A Disentangled Adversarial Neural Topic Model for Separating Opinions from Plots in User Reviews

Gabriele Pergola, Lin Gui, Yulan He
Department of Computer Science, University of Warwick, UK
{gabriele.pergola, lin.gui, yulan.he}@warwick.ac.uk

Abstract

The flexibility of the inference process in Variational Autoencoders (VAEs) has recently led to revising traditional probabilistic topic models giving rise to Neural Topic Models (NTM). Although these approaches have achieved significant results, surprisingly very little work has been done on how to disentangle the latent topics. Existing topic models when applied to reviews may extract topics associated with writers' subjective opinions mixed with those related to factual descriptions such as plot summaries in movie and book reviews. It is thus desirable to automatically separate opinion topics from plot/neutral ones enabling a better interpretability. In this paper, we propose a neural topic model combined with adversarial training to disentangle opinion topics from plot and neutral ones. We conduct an extensive experimental assessment introducing a new collection of movie and book reviews paired with their plots, namely MOBO dataset, showing an improved coherence and variety of topics, a consistent disentanglement rate, and sentiment classification performance superior to other supervised topic models.

1 Introduction

Variational Autoencoders (VAEs) (Kingma et al., 2014) allow to design complex generative models of data. In the wake of the renewed interest for VAEs, traditional probabilistic topic models (Blei et al., 2003) have been revised giving rise to several Neural Topic Model (NTM) variants, such as NVDM (Miao et al., 2016), ProdLDA (Srivastava et al., 2017), and NTM-R (Ding et al., 2018). Although these approaches have achieved significant results via the neural inference process, existing topic models when applied to user reviews may extract topics with writers' subjective opinions mixed with those related to factual descriptions such as plot summaries of movies and books (Lin et al., 2012). Yet surprisingly very little work has been done on how to disentangle the inferred topic representations.

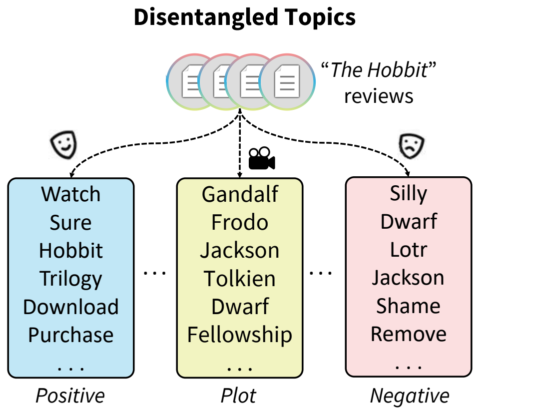

Figure 1: Disentangled topics extracted by DIATOM from the Amazon reviews for "The Hobbit".

Disentangled representations can be defined as representations where individual latent units are sensitive to variations of a single generative factor, while being relatively invariant to changes of other factors (Bengio et al., 2013; Higgins et al., 2017). Inducing such representations has been shown significantly beneficial for their generalization and interpretability (Achille and Soatto, 2018; Peng et al., 2019). In the context of topic modeling, we propose to consider latent topics as generative factors to be disentangled to improve their interpretability and discriminative power. Disentangled topics are topics invariant to the factors of variation of text, which for instance, in the context of book and movie reviews could be the author's opinion (e.g. positive/negative), the salient parts of a plot or other auxiliary information reported. An illustration of this is shown in Figure 1 in which opinion topics are separated from plot topics.

However, models relying solely on sentiment information are easily misled and not suitable to disentangle opinion from plots, since even plot descriptions frequently make large use of sentiment expressions (Pang and Lee, 2004a). Consider for example the following sentence: "The ring holds a

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2870–2883
June 6–11, 2021.

dark power, and it soon begins to exert its *evil influence* on Bilbo", an excerpt from a strong positive Amazon's review.

