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:
Tuan V. M. Le and Leman Akoglu. 2019. ContraVis: Contrastive and Visual Topic Modeling for Comparing Document Collections. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308558.3313617

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313617

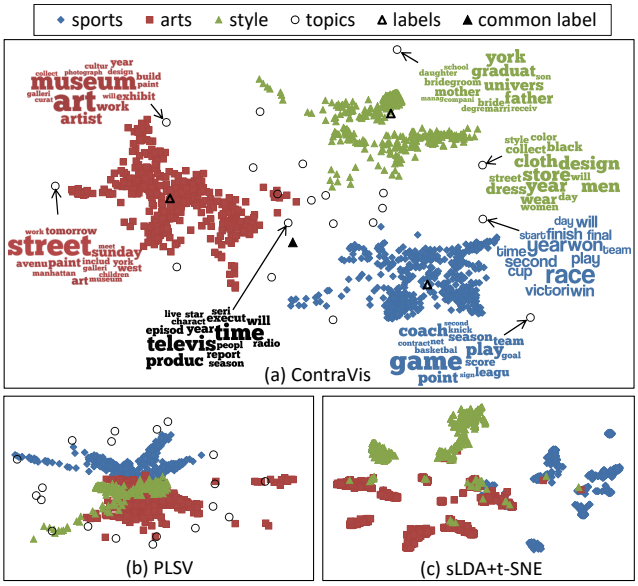

Figure 1: (best in color) Contrastive analysis of `sports`, `arts`, `style` documents from NYTnews (K=20 topics); (top) ContraVis (supervised joint method) not only finds latent discriminative topics but also gives visual embedding of documents and topics. Two discriminative topics per label shown in wordclouds (w/ respective colors) + one common topic (black); (bottom) *PLSV* (unsupervised joint method) gives visual embedding of documents and topics. However, the topics are not as discriminative, indicated by a mixing of labels. *sLDA*+tSNE (supervised topic modeling followed by embedding, document-only) has a mixing of labels and gives no means to visually explore topics. (See details in §5)

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: (1) 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/tuanlvm/ContraVis.

2 PROBLEM DEFINITIONS

In this problem 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 $\mathcal{D} = \{(d_1, s_1), \ldots, (d_N, s_N)\}$, where $s_n \in \mathcal{L}$ is the label of document d_n and $\mathcal{L}$ 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 $\mathcal{D}$ containing words from a finite vocabulary $\mathcal{V}$ and labels from set $\mathcal{L}$, 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.

	Notation	Description
Input Corpus	$\mathcal{D}$	document corpus
	N	number of documents, $\lvert\mathcal{D}\rvert = N$
	$\mathcal{V}$	vocabulary, $\lvert\mathcal{V}\rvert = W$
	d_n	a specific document
	M_n	number of words in document d_n
	$\mathcal{L}$	set of unique labels, $\lvert\mathcal{L}\rvert = L$
	s_n	single label of document d_n
	c	common label, which is applied to all documents
	Λ_n	$\Lambda_n := \{s_n, c\}$
Topics	K	total number of topics
	z	a specific latent topic
	β_z	word distribution of topic z
	$\boldsymbol{\beta}$	collection of β_z's for all topics
Visualization	d	visualization dimension (2 or 3)
	x_n	latent coordinate of doc. d_n in visualization space
	ϕ_z	latent coordinate of topic z in visualization space
	μ_l	latent coordinate of label l in visualization space
	$\boldsymbol{\chi}$	collection of x_n's for all documents
	$\boldsymbol{\Phi}$	collection of ϕ_z's for all topics
	$\boldsymbol{\mu}$	collection of μ_l for all labels

(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_l mutual topics per label $l \in \mathcal{L}$ and K_c 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_n of each document d_n and ϕ_z for each topic. In addition, each label l is also assumed to have a latent coordinate μ_l in the same visualization space. Let $\boldsymbol{\mu} = \{\mu_l\}_1^L$. The label distribution of d_n is expressed by its Euclidean distances to labels as follows:

$$\mathrm{P}(l|x_n, \boldsymbol{\mu}) = \frac{\exp(-\frac{1}{2}||x_n - \mu_l||^2)}{\sum_{l'=1}^{L} \exp(-\frac{1}{2}||x_n - \mu_{l'}||^2)} \tag{1}$$

According to Eq. (1), $\mathrm{P}(l|x_n, \boldsymbol{\mu})$ is high when x_n is close to μ_l. Since we observe the label s_n of document d_n, d_n should be placed close to μ_{s_n} in the visualization space so that $\mathrm{P}(s_n|x_n, \boldsymbol{\mu})$ is high. Note that each document also has a probability $\mathrm{P}(c|x_n, \boldsymbol{\mu})$ 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 $\mathrm{P}(z|l, x_n, \boldsymbol{\Phi})$, which is determined by:

$$\mathrm{P}(z|l, x_n, \boldsymbol{\Phi}) = \frac{\exp(-\frac{1}{2}||\mu_l - \phi_z||^2)\exp(-\frac{1}{2}||x_n - \phi_z||^2)}{\sum_{z'=1}^{K} \exp(-\frac{1}{2}||\mu_l - \phi_{z'}||^2)\exp(-\frac{1}{2}||x_n - \phi_{z'}||^2)} \tag{2}$$

