julien. In our experiments, we use DiscLDA with fixed transformation matrix $T$.}

\footnote{http://scikit-learn.org (sklearn.decomposition.LatentDirichletAllocation)}

\footnote{We use the implementation at \url{https://github.com/tuanlvm/SEMAFORE}}
With this intuition, we design two subtasks for evaluation, named Common and Discriminative, described as follows.

**Setup:** To setup, we identify the top three most frequent labels from each dataset, which we refer to as C, B, and A (without loss of generality). For the Common task, we sample 250 documents each from label A and B, and 100 from C. We split C documents in half and mix with those from A and B. We consider 300 documents (250 A + 50 C) as class 1 and the other 300 (250 B + 50 C) as class 2. We then apply the methods to compare class 1 and class 2. The goal is to identify the documents with common topics among the two classes, i.e. the hidden label C ones which are considered as ground truth for the Common task.

For the Discriminative task, we sample 500 documents from C, and 50 each from A and B. We consider 300 documents (250 C + 50 A) as class 1 and the other 300 (250 C + 50 B) as class 2. We then apply the methods to compare class 1 and class 2. This time the goal is to identify the documents with discriminative topics among the two classes, i.e. the hidden label A and B ones which are considered as ground truth for the Discriminative task.

We quantify the ranking performance of the methods. Concretely, we rank the documents by their distance to the common label c, which enables visual contrastive analysis. To setup, we identify the top three most frequent labels (in increasing order for Common, and decreasing order for Discriminative). Since the baseline methods do not have such a c, we first find the mean coordinate of all documents from each class and take the mean of these means to designate as the common label’s coordinate. We repeat the experiments for 10 randomly selected samples per dataset, and report average precision and recall on each subtask. Basically, a good contrastive visualization will have a better ranking performance because it separates well the documents with common topics from the ones with discriminative topics.

**Results:** Figure 3 shows the precision versus recall curves for all methods on each dataset for the Common task, and Figure 4 the respective plots for the Discriminative task. We see that ContraVis produces a better ranking for both tasks and outperforms the competing methods.

More concretely, Table 4 and Table 5 provide the mean average precision (area under PR curves) respectively for both tasks. We find that ContraVis is significantly better than the baselines in many cases according to the paired Wilcoxon signed rank test. An example illustration on NYTNews is given in Figure 5 comparing ContraVis with sLDA for these two subtasks. These results show the contrastive power of ContraVis in the visualization space. In addition, the ContraVis’s ranking performance indicates that users can easily spot the similarities or differences among document collections by simply observing the distances between documents and the common label c, which enables visual contrastive analysis. Many case studies will be presented in §5 to show the potential of ContraVis for contrasting and comparing document collections.

**4.4 Task 2: Word-to-Label Relevances**

We find that the ContraVis’s contrastive power in visualization space (§4.3) does not come at the expense of discriminative power and semantic coherence. To show this, we measure semantic discrimination quality, particularly, how well ContraVis can identify the most relevant words to a label (or class) as compared to the baseline methods. For instance, humans would frequently use words like ‘life’ and ‘baby’ when talking about abortion, and the words ‘god’ and ‘faith’ for religion (See Figure 10).

![Figure 3: Precision-Recall curves (avg’ed over 10 samples) for task Common.](https://example.com/fig3.png)

![Figure 4: Precision-Recall curves (avg’ed over 10 samples) for task Discriminative.](https://example.com/fig4.png)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ContraVis</th>
<th>sLDA</th>
<th>DiscLDA</th>
<th>LDA</th>
<th>PLSV</th>
</tr>
</thead>
<tbody>
<tr>
<td>PolForum</td>
<td>0.599 ± 0.06</td>
<td>0.374 ± 0.06</td>
<td>0.157 ± 0.06</td>
<td>0.158 ± 0.06</td>
<td>0.213 ± 0.12</td>
</tr>
<tr>
<td>PeerReview</td>
<td>0.423 ± 0.06</td>
<td>0.344 ± 0.06</td>
<td>0.207 ± 0.06</td>
<td>0.276 ± 0.09</td>
<td>0.152 ± 0.02</td>
</tr>
<tr>
<td>NYTNews</td>
<td>0.410 ± 0.04</td>
<td>0.156 ± 0.03</td>
<td>0.126 ± 0.03</td>
<td>0.121 ± 0.02</td>
<td>0.112 ± 0.01</td>
</tr>
</tbody>
</table>

**Ground truth:** There is no existing repository of word-to-subject-matter relevances that we could directly use as ground truth. Therefore, we use 3 different means to create a ranked list of words by relevance to each label.

1) **tf-idf:** We rank the words that appear in documents of a certain label by their frequency.