Therefore, we propose to distinguish opinion-bearing topics from plot/neutral ones combining a neural topic model architecture with adversarial training. In this study, we present the DIsentangled Adversarial TOpic Model (DIATOM)[1], aiming at disentangling information related to the target labels (i.e. the review score), from other distinct aspects yet possibly still polarised (e.g. plot descriptions). We also introduce a new dataset, namely the MOBO dataset[1], made up of movie and book reviews, paired with their related plots. The reviews come from different publicly available datasets: IMDB (Maas et al., 2011), GoodReads (Wan et al., 2019) and Amazon reviews (McAuley et al., 2015), and encompass a wide spectrum of domains and styles. We conduct an extensive experimental assessment of our model. First, we assess the topic quality in terms of topic coherence and diversity and compare DIATOM with other supervised topic models on the sentiment classification task; then, we analyse the disentangling rate of topics to quantitatively assess the degree of separation between actual opinion and plot/neutral topics.
Our contributions are summarized below:

- We propose a new model, DIATOM, which is able to generate disentangled topics through the combination of VAE and adversarial learning;
- We introduce the MOBO dataset, a new collection of movie and book reviews paired with their plots;
- We conduct an experimental assessment of our model, highlighting more interpretable topics with better topic coherence and diversity scores compared to others state-of-the-art supervised topic models, and improved discriminative power on sentiment classification, and a consistent topic-disentanglement rate.

2 Related Work

Our work is closely related to three lines of research: sentiment-topic models, neural topic models and learning disentangled representations.

Sentiment-Topic Models. Probabilistic graphical models for topic extraction have been extensively studied. Beyond LDA (Blei et al., 2003), a wide spectrum of models has specialised it to more particular tasks using contextual information (Rosen-Zvi et al., 2004; Wang et al., 2006; Blei et al., 2006; Pergola et al., 2018, 2019). Supervised-LDA (sLDA) (Mcauliffe et al., 2008) is a general-purpose supervised extension which builds on top of LDA by adding a response variable associated with each document (e.g. a review's rating). A category of extensions particularly relevant for this work is the sentiment-topic models. Examples include the Joint Sentiment-Topic (JST) model (Lin and He, 2009; Lin et al., 2012) and Aspect and Sentiment Unification Model (ASUM) (Jo and Oh, 2011). These models are able to extract informative topics grouped under different sentiment classes. Although they do not rely on document labels, they require word prior polarity information to be incorporated into the learning process in order to generate consistent results. Nevertheless, The possibility to supervise the learning process with document labels makes JST suitable for a fair comparison. Compared to DIATOM, the discussed sentiment topic models can only distinguish between *polarity*-bearing topics and neutral ones, remaining strictly aligned to the provided labels. Instead, DIATOM is able to generate opinion-bearing topics and plot topics which may still be polarized but not carrying any user's opinion.

Neural Topic Models. Neural models provide a more generic and extendable alternative to topic modeling, and therefore, have recently gained increasing interest. Some of them use belief networks (Mnih et al., 2014), or enforce the Dirichlet prior on the document-topic distribution via Wasserstein Autoencoders (Nan et al., 2019). Others adopt continuous representations to capture long-term dependencies or preserve word order via sequence-to-sequence VAE (Dieng et al., 2017; Xu et al., 2017; Bowman et al., 2016; Yang et al., 2017) whose time complexity and difficulty of training, however, have limited their applications. Neural Variational Document Model (NVDM) (Miao et al., 2016) is a direct extension of VAE used for topic detection in text. In NVDM, the prior of latent topics is assumed to be a Gaussian distribution. This is not ideal since it cannot mimic the simplex in the latent topic space. To address this problem, LDA-VAE (Srivastava et al., 2017) instead used the logistic normal distribution to approximate the Dirichlet distribution. ProdLDA (Srivastava et al., 2017) extended LDA-VAE by replacing the mixture model of LDA with a product of experts. SCHOLAR is a

[1]Source code and dataset available at: https://github.com/gabrer/diatom.