We can see that for a label l, x_n has high $\mathrm{P}(z|l, x_n, \boldsymbol{\Phi})$ when it is close to topic ϕ_z and ϕ_z is close to label μ_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: $\mu_l \sim \mathrm{Normal}(0, \sigma_0^{-1} I)$
(2) For each topic $z = 1, \ldots, K$:
  (a) Draw z's word distribution: $\beta_z \sim \mathrm{Dirichlet}(\lambda)$
  (b) Draw z's coordinate: $\phi_z \sim \mathrm{Normal}(0, \varphi^{-1} I)$
(3) For each document d_n, where $n = 1, \ldots, N$:
  (a) Draw d_n's coordinate: $x_n \sim \mathrm{Normal}(\mu_{s_n}, \gamma^{-1} I)$
  (b) For each word $w_{nm} \in d_n$:
    (i) Draw a label: $l \sim \mathrm{Multi}(\{\mathrm{P}(l|x_n, \boldsymbol{\mu})\}_{l=1}^{L})$
    (ii) Draw a topic: $z \sim \mathrm{Multi}(\{\mathrm{P}(z|l, x_n, \boldsymbol{\Phi})\}_{z=1}^{K})$
    (iii) Draw a word: $w_{nm} \sim \mathrm{Multi}(\beta_z)$
  (c) Set Λ_n = unique labels of w_{nm}

Above, I is the $d \times d$ identity matrix. σ_0, φ, and γ are hyperparameters that control the variance of the Normal distributions. Similar to LDA [2], ContraVis associates each topic with a word probability distribution β_z where $\mathrm{P}(\beta_z) = \frac{\Gamma((\lambda+1)W)}{\Gamma(\lambda+1)^W} \prod_{w=1}^{W} \beta_{zw}^{\lambda}$ and λ 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 μ_{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 $\mathrm{P}(z|l, x_n, \boldsymbol{\Phi})$, w_{nm} is then drawn from multinomial β_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 $\boldsymbol{\beta} = \{\beta_z\}_{z=1}^{K}$, document coordinates $\boldsymbol{\chi} = \{x_n\}_{n=1}^{N}$, topic coordinates $\boldsymbol{\Phi} = \{\phi_z\}_{z=1}^{K}$, and label coordinates $\boldsymbol{\mu} = \{\mu_l\}_{l=1}^{L}$ (including common label c). Let $\Omega = \langle \boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\Phi}, \boldsymbol{\mu} \rangle$. 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

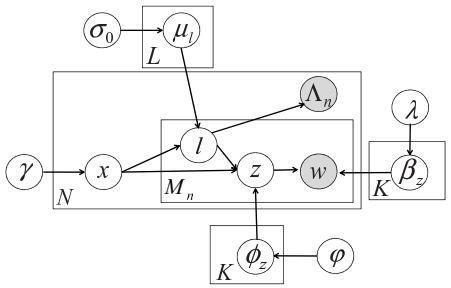

Figure 2: Graphical model representation of ContraVis's generative process (See §3.1).

to infer Ω from $\mathcal{D}$ based on maximum a posteriori estimation. Data likelihood can be written as follows (hyperparameters omitted for simplicity).

$$\begin{aligned}\mathrm{P}(\mathcal{D}|\Omega) &= \sum_{\mathcal{L}_\mathcal{W}}\sum_{\mathcal{Z}_\mathcal{W}} \mathrm{P}(\mathcal{D},\, \mathcal{L}_\mathcal{W},\, \mathcal{Z}_\mathcal{W}|\Omega) = \prod_{n=1}^{N} \mathrm{P}(\Lambda_n)\mathrm{P}(d_n|x_n,\, \boldsymbol{\mu},\, \boldsymbol{\Phi},\, \boldsymbol{\beta}) \\ &= \prod_{n=1}^{N}\Big[\mathrm{P}(\Lambda_n)\prod_{m=1}^{M_n}\sum_{l=1}^{L}\sum_{z=1}^{K}\big[\mathrm{P}(l|x_n,\, \boldsymbol{\mu})\mathrm{P}(z|l,\, x_n,\, \boldsymbol{\Phi})\mathrm{P}(w_{nm}|\beta_z)\big]\Big]\end{aligned} \tag{3}$$

where $\mathcal{L}_\mathcal{W}, \mathcal{Z}_\mathcal{W}$ are respectively label assignments and topic assignments to all the words $\mathcal{W}$ in $\mathcal{D}$, and

$$\mathrm{P}(\Lambda_n) = \prod_{m=1}^{M_n}\sum_{l\in\Lambda_n}\mathrm{P}(l|x_n,\, \boldsymbol{\mu}) \tag{4}$$

The EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying E and M steps that we outline below.