2) **Google 5-gram frequencies:** We take a large collection of 5-grams generated by Google Inc. from around 1 trillion word tokens of text from public Web pages. Each label l (e.g. religion), we count the number of times the word pair (w, l) co-occurs across the 5-grams, and sort the words (w’s) by their frequency.

3) **Yahoo 5-gram frequencies:** We obtain a 3rd ranking of words by label relevance based on Yahoo 5-grams, in the same fashion as the above.

---

1. https://catalog.ldc.upenn.edu/ldc2006t13
Figure 5: (best in color) Visualization of documents from NYTnews ($K=30$); Common task: (a) ContraVis (CV) places docs with common topics (sports) near the common label, (b) sLDA+: it is not as easy to spot common docs; Discriminative task: (c) CV places docs with discr. topics (business, arts) far from the common label as well as one another, (d) sLDA+: it is not easy to spot these docs. + is short for +t-SNE.

Setup: We quantify the word-to-label relevances of models by computing $p(w|l)$, which is given in Eq. (3.2) for ContraVis. For unsupervised models, LDA and PLSV, we compute $p(w|l) = \sum_z p(w|z)p(z|l)$, where $p(w|z)$ is readily output by the model and $p(z|l) = \frac{\sum_{z,n=1}^{z,N} P(z|d_{nl})}{\sum_{z,n=1}^{z,N} P(z|d_{nl})}$. For sLDA and DiscLDA, $p(w|l) \propto \sum_{z,n=1}^{z,N} P(w|z)p(z|d_{nl})p(l|d_{nl})p(d_{nl}) = \frac{1}{N}$, where $p(l|d_{nl})$ for sLDA is computed using softmax [5], and is approximated for DiscLDA using bridge sampling [9].

We measure the quality of the top $t$ words in a given model’s ranking, based on the Normalized Discounted Cumulative Gain: $NDCG = \frac{DCG@t}{DCG@t^*}$, $DCG@t = \sum_{i=1}^{t} \frac{rel_i}{\log(i+1)}$ (the higher, the better), where $rel_i$ is the “true” relevance of top $i$th word. We obtain $rel_i$ from the ground truth using corresponding word’s respective score used for ranking (i.e., frequencies or tf-idf). $IDCG@t$ is the “ideal” $DCG$, that of the top $t$ words of the ground truth itself.

Note that Yahoo and Google rankings contain many out-of-corpus words since they are based on external corpora, where we expect all methods to obtain lower $NDCG$ relative to the tf-idf ground truth with the ranked list of in-corpus words.

Results: Figure 6 shows $IDCG@t$ performance of the methods for $t \in \{10, 20, \ldots, 100\}$ on each dataset based on the tf-idf in-corpus ground truth. ContraVis consistently ranks among the top across all $t$ thresholds, closely followed by sLDA. To quantify the differences more concretely, we perform the paired Wilcoxon signed rank test across 10 samples and all labels of each dataset, as presented in Table 6 for all three ground truth. As one can see, ContraVis significantly outperforms the competing methods in many cases. On NYTnews there is no significant difference between ContraVis and sLDA, but, ContraVis is significantly superior to sLDA on the other two datasets across all ground truth.

Pairwise significance tests according to the paired Wilcoxon signed rank test reveal the (total or partial) orderings in Table 7. We find that ContraVis is significantly superior to all methods in many cases and ties with sLDA and LDA in some cases. We wrap up with Table 8 that provides qualitative evidence to the quality of $p(w|l)$ rankings by ContraVis. It ranks words like ‘white’ and ‘black’ at top for $l$ : race from PolForum, ‘candidate’ and ‘recruit’ for $l$ : people-analytics from PeerReview, and ‘building’ and ‘house’ for $l$ : real-estate from NYTnews, among others, which all agree with human intuition.
4.5 Task 3: Classification in 2-d
Since learning contrastive visualization is supervised, we can expect that the labels should be well separated in the visualization space. However, we find it is still valuable to show what is the extent that ContraVis separates the labels compared to the baselines. To investigate this, we perform \( k \) Nearest Neighbors (kNN) classification with visualization coordinates as inputs, i.e., each document will be assigned to the dominant label among its \( k \) nearest neighbors in the visualization space. The accuracy is the fraction of documents with predicted label matching the truth. We report the accuracy on each dataset, averaged across 10 collections. Intuitively, higher accuracy means that labels are more separated in the visualization space.

Results: Figure 7 shows kNN accuracy of all methods on each dataset, for varying (left) number of topics \( K \) and (right) number of nearest neighbors \( k \) and an example illustration on NYTnews is given in Figure 8 comparing ContraVis with DiscLDA (the second best performing baseline on this task). We note the near-perfect accuracy that ContraVis achieves across all datasets and settings which may imply that ContraVis aggressively separate the labels to learn discriminative topics and thus the learned topic model may be too overfitted. However, as discussed in §6, ContraVis focuses on comparing document collections, which aims to learn a discriminative topic model that best explains the similarities and differences among these document collections. Its focus is not on prediction tasks such as classification or regression. Therefore, overfitting may not be a big issue to ContraVis, but conversely, in some cases it may be a good trait (e.g., when we try to extract the differences between two very similar labels).