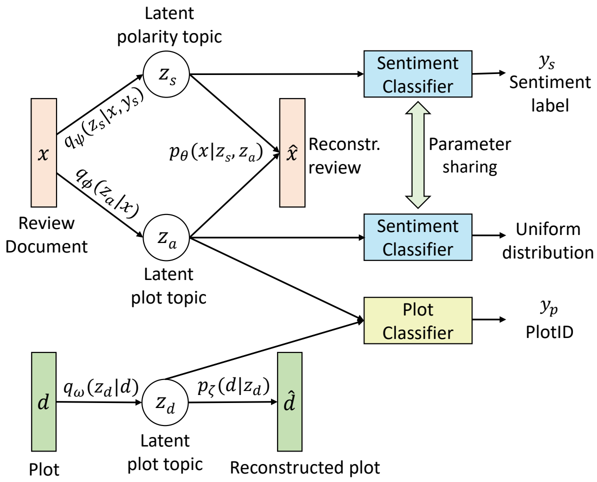

Figure 2: The DIATOM Architecture.

neural framework for topic models with metadata incorporation (Card et al., 2018). When metadata are document labels, the model infers topics that are relevant to those labels. Although some studies have applied the adversarial approach (Goodfellow et al., 2014) and reinforcement learning (Gui et al., 2019) to topic models setting a Dirichlet prior on the generative network (Wang et al., 2019; Masada et al., 2018), it is still unexplored how to use this mechanism to disentangle opinion-bearing topics from plot or neutral topics.

Representation Disentanglement. Among the slightly different versions of representation disentanglement proposed (Bengio et al., 2013; Higgins et al., 2018; Gao et al., 2019), the one achieved in DIATOM is analogous to Thomas et al. (2017) and Bengio et al. (2017), where they impose additional constraints to the representations controlled using a reinforcement learning mechanism determining the disentangled factors. Alternatively, in DIATOM we make use of an adversarial approach over the available target labels. Application in text processing has shown promising results (John et al., 2019; Kumar et al., 2017; Hoang et al., 2019; Esmaeili et al., 2019), yet applications to topic modeling are still limited (Wilson et al., 2016) and to the best of our knowledge, there is no work in separating opinion-bearing topics from plot/neutral topics.

3 DIATOM architecture

Our proposed DIATOM model is shown in Figure 2. Assuming a document $\boldsymbol{x}$ is associated with a sentiment label y_s, and each document can be represented by latent topics associated with sentiments ($\boldsymbol{z}_s$) and plots[2] ($\boldsymbol{z}_a$), we aim to learn a model maximizing the joint data-label log-likelihood, $\log p(\boldsymbol{x}, y_s)$:

$$\begin{aligned}\log p(\boldsymbol{x}, y_s) &= \log \int\int p(\boldsymbol{x}, y_s, \boldsymbol{z}_a, \boldsymbol{z}_s) d\boldsymbol{z}_a d\boldsymbol{z}_s \\ &\geq \mathbb{E}_{q_\phi(\boldsymbol{z}_a|\boldsymbol{x}), q_\psi(\boldsymbol{z}_s|\boldsymbol{x}, y_s)}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}_a, \boldsymbol{z}_s)] \\ &+ \mathbb{E}_{q_\phi(\boldsymbol{z}_a|\boldsymbol{x}), q_\psi(\boldsymbol{z}_s|\boldsymbol{x}, y_s)}[\log p_\pi(y_s|\boldsymbol{x})] \\ &- \mathrm{KL}\left(q_\phi(\boldsymbol{z}_a|\boldsymbol{x})||p(\boldsymbol{z}_a)\right) \\ &- \mathrm{KL}\left(q_\psi(\boldsymbol{z}_s|\boldsymbol{x}, y_s)||p(\boldsymbol{z}_s)\right) \qquad (1)\end{aligned}$$

Inspired by Miao et al. (2016) and Card et al. (2018), we assume the document-level topic distribution for plots can be approximated by a multi-layer perceptron (MLP) taking as input a multivariate Gaussian distribution, and similarly for the topic distribution for sentiments. The multinomial distribution over words under a plot topic and an opinion topic can be parametrised by a weight matrix $\boldsymbol{W}$. The generative process is shown below.