E step: The conditional expectation of the data loglikelihood with priors under the current estimate of the parameters $\hat{\Omega}$ is

$$\begin{aligned}C(\Omega|\hat{\Omega}) &= \mathrm{E}_{(\mathcal{L}_\mathcal{W},\mathcal{Z}_\mathcal{W}|\mathcal{D},\hat{\Omega})}[\log \mathrm{P}(\mathcal{D},\, \mathcal{L}_\mathcal{W},\, \mathcal{Z}_\mathcal{W}|\Omega)] \\ &= \sum_{n=1}^{N}\sum_{m=1}^{M_n} log \sum_{l\in\Lambda_n}\mathrm{P}(l|x_n,\, \boldsymbol{\mu}) \\ &+ \sum_{n=1}^{N}\sum_{m=1}^{M_n}\sum_{l=1}^{L}\sum_{z=1}^{K}\mathrm{P}(l,\, z|n,\, m,\, \hat{\Omega})\log\big[\mathrm{P}(l|x_n,\, \boldsymbol{\mu})\mathrm{P}(z|l,\, x_n,\, \boldsymbol{\Phi})\mathrm{P}(w_{nm}|\beta_z)\big] \\ &+ \sum_{n=1}^{N}\log\mathrm{P}(x_n) + \sum_{z=1}^{K}\log\mathrm{P}(\phi_z) + \sum_{z=1}^{K}\log\mathrm{P}(\beta_z) + \sum_{l=1}^{L}\log\mathrm{P}(\mu_l)\end{aligned} \tag{5}$$

where

$$\mathrm{P}(x_n) = \Big(\frac{\gamma}{2\pi}\Big)^{\frac{D}{2}}\exp\Big(-\frac{\gamma}{2}\parallel x_n - \mu_{s_n}\parallel^2\Big), \tag{6}$$

$$\mathrm{P}(\phi_z) = \Big(\frac{\varphi}{2\pi}\Big)^{\frac{D}{2}}\exp\Big(-\frac{\varphi}{2}\parallel \phi_z\parallel^2\Big), \tag{7}$$

$$\mathrm{P}(\mu_l) = \Big(\frac{\sigma_0}{2\pi}\Big)^{\frac{D}{2}}\exp\Big(-\frac{\sigma_0}{2}\parallel \mu_l\parallel^2\Big). \tag{8}$$

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 m'th word in the n'th document under current estimate of the parameters is given as

$$\mathrm{P}(l,\, z|n,\, m,\, \hat{\Omega}) = \frac{\mathrm{P}(l|\hat{x}_n)\mathrm{P}(z|l,\, x_n,\, \hat{\boldsymbol{\Phi}})\mathrm{P}(w_{nm}|\hat{\beta}_z)}{\sum_{l'=1}^{L}\sum_{z'=1}^{K}\mathrm{P}(l'|\hat{x}_n)\mathrm{P}(z'|l',\, x_n,\, \hat{\boldsymbol{\Phi}})\mathrm{P}(w_{nm}|\hat{\beta}_{z'})}. \tag{9}$$

M step: We update the entries of each β_z as follows:

$$\hat{\beta}_{zw} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M_n}\mathbb{1}(w_{nm}=w)\mathrm{P}(z|n,\, m,\, \hat{\Omega}) + \lambda}{\sum_{w'=1}^{W}\sum_{n=1}^{N}\sum_{m=1}^{M_n}\mathbb{1}(w_{nm}=w')\mathrm{P}(z|n,\, m,\, \hat{\Omega}) + \lambda W} \tag{10}$$

where $\mathbb{1}(X) = 1$ if X is true and 0 otherwise. As for the other parameters, x_n, ϕ_z, μ_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, ϕ_z, μ_l.

$$\begin{aligned}\frac{\partial C(\Omega|\hat{\Omega},\, g)}{\partial \phi_z} &= \sum_{n=1}^{N}\sum_{m=1}^{M_n}\sum_{l=1}^{L}\big(\mathrm{P}(l|n,\, m,\, \hat{\Omega})\mathrm{P}(z|l,\, x_n,\, \Phi) - \mathrm{P}(l,\, z|n,\, m,\, \hat{\Omega})\big)(2\phi_z - \mu_l - x_n) \\ &\quad - \varphi\phi_z\end{aligned} \tag{11}$$

$$\begin{aligned}\frac{\partial C(\Omega|\hat{\Omega},\, g)}{\partial x_n} &= M_n\Big(\sum_{l=1}^{L}(x_n-\mu_l)\mathrm{P}(l|x_n,\, \boldsymbol{\mu}) - \sum_{l=1}^{\Lambda_n}(x_n-\mu_l)\frac{\mathrm{P}(l|x_n,\, \boldsymbol{\mu})}{\sum_{l=1}^{\Lambda_n}\mathrm{P}(l|x_n,\, \boldsymbol{\mu})}\Big) \\ &\quad - \gamma(x_n - \mu_{s_n}) + \sum_{m=1}^{M_n}\sum_{l=1}^{L}\big(\mathrm{P}(l|x_n,\, \boldsymbol{\mu}) - \mathrm{P}(l|n,\, m,\, \hat{\Omega})\big)(x_n-\mu_l) \\ &\quad + \sum_{m=1}^{M_n}\sum_{l=1}^{L}\sum_{z=1}^{Z}\big(\mathrm{P}(l|n,\, m,\, \hat{\Omega})\mathrm{P}(z|x_n,\, l,\, \Phi) - \mathrm{P}(l,\, z|n,\, m,\, \hat{\Psi})\big)(x_n-\phi_z)\end{aligned} \tag{12}$$