5 QUALITATIVE EVALUATION
In this section, we evaluate ContraVis qualitatively through many case studies where we apply our method to contrast and compare 2 or 3 document collections. These case studies represent common application scenarios. Our case studies are designed to answer two questions: (1) How well does ContraVis capture the underlying discriminative and common topics? Do these topics agree with human intuition? and (2) How good are the visualizations that ContraVis produces?

5.1 NYTnews
We perform our first case study on NYTnews using three labels (sports, arts, style). As shown in Figure 1(a), ContraVis well-separates the collections and the learned topics are also discriminative. For sports, we see topics about basketball and racing game. Topics exclusive to arts include `museum`, `art`, `exhibit` and `street`, `art`, `paint`. For style, we see a topic on fashion (`cloth`, `design`, `women`) and another about lifestyle (`graduate`, `university`, `father`). Figures 1(b) and 1(c) show the visualizations by PLSV and sLDA+t-SNE which are not as comprehensive and documents are not as well separated.

The next study compares documents across time. Specifically, we apply ContraVis to contrast technology articles in NYTnews from 1990’s vs. 2000’s. As shown in Figure 9, in 1990’s, tech-news talked about `website`, `travel`, `find`, `book`, etc. Notice that WWW was invented in 1989, after which computers started being used for finding information, booking travel, etc. In contrast, in 2000’s news started to involve `video`, `game`, and `xbox`, that is around the period when Xbox was introduced in 2001. In both time periods, tech-news discuss about `system`, `machine`, `power`, and `technology`. The number of articles also reflects the development of technology. We have only 79 articles on technology in 1990-1997, with a lot more in 2000 and onward.

5.2 PolForum
Next, we apply ContraVis to see what topics are discussed in different threads of PolForum. Figure 10 shows a contrastive visualization of abortion vs. religion posts. PLSV and sLDA+t-SNE
embeddings in Figure 10(b) and 10(c) do not provide a well visual comparison of the two threads. CONTRAVIS embedding in Figure 10(a), by contrast, shows separated clusters. Also shown are two most discriminative topics for each thread as well as two common topics. Abortion topics are intuitive; with words like ‘abort’, ‘plan’, ‘parenthood’ and ‘human’, ‘cell’, ‘zygote’. Under religion, people talk about Christianity and Islam, respectively involving words ‘god’, ‘jesus’, ‘bible’ and ‘islam’, ‘muslim’, ‘allah’. They share common topics around ‘peopl’, ‘opinion’, ‘position’, etc.

Figure 11 shows another use case, comparing posts on gun-control, taxes, & military from PolForum using CONTRAVIS (K=20). That CONTRAVIS still can detect discriminative topics which are closely related to the role of each department. For example, from CONTRAVIS embedding in Figure 12(a), one can tell that people-analytics department is responsible for recruiting, engineering deals with project design and testing, and operations is responsible for managing business in different regions. These insights cannot be gained from the embeddings by PLSV and sLDA+t-SNE embeddings mix the labels. sLDA+tSNE embedding is doc-only (no topics). Thru several case studies from different domains above, we have shown that CONTRAVIS finds meaningful topics that agree with human understanding. In addition, it provides an intuitive visualization that helps quickly identify common and discriminative topics and their relationship to documents and labels. Overall, CONTRAVIS shows promise for exploratory and comparative text analysis, and is superior to existing topic models like PLSV (unsupervised joint method) and sLDA which necessitates post-hoc embedding with subpar results.

6 RELATED WORK

Unsupervised Topic Modeling and Visualization: There are various unsupervised topic models [2, 6, 22]. LDA [2] is arguably the most popular one that represents each document as a mixture of topics and each topic as a probability distribution over words. For visualization, one needs to pipeline these with dimensionality reduction
methods such as t-SNE [14], LargeVis [26]. Recently, methods that jointly model topics and visualization using a single generative model have been developed. This line of research is pioneered by PLSV [7] and is further developed by later works that incorporate the neighborhood graph [12], use document network/hyperlinks information [10], or model documents’ spherical representation [11]. Similar to CONTRAVis, these methods assume that all documents and topics are represented as points in the same visualization space. The main differences between CONTRAVis and these methods are: (1) CONTRAVis is a supervised method where labels are used to extract the discriminative topics and produce contrastive visualization; (2) CONTRAVis assumes that each document has a label distribution and labels are also located as points in the visualization space. The distances from documents to labels encode the label distributions; (3) CONTRAVis introduces label-dependent topic distributions for each document. Basically, a document tends to pick topics belonging to its label for generating its words; (4) CONTRAVis explicitly models common topics by introducing the common label which is shared across documents. By modeling label distributions and label-dependent topic distributions, CONTRAVis can effectively extract discriminative and common topics. We extensively compare to LDA+t-SNE and PLSV. Other joint methods [10–12] are not directly comparable to CONTRAVis, as they use additional input such as neighborhood information or the document network. We can extend CONTRAVis to incorporate such information if available.