- For each document $d \in \{1, .., D\}$,
  - Draw the latent plot-topics,
    $\hat{\phi} \sim \mathcal{N}(\mu_\phi, \Sigma_\phi), \quad \boldsymbol{z}_a = f_{\hat{\phi}}(\hat{\phi})$
  - Draw the latent opinion-topics,
    $\hat{\psi} \sim \mathcal{N}(\mu_\psi, \Sigma_\psi), \quad \boldsymbol{z}_s = f_{\hat{\psi}}(\hat{\psi})$
  - For each word $n \in \{1, .., N_d\}$ in d,
    - Draw $x_{d,n} \sim p(x_{d,n}|\boldsymbol{W}, \boldsymbol{z}_a, \boldsymbol{z}_s)$
  - Generate the document-level sentiment label, $y_s \sim p(y_s|f_y(\boldsymbol{z}_s))$

where $f_{\hat{\phi}}$, $f_{\hat{\psi}}$ and f_y are MLPs, $\boldsymbol{z}_a$ is a K-dimensional latent topic representation of plots for document d, $\boldsymbol{z}_s$ is a S-dimensional latent topic representation of sentiments for document d. The probability of word $x_{d,n}$ can be parametrised by another network:

$$p(x_{d,n}|\boldsymbol{W}, \boldsymbol{z}_a, \boldsymbol{z}_s) \propto \exp\left(\boldsymbol{m}_d + \boldsymbol{W} \cdot (\boldsymbol{z}_a \parallel \boldsymbol{z}_s)\right) \qquad (2)$$

where $\boldsymbol{m}_d$ is the V-dimensional background log-frequency word distribution, and $\boldsymbol{W} \in \mathbb{R}^{V \times (K+S)}$, while $\boldsymbol{z}_a \parallel \boldsymbol{z}_s$ is the concatenation of the two latent topic vectors.

Plot Inference Network. Following the idea of VAE which computes a variational approximation to an intractable posterior using MLPs, we define two inference networks f_{μ_ϕ} and f_{Σ_ϕ} which takes as input the word counts in documents:

$$\mu_\phi = f_{\mu_\phi}(\boldsymbol{x}) \quad \Sigma_\phi = \mathrm{diag}(f_{\Sigma_\phi}(\boldsymbol{x})) \qquad (3)$$

[2]These are the topics not associated with the target sentiments, which can be either plot topics or neutral topics (not about plots). For notational convenience, we call both plot topics.

The outputs of both networks are vectors in $\mathbb{R}^K$. Here, 'diag' converts a column vector to a diagonal matrix. For a document $\boldsymbol{x}$, $q(\phi) \simeq \mathcal{LN}(\mu_\phi, \Sigma_\phi)$. With such a formulation, we can generate samples from $q(\phi)$ by first sampling $\epsilon \sim \mathcal{N}(0, I)$ and then computing $\hat{\phi} = \sigma(\mu_\phi + \Sigma_\phi{}^{1/2}\epsilon)$.

Sentiment Inference Network. Similarly, to compute a variational approximation to $q(\psi)$, we define two inference networks f_{μ_ψ} and f_{Σ_ψ} which takes as input the word counts in documents:

$$\mu_\psi = f_{\mu_\psi}(\boldsymbol{x}) \quad \Sigma_\psi = \text{diag}(f_{\Sigma_\psi}(\boldsymbol{x})) \tag{4}$$

The outputs of both networks are vectors in $\mathbb{R}^S$. For a document $\boldsymbol{x}$, $q(\psi) \simeq \mathcal{LN}(\mu_\psi, \Sigma_\psi)$. We then generate samples from $q(\psi)$ by first sampling $\epsilon \sim \mathcal{N}(0, I)$ and then computing $\hat{\psi} = \sigma(\mu_\psi + \Sigma_\psi{}^{1/2}\epsilon)$.