$$\begin{aligned}\frac{\partial C(\Omega|\hat{\Omega},\, g)}{\partial \mu_l} &= -\sigma_0\mu_l + \sum_{n=1,\, l=s_n}^{N}\gamma(x_n-\mu_{s_n}) - \sum_{n=1,\, l\notin\Lambda_n}^{N}M_n\mathrm{P}(l|x_n,\, \boldsymbol{\mu})(x_n-\mu_l) \\ &\quad + \sum_{n=1,\, l\in\Lambda_n}^{N}M_n\Big(\frac{\mathrm{P}(l|x_n,\, \boldsymbol{\mu})}{\sum_{l'=1}^{\Lambda_n}\mathrm{P}(l'|x_n,\, \boldsymbol{\mu})} - \mathrm{P}(l|x_n,\, \boldsymbol{\mu})\Big)(x_n-\mu_l) \\ &\quad + \sum_{n=1}^{N}\sum_{m=1}^{M_n}\big(\mathrm{P}(l|n,\, m,\, \hat{\Omega}) - \mathrm{P}(l|x_n,\, \boldsymbol{\mu})\big)(x_n-\mu_l) \\ &\quad + \sum_{n=1}^{N}\sum_{m=1}^{M_n}\sum_{z=1}^{Z}\big(\mathrm{P}(l|n,\, m,\, \hat{\Omega})\mathrm{P}(z|x_n,\, l,\, \Phi) - \mathrm{P}(l,\, z|n,\, m,\, \hat{\Omega})\big)(\mu_l-\phi_z)\end{aligned} \tag{13}$$

Our estimated model parameters, $\Omega = \langle\boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\Phi}, \boldsymbol{\mu}\rangle$, respectively capture topic word distributions (Problem 1 *(i)*) as well as document, topic, and label coordinates (Problem 2). Using Ω 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.

$$\mathrm{P}(z|d_n,\, \Omega) = \sum_{l=1}^{L}\mathrm{P}(l|x_n)\mathrm{P}(z|l,\, x_n,\, \boldsymbol{\Phi}) \tag{14}$$

$$\mathrm{P}(z|l,\, \Omega) = \frac{\sum_{n=1}^{N}\mathrm{P}(d_n)\mathrm{P}(l|x_n)\mathrm{P}(z|l,\, x_n,\, \boldsymbol{\Phi})}{\sum_{n=1}^{N}\mathrm{P}(d_n)\mathrm{P}(l|x_n)} \tag{15}$$

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

3.2.1 Quantifying Common and Discriminative Topics.

Based on $p(z|l)$ as in Eq. (15), we define $cmn\_score = p(z|c)$ and $disc\_score = \frac{p(z|l)}{max_{l'\neq l}p(z|l')}$ 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 $\mathcal{D}$ 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.

$$\mathrm{P}(\mathcal{D}|\boldsymbol{\mu}, \Phi, \boldsymbol{\beta}) = \prod_{n=1}^{N} \int_{x_n} \mathrm{P}(\Lambda_n)\, \mathrm{P}(d_n|x_n, \boldsymbol{\mu}, \Phi, \boldsymbol{\beta})\, \mathrm{P}(x_n) dx_n \,, \qquad (16)$$

where $\mathrm{P}(d_n|x_n, \boldsymbol{\mu}, \Phi, \boldsymbol{\beta})$ 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 [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:

$$\mathrm{P}(\mathcal{D}|\boldsymbol{\mu}, \Phi, \boldsymbol{\beta}) \simeq \prod_{n=1}^{N} \Big[\frac{1}{S} \sum_{s} \mathrm{P}(\Lambda_n)\, \mathrm{P}(d_n|x_n^{(s)}, \boldsymbol{\mu}, \Phi, \boldsymbol{\beta})\Big] \,. \qquad (17)$$

3.2.3 **Computational Complexity.** In ContraVis inference algorithm, E-step is performed in $O(NLKW)$ where we evaluate Eq. (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 $O(qp)$ plus the cost to evaluate $C(\Omega|\hat{\Omega})$ in Eq. (5) and its gradients [17]. Here, q is the number of steps stored in the memory and $p = (N + L + K)d$ is the number of variables. The evaluations of $C(\Omega|\hat{\Omega})$ and its gradients can be performed in $O(NLKW)$. Since modest values of q are often between 3 and 20 [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.[1] 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 $L = 9$ categories: `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.[2] 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`.

[1] Publicly available at https://catalog.ldc.upenn.edu/ldc2008t19

[2] Publicly available at https://github.com/tuanlvm/ContraVis/tree/master/data

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

	#labels L	#documents N	# words W
NYTnews	9	515252	8666
PolForum	8	18732	6774
PeerReview	4	5336	4032

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*[3][16] (s for supervised)
(2) (Supervised) *DiscLDA*[4] [9] (Disc for discriminative)
(3) (Unsupervised): *LDA*[5] [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 [14] to embed them into 2-d. (See for e.g., Fig. 8b)

(4) (Unsupervised & Joint (Topics+Vis)): *PLSV*[6] [7]
(5) (Supervised & Joint (Topics+Vis)): ContraVis [this paper].

Table 3: Comparison of methods by properties.