Supervised Topic Modeling and Visualization: Several supervised topic models have been proposed such as sLDA [16], DiscLDA [9] for single-label documents and LabeledLDA [20], PLDA [21] for multi-label and partially labeled documents respectively. Recently, MedLDA integrates the max-margin principle into supervised topic models by optimizing the goodness of fit of the learned topic model and its prediction accuracy on a max-margin classifier [29]. These models use document side information such as categories or review rating scores to steer the model learning towards extracting more discriminative topics for prediction tasks such as classification or regression. Unlike these supervised topic models, CONTRAVis focuses on the task of comparing document collections whose objective is to extracting common and discriminative topics for exploring both the similarities and the differences among these collections. Its focus is not on document classification tasks. Moreover, CONTRAVis also jointly produces a contrastive visualization, which enable visual comparison of document collections. For supervised topic models mentioned above, one needs to pipeline these methods with a post-hoc embedding using a dimensionality reduction technique (like t-SNE [14]), which is not effective for visually contrasting document collections as we demonstrated in our experiments. Among these models, sLDA and DiscLDA are more closely related to our work as they are for single-label documents and in principle we can extend CONTRAVis to integrate the max-margin mechanism as in MedLDA. In our experiments, we extensively compared to two state-of-the-art supervised topic models for single-label documents, sLDA+t-SNE and DiscLDA+t-SNE.

Comparative Text Mining: There have been works that directly model common and discriminative topics for comparative text mining. CCMix [28] introduces a probabilistic mixture model to identify $k$ common themes across all collections and $k$ collection-specific themes for each collection. ccLDA [19] improves over CCMix [28] by introducing an LDA-based analog, which automatically learns the probability of using the common vs. collection-specific word distributions while generating documents. While ccLDA maintains common and collection-specific word probability vectors for each topic, differential topic models [4] propose to use a hierarchy of topics across collections. Recently, there are works that propose using spectral methods [30] and nonnegative matrix factorization [8] for comparative text mining. However, they mainly focus on scenarios involving two collections while our method can handle multiple corpora. Moreover, although these methods can infer common and discriminative topics across collections, they do not provide any means for visual comparison, unlike CONTRAVis.

Visualization for Comparison: CONTRAVis uses scatterplot visualization to display the relationship between documents, topics, and labels. There exist work that propose other visualization forms for making comparisons. For example, DiTop-View, in which topics are represented as glyphs on a 2D plane, is used for comparing document collections [18]. Buddy plots, on the other hand, represent each document as a row and other documents are encoded as circular glyphs along the row for comparison of topic models [1]. TextDNA [25] use configurable colorfields to visualize word usage patterns across different text collections. Other work focus on comparing two or more documents using word clouds [3, 13]. Compared to these visualization forms, scatterplot by CONTRAVis provides an effective way to get an overview of multiple corpora as it explicitly shows the relationship between documents, topics and labels. A challenge with scatterplots is that they are susceptible to overdraw (overlapping glyphs); for very large document corpora one could employ several tackling strategies [15].

We note that work on summarizing and extracting phrases for contrasting opinions in text documents [23, 24] are different in focus, without specific emphasis on topic modeling or visual analysis.

7 CONCLUSION

We propose CONTRAVis, a new supervised topic model for contrasting and visualizing multiple document collections. Our model jointly learns common and discriminative topics as well as embeddings of documents, topics, and labels for visualization. Due to its joint nature of contrastive topic modeling and visualization, the learned topics and their relationships to documents and labels are reflected faithfully in the visualization, which facilitates exploring and comparing the collections. To the best of our knowledge, CONTRAVis is the first to jointly address the contrastive topic modeling and visualization problems. We conduct comprehensive experiments on real-life datasets and the results show that our method significantly outperforms the existing techniques in terms of contrastive power as well as visual and semantic coherence.

ACKNOWLEDGMENTS

This research is sponsored by NSF CAREER 1452425 and IIS 1408287. We thank Daniel Bay from Uber Technologies Inc. for helping us in our data analysis. Any conclusions expressed in this material are of the authors and do not necessarily reflect the views, expressed or implied, of the funding parties or our industrial partners.

---

Code and public-domain datasets available at https://github.com/tuanlvm/ContraVis
REFERENCES