Overall Objective. With the sampled $\hat{\phi}$ and $\hat{\psi}$, for each document $\boldsymbol{x}$, we can minimise the reconstruction loss with a Monte Carlo approximation using L independent samples:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{x}} \approx \frac{1}{L}\sum_{l=1}^{L}\sum_{n=1}^{N_d} \log p_\theta(x_{d,n}|\hat{\phi}^{(l)}, \hat{\psi}^{(l)}) \\ - \text{KL}\left(q(\boldsymbol{z}_a|\boldsymbol{x}) \;||\; p(\boldsymbol{z}_a)\right) \\ - \text{KL}\left(q(\boldsymbol{z}_s|\boldsymbol{x}, y_s) \;||\; p(\boldsymbol{z}_s)\right)\end{aligned} \tag{5}$$

where the first term in the RHS is given by Eq. (2). It has been previously shown in Kingma et al. (2014), if a standard multivariate normal prior is placed on the latent variables $\boldsymbol{z}_a$ and $\boldsymbol{z}_s$, then there is a closed form solution to the KL divergence terms above.

We assume that the latent topics associated with plots, $\boldsymbol{z}_a$, are independent of sentiment classes, and hence, when fed into a sentiment classifier, should generate a uniform sentiment class distribution (similar to adversarial learning). On the contrary, the latent topics associated with sentiments, $\boldsymbol{z}_s$, should bear essential information to discriminate between sentiment classes. Therefore, we define the following two objectives for sentiment classification; the former being the expected KL divergence with the uniform distribution $\mathcal{U}$, and the latter a cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{adv} &= -\mathbb{E}_{q_\phi(\boldsymbol{z}_a)}\left[\text{KL}\left(\mathcal{U}(0, M) \;||\; p(\hat{y}|\boldsymbol{z}_a)\right)\right] \\ \mathcal{L}_{sent} &= -\mathbb{E}_{q_\psi(\boldsymbol{z}_s)}\sum_{c=1}^{M} y_c \log\left(p(\hat{y}_c|\boldsymbol{z}_s)\right)\end{aligned} \tag{6}$$

where M is the total number of sentiment classes, and $\mathcal{U}(0, M)$ represents the uniform sentiment class distribution.

To further disentangle the latent topics associated with plots and sentiments, while concurrently minimise the redundancy in the final topic matrix, we apply an orthogonal regularizer over the decoder matrix $\boldsymbol{W}$. $\mathcal{L}_{orth}$ reaches its minimum value when the dot product between different topic distributions goes close to zero:

$$\mathcal{L}_{orth} = ||\, \boldsymbol{W} \cdot \boldsymbol{W}^T - \mathbb{I}\, || \tag{7}$$

Our final objective function is:

$$\mathcal{L} = -\alpha\mathcal{L}_{\boldsymbol{x}} + \beta\mathcal{L}_{adv} + \gamma\mathcal{L}_{sent} + \eta\mathcal{L}_{orth} \tag{8}$$

where α, β, γ and η control the relative contribution of various loss functions.

Plot Network. An additional VAE is plugged to the model providing a supplementary signal for the latent plot topic extraction. This mechanism preserves the plot information that might contain some sentiment words and thus, be wrongly regard as a user's opinion. The inference network is defined analogously to Eq. 3, which instead of taking a review document, takes a plot summary as input. An additional cross-entropy objective is minimized to drive the latent plot topics ($\boldsymbol{z}_a$) which would have a similar discriminative power as the features ($\boldsymbol{z}_d$) directly derived from the plots when used for plot classification:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{d}} &= \mathbb{E}_{q_\omega(\boldsymbol{z}_d|d)}[p_\zeta(d|\boldsymbol{z}_d)] - \text{KL}\left(q(\boldsymbol{z}_d|\boldsymbol{d}) \;||\; p(\boldsymbol{z}_d)\right) \\ \mathcal{L}_{plot_{z_a}} &= -\mathbb{E}_{q_\phi(\boldsymbol{z}_a)}\sum_{p=1}^{P} y_p \log(p(\hat{y}_p|\boldsymbol{z}_a)) \\ \mathcal{L}_{plot_{z_d}} &= -\mathbb{E}_{q_\omega(\boldsymbol{z}_d)}\sum_{p=1}^{P} y_p \log\left(p(\hat{y}_p|\boldsymbol{z}_d)\right)\end{aligned} \tag{9}$$

where P denotes the total number of plots in each dataset. Finally, $-\mathcal{L}_{\boldsymbol{d}}$ and $\mathcal{L}_{plot}$ are added to the overall loss defined in Eq. 8.