Properties vs. Methods	*LDA*	*sLDA*, *DiscLDA*	*PLSV*	ContraVis
Contrastive (Supervised)		✔		✔
Visually-expressive			✔	✔
Unified/Jointly modeled			✔	✔

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.

[3] We use author implementation at https://github.com/blei-lab/class-slda. This is a variant of *sLDA* with a categorical response [5].

[4] We obtained the implementation from the author Prof. Simon Lacoste-Julien. In our experiments, we use *DiscLDA* with fixed transformation matrix T.

[5] http://scikit-learn.org (sklearn.decomposition.LatentDirichletAllocation)

[6] We use the implementation at 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 (in 2-d) for our ContraVis (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).

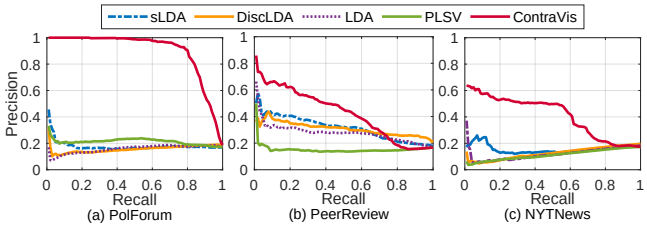

Figure 3: Precision-Recall curves (avg'ed over 10 samples) for task *Common*.

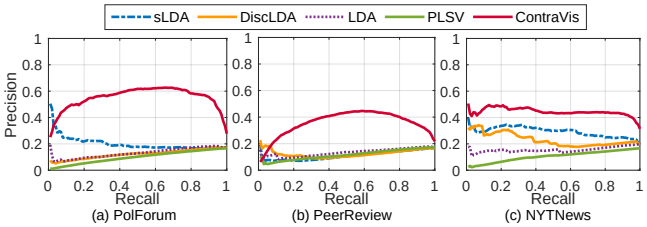

Figure 4: Precision-Recall curves (avg'ed over 10 samples) for task *Discriminative*.

Table 4: Mean Average Precision (MAP) (avg'ed across 10 samples) ± stan.dev. on task *Common*. ▲ (p<0.001) and △ (p<0.005) are cases where ContraVis is significantly better than the baseline w.r.t. the paired Wilcoxon signed rank test.

	ContraVis	*sLDA*	*DiscLDA*	*LDA*	*PLSV*
PolForum	**0.897**±0.03	0.178±0.04▲	0.157±0.08▲	0.158±0.03▲	0.213±0.14▲
PeerReview	**0.423**±0.16	**0.314**±0.10	**0.307**±0.08	**0.276**±0.09	0.152±0.02▲
NYTnews	**0.410**±0.24	0.156±0.03△	0.126±0.03▲	0.121±0.02△	0.112±0.01▲

Table 5: Mean Average Precision (MAP) of methods (avg'ed across 10 samples) ± stan.dev. on task *Discriminative*.

	ContraVis	*sLDA*	*DiscLDA*	*LDA*	*PLSV*
PolForum	**0.548**±0.07	0.201±0.08▲	0.122±0.02▲	0.132±0.02▲	0.099±0.00▲
PeerReview	**0.356**±0.07	0.111±0.01▲	0.125±0.03▲	0.134±0.03▲	0.116±0.01▲
NYTnews	**0.444**±0.13	**0.293**±0.18	**0.232**±0.19	0.156±0.04▲	0.105±0.01▲

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 total tf-idf values. As such, this ranking is obtained based on in-corpus information.

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.[7] For 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[8], in the same fashion as the above.

[7] https://catalog.ldc.upenn.edu/ldc2006t13
[8] https://webscope.sandbox.yahoo.com/catalog.php?datatype=l

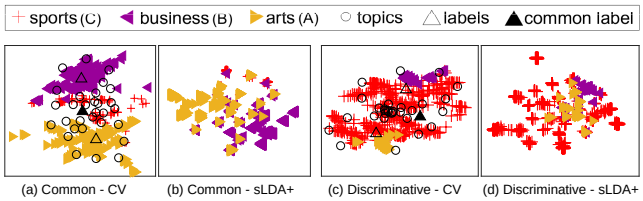

Figure 5: (best in color) Visualization of documents from NYTnews (K=30); *Common* task: (a) ContraVis (CV) places docs with common topics (`sports`) near the common label, (b) *sLDA+*: it is not as easy to spot common docs; *Discriminative* task: (c) CV places docs with discr. topics (`business`, `arts`) 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 $p(z|l) = \frac{\sum_{d_n:s_n=l} p(z|d_n)}{\sum_{z=1}^{K}\sum_{d_n:s_n=l} p(z'|d_n)}$. For *sLDA* and *DiscLDA*, $p(w|l) \propto \sum_{n=1}^{N}\sum_{z=1}^{K} p(w|z)p(z|d_n)p(l|d_n)p(d_n)$ with $p(d_n) = \frac{1}{N}$, where $p(l|d_n)$ 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: $NDCG = \frac{DCG@t}{IDCG@t}$. $DCG@t = \sum_{i=1}^{t} \frac{rel_i}{\log_2(i+1)}$ (the higher, the better), where rel_i is the "true" relevance of top ith 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 $NDCG@t$ 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.

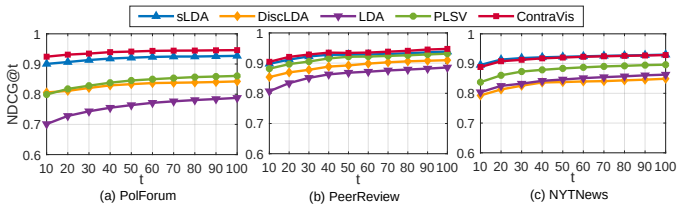

Figure 6: $p(w|l)$ ranking quality w.r.t. tf-idf ground-truth and $NDCG@t$ at varying t: number of top words. Results are averaged across 10 samples and all labels for each dataset.

Table 6: $p(w|l)$ ranking quality w.r.t $NDCG@100$ avg'ed across 10 samples and all labels from each dataset (K = 30). Symbols ▲ (p<0.001) and △ (p<0.005) denote the cases where ContraVis is significantly better than the baseline w.r.t. the paired Wilcoxon signed rank test.

	ContraVis	*sLDA*	*DiscLDA*	*LDA*	*PLSV*
PolForum					
tf-idf	**0.9460**±0.00	0.9270±0.01▲	0.8424±0.01▲	0.7878±0.07▲	0.8605±0.01▲
Google	**0.5494**±0.02	0.5151±0.03▲	0.4681±0.02▲	0.3770±0.06▲	0.4562±0.01▲
Yahoo	**0.5635**±0.02	0.5291±0.03△	0.4806±0.01▲	0.3875±0.06▲	0.4686±0.02▲
PeerReview					
tfidf	**0.9471**±0.00	0.9382±0.00▲	0.9096±0.01▲	0.8855±0.04▲	0.9313±0.00▲
Google	**0.2388**±0.01	0.2128±0.02▲	0.1805±0.02▲	0.2001±0.02▲	0.1971±0.02▲
Yahoo	**0.2461**±0.01	0.2193±0.02▲	0.1855±0.02▲	0.2061±0.02△	0.2022±0.02▲
NYTnews					
tf-idf	**0.9286**±0.00	**0.9300**±0.01	0.8490±0.01▲	0.8628±0.04▲	0.8962±0.01▲
Google	**0.2483**±0.03	**0.2384**±0.00	0.1947±0.01▲	**0.2348**±0.03	0.2119±0.01△
Yahoo	**0.2574**±0.03	**0.2479**±0.02	0.2062±0.01▲	**0.2450**±0.03	0.2219±0.01△

Table 7: Relative $p(w|l)$ ranking quality comparison w.r.t. $NDCG@100$ (K = 30). ≫ denotes p<0.001, > denotes p<0.005 and ≈ denotes indifference w.r.t. the paired Wilcoxon signed rank test over 10 samples and all labels per dataset.

Dataset	GTruth	Relative Order
PolForum	tf-idf	CV ≫ *sLDA* ≫ *PLSV* ≫ *DiscLDA* > *LDA*
	Google	CV ≫ *sLDA* > *DiscLDA* ≈ *PLSV* ≫ *LDA*
	Yahoo	CV > *sLDA* > *DiscLDA* ≈ *PLSV* ≫ *LDA*
PeerReview	tf-idf	CV ≫ *sLDA* ≫ *PLSV* ≫ *DiscLDA* ≈ *LDA*
	Google	CV ≫ *sLDA* > *PLSV* ≫ *DiscLDA*; CV≫*LDA*
	Yahoo	CV ≫ *sLDA* > *PLSV* ≫ *DiscLDA*; CV>*LDA*
NYTnews	tf-idf	CV ≈ *sLDA* ≫ *PLSV* ≫ *DiscLDA*; CV≫*LDA*
	Google	CV ≈ *sLDA* > *PLSV* > *DiscLDA*; CV≈*LDA*
	Yahoo	CV ≈ *sLDA* > *PLSV* > *DiscLDA*; CV≈*LDA*

Table 8: Top 5 most relevant/discriminative words for a selection of three labels from each dataset w.r.t. $p(w|l)$ by ContraVis closely agree with human understanding.

Dataset	Label	Top 5 Words
PolForum	`abortion`	abort, human, life, peopl, person
	`taxes`	tax, govern, incom, pay, will
	`race`	black, white, peopl, race, racist
PeerReview	`operations`	help, citi, idea, busi, peopl
	`engineering`	help, project, engin, issu, good
	`people-analytics`	candid, help, recruit, hire, manag
NYTnews	`sports`	game, team, play, season, year
	`business`	compani, year, percent, will, market
	`real-estate`	build, hous, year, number, apart

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).