4 Experimental Setup

We conduct thorough experimental evaluations to assess the quality and disentanglement rate of extracted topics. To assess the quality of topics, we compute their topic coherence (Röder et al., 2015) coupled with their topic uniqueness. Then, we additionally look at the discriminative power of the disentangled features on the sentiment classification task. To fully assess the disentanglement rate of different methods, we perform topic labeling to compute the sentiment polarity of each topic (if any) and then measure the overall disentanglement rate (§5.3). As a result, we obtain an estimate of the extent to which different models can accurately control the topic disentanglement rate. We introduce and use a new dataset, named the *MOBO* dataset, pairing movie/book plots with their users' reviews, and including human-annotated sentences.

MOBO Dataset. The MOBO dataset is a collection of reviews and plots about **MO**vie and **BO**ok, associated to human-annotated sentences: while the pairs of reviews and plots are used to enhance the generation of plot topics, the human-annotated sentences provide the necessary ground-truth to automatically evaluate the topics' polarity.

Movie and book reviews were collected and paired from 3 public datasets: the Stanford's IMDB movie reviews (Maas et al., 2011), the GoodReads (Wan et al., 2019) and the Amazon reviews dataset (McAuley et al., 2015). Among all the available reviews in the IMDB dataset, we keep the ones with a corresponding plot in the MPST dataset (Kar et al., 2018), a corpus of movie synopses. The Goodreads dataset comes already with books' reviews paired with the related plots; while from the Amazon dataset, among all the product reviews, we keep only the ones related to movies available on the store and whose descriptions consist of the movie plots[3]. With the help of 15 annotators we further labeled more than 18,000 reviews' sentences ($\sim$ 6000 per corpus), marking the sentence polarity (`Positive`, `Negative`), or whether a sentence describes its corresponding movie/book `Plot`, or none of the above (`None`)[4]. We ensured that each sentence was labelled by at least 2 annotators by assigning overlapping subsets of $\sim$ 2400 sentences. In case of disagreement, when no consensus was reached, a final choice was made through a majority vote involving a third annotator. The final inter-annotator agreement (Cohen's kappa) was computed between each pair of annotators sharing a common subsets, with a minimum value of 0.572 and maximum of 0.831, for a resulting average of 0.758[5]. We report the dataset statistics in Table 1.
We report the dataset statistics in Table 1.

Baselines. We compare the experimental results with the following baselines:

[3]The dataset provides a predefined split of the corpus which preserves on train, development and test sets the same distribution of reviews based on their corresponding plots.

[4]We use *Doccano* as framework for collaborative labelling: https://github.com/doccano/

[5]We publicly release the full set of sentences with and without annotations for future expansion.

Statistics	IMDB	GoodReads	Amazon
# plots	1,131	150	100
# reviews	25,836	83,852	32,375
% Pos. reviews	0.46	0.33	0.32
% Neg. reviews	0.54	0.50	0.46
% Neu. reviews	0	0.17	0.22
Training set	20,317	65,816	25,883
Dev. set	2,965	9,007	3,275
Test set	2,554	9,029	3,217
# annotated sent.	6,000	6,000	6,000

Table 1: The MOBO dataset statistics.

sLDA (Mcauliffe et al., 2008): a supervised extension of LDA adding a response variable associated with each document.

JST (Lin and He, 2009): Joint Sentiment-Topic model built on LDA which is able to extract polarity-bearing topics.

NVDM (Miao et al., 2016): a VAE with an encoder network mapping the bag-of-words representations into continuous latent distributions, and a generative network for the document reconstruction.

GSM (Miao et al., 2017): based upon NVDM, the Gaussian Softmax topic model generating the topic distribution by applying a softmax function on the hidden representations of documents.

NTM (Ding et al., 2018): Neural Topic Model is a variation of NVDM by plugging the topic coherence metric directly into the model's objective.

PRODLDA (Srivastava et al., 2017): ProdLDA introduces a Dirichlet prior in place of Gaussian prior over the latent topic variable.