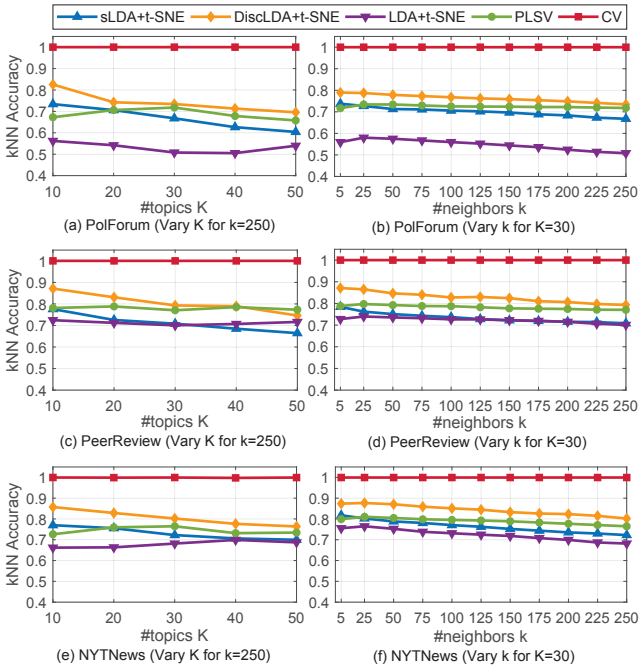

Figure 7: kNN classification accuracy in 2-d (avg'ed across 10 random samples per dataset) on (from top to bottom) PolForum, PeerReview, and NYTnews. ContraVis (denoted as CV) consistently and significantly outperforms competition in all settings.

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 one 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

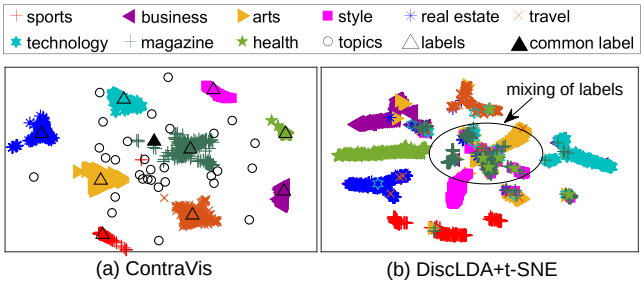

Figure 8: (best in color) 2-d visualization of documents from NYTnews ($K = 30$); (left) ContraVis embedding well-separates documents with different labels, in contrast (right) *DiscLDA*+tSNE embedding is not as pure.

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

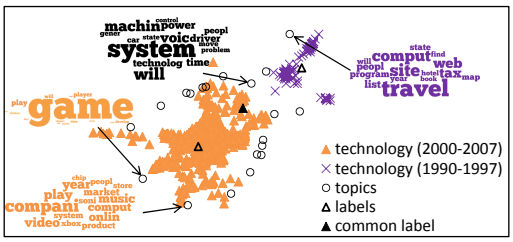

Figure 9: (best in color) Comparing `technology` articles from different time periods in NYTnews using ContraVis (K=20).

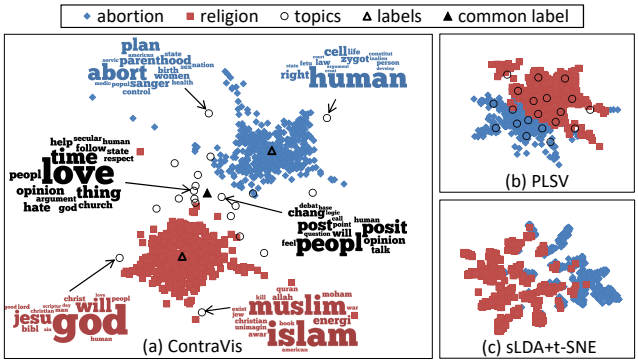

Figure 10: (best in color) Contrastive visualization of `abortion` and `religion` posts from PolForum (K=20); (left) ContraVis embedding well-separates the collections and two most discriminative topics per label and two common topics shown are intuitive, while (right) *PLSV* and *sLDA*+tSNE embeddings are not on par.

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 from `taxes`, `military` and `gun-control` threads. We see that a common topic has frequently used words such as 'govern', 'right', 'rule' and 'power'. Also shown are one discriminative topic for each thread, which are right on subject and coherent. For example, topic exclusive to `gun-control` has representative words 'gun', 'furious', 'law', 'firearm' and 'atf'. ATF is a United States law enforcement agency for regulating use of firearms. In the embedding, `taxes` and `gun-control` posts are close to each other which suggests that they may be more related to each other than they are to `military`.

5.3 PeerReview

Finally, we present a case study where we compare reviews of employees from different departments in PeerReview. Since documents are reviews, we expect to see very frequent words like 'help(ful)', 'good' and 'idea' in the topics. In spite of

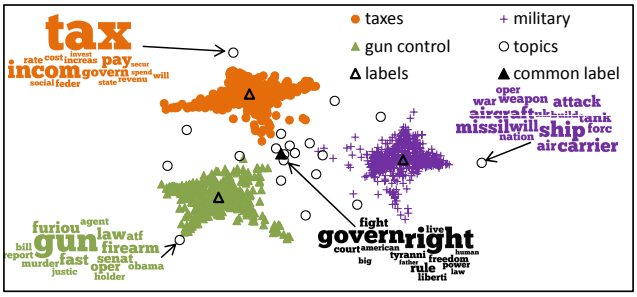

Figure 11: (best in color) Comparing posts on `gun-control`, `taxes`, & `military` from PolForum using ContraVis (K=20).

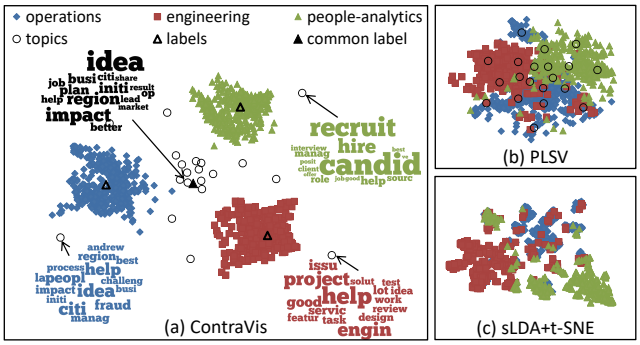

Figure 12: (best in color) Contrastive visualization of reviews for employees from different departments in PeerReview (K=20); (left) In ContraVis embedding, departments are well-separated and disc. topics reflect the role of each department; (right) *PLSV* and *sLDA*+tSNE embeddings mix the labels. *sLDA*+tSNE embedding is doc-only (no topics).

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 in Figure 12(b) and 12(c).

Through 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 as well as 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[9], 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.

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

REFERENCES

[1] Eric Alexander and Michael Gleicher. 2016. Task-driven comparison of topic models. *IEEE transactions on visualization and computer graphics* 22, 1 (2016).

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

[3] Quim Castella and Charles Sutton. 2014. Word storms: Multiples of word clouds for visual comparison of documents. In *Proceedings of the 23rd international conference on World wide web*. ACM, 665–676.

[4] Changyou Chen, Wray L. Buntine, Nan Ding, Lexing Xie, and Lan Du. 2015. Differential Topic Models. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 2 (2015).

[5] Wang Chong, David Blei, and Fei-Fei Li. 2009. Simultaneous image classification and annotation. In *CVPR*. 1903–1910.

[6] Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *UAI*. 289–296.

[7] Tomoharu Iwata, Takeshi Yamada, and Naonori Ueda. 2008. Probabilistic latent semantic visualization: topic model for visualizing documents. In *KDD*. 363–371.

[8] Hannah Kim, Jaegul Choo, Jingu Kim, Chandan K Reddy, and Haesun Park. 2015. Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization. In *KDD*. 567–576.

[9] Simon Lacoste-Julien, Fei Sha, and Michael I Jordan. 2009. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *NIPS*. 897–904.

[10] Tuan M. V. Le and Hady W. Lauw. 2014. Probabilistic latent document network embedding. In *ICDM*. 270–279.

[11] Tuan M. V. Le and Hady W. Lauw. 2014. Semantic visualization for spherical representation. In *KDD*. 1007–1016.

[12] Tuan M. V. Le and Hady W. Lauw. 2016. Semantic visualization with neighborhood graph regularization. *Jour. of AI Res.* 55 (2016), 1091.

[13] Tuan M. V. Le and Hady W. Lauw. 2016. Word Clouds with Latent Variable Analysis for Visual Comparison of Documents. In *IJCAI*.

[14] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

[15] Adrian Mayorga and Michael Gleicher. 2013. Splatterplots: Overcoming overdraw in scatter plots. *IEEE transactions on visualization and computer graphics* 19, 9 (2013), 1526–1538.

[16] Jon D Mcauliffe and David M Blei. 2008. Supervised topic models. In *NIPS*.

[17] J. Nocedal and S. J. Wright. 2006. *Numerical Optimization*. Springer.

[18] Daniela Oelke, Hendrik Strobelt, Christian Rohrdantz, Iryna Gurevych, and Oliver Deussen. 2014. Comparative exploration of document collections: a visual analytics approach. In *Computer Graphics Forum*, Vol. 33. Wiley Online Library.

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

[20] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*. 248–256.

[21] Daniel Ramage, Christopher D Manning, and Susan Dumais. 2011. Partially labeled topic models for interpretable text mining. In *KDD*. 457–465.

[22] Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J Mooney. 2010. Spherical topic models. In *ICML*. 903–910.

[23] Xiang Ren, Yuanhua Lv, Kuansan Wang, and Jiawei Han. 2017. Comparative Document Analysis for Large Text Corpora.. In *WSDM*. ACM, 325–334.

[24] Zhaochun Ren and Maarten de Rijke. 2015. Summarizing Contrastive Themes via Hierarchical Non-Parametric Processes.. In *SIGIR*. ACM, 93–102.

[25] Danielle Albers Szafir, Deidre Stuffer, Yusef Sohail, and Michael Gleicher. 2016. Textdna: Visualizing word usage with configurable colorfields. In *Computer Graphics Forum*, Vol. 35. Wiley Online Library, 421–430.

[26] Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. 2016. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 287–297.

[27] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *ICML*. 1105–1112.

[28] ChengXiang Zhai, Atulya Velivelli, and Bei Yu. 2004. A Cross-collection Mixture Model for Comparative Text Mining. In *KDD*. 743–748.

[29] Jun Zhu, Amr Ahmed, and Eric P Xing. 2012. MedLDA: maximum margin supervised topic models. *Journal of Machine Learning Research* 13, Aug (2012), 2237–2278.

[30] James Y. Zou, Daniel J. Hsu, David C. Parkes, and Ryan Prescott Adams. 2013. Contrastive Learning Using Spectral Methods.. In *NIPS*. 2238–2246.