SCHOLAR (Card et al., 2018): a neural framework based on variational inference for the generation of topic while incorporating metadata information.

Parameter Setting. We perform tokenization and sentence splitting with SpaCy[6]. When available, we keep the default preprocessing, as it is the case for sLDA and SCHOLAR. Along with stopwords, we also remove tokens shorter than three characters and those with just digits or punctuation. We set the vocabulary to the 2,000 most common words as the best trade-off for each dataset. The 300-dimensional word vectors are initialized with a pretrained BERT embedding (Devlin et al., 2019). Sentence embeddings are generated from the Sentence-BERT using a pretrained BERT-large with mean-tokens pooling (Reimers and Gurevych, 2019). We use the predefined split of the MOBO dataset into training, development and test set in the proportion of 80/10/10 and average all the results over 5 runs.[7]

[6]https://spacy.io/

[7]Hyperparameter setting and training details are in Appendix.

Datasets	Models	Topic Coherence / Topic Uniqueness 25	50	100	200
IMDB	LDA	0.395 / 20.3	0.387 / 30.1	0.383 / 33.9	0.391 / 34.4
	sLDA	0.421 / 15.8	0.376 / 18.9	0.291 / 13.5	0.288 / 14.6
	JST	0.472 / 22.7	0.526 / 26.8	0.527 / 29.3	0.530 / 31.1
	NVDM	0.281 / 15.8	0.284 / 30.2	0.273 / **50.3**	0.266 / **54.8**
	GSM	0.384 / 22.4	0.402 / 21.0	0.410 / 39.7	0.394 / 42.4
	NTM	0.423 / 28.8	0.508 / 28.6	0.513 / 24.1	0.523 / 23.5
	PRODLDA	0.502 / 31.1	0.543 / 30.8	0.566 / 27.7	0.558 / 29.2
	SCHOLAR	**0.550** / 28.4	0.616 / 27.0	0.618 / 29.7	0.593 / 31.5
	DIATOM	0.544 / **37.1**	**0.639** / **38.1**	**0.626** / 36.5	**0.615** / 30.7
GoodReads	LDA	0.441 / 19.6	0.463 / 33.5	0.455 / 41.6	0.462 / 40.3
	sLDA	0.432 / 34.4	0.387 / 47.3	0.313 / 25.6	0.315 / 23.8
	JST	0.465 / 43.5	0.549 / 46.2	0.560 / 47.6	0.551 / 45.2
	NVDM	0.294 / 40.8	0.323 / 30.2	0.287 / 48.3	0.264 / 46.9
	GSM	0.411 / 24.8	0.481 / 40.1	0.482 / 38.1	0.473 / 41.4
	NTM	0.421 / 23.5	0.523 / 47.6	0.493 / 33.4	0.465 / 38.7
	PRODLDA	0.551 / 30.3	0.562 / 41.8	0.564 / 39.8	0.556 / 37.7
	SCHOLAR	0.545 / 38.3	0.603 / 42.0	**0.681** / 41.2	**0.664** / 38.4
	DIATOM	**0.582** / **54.0**	**0.634** / **52.9**	0.628 / **54.9**	0.609 / **48.7**
Amazon	LDA	0.430 / 28.9	0.447 / 47.5	0.438 / 64.8	0.445 / 59.3
	sLDA	0.421 / 67.7	0.393 / 62.1	0.323 / 87.5	0.331 / 74.8
	JST	0.450 / 73.0	0.558 / 71.2	0.544 / 78.8	0.518 / 70.9
	NVDM	0.278 / 42.4	0.310 / 32.5	0.281 / 38.4	0.261 / 49.1
	GSM	0.441 / 53.2	0.451 / 60.0	0.433 / 61.7	0.427 / 64.4
	NTM	0.493 / 52.8	0.501 / 53.1	0.547 / 55.3	0.508 / 59.3
	PRODLDA	0.492 / 63.4	0.543 / 51.4	0.528 / 58.7	0.551 / 62.1
	SCHOLAR	0.548 / 60.5	0.587 / 65.1	**0.641** / 63.2	0.629 / 68.2
	DIATOM	**0.563** / **82.0**	**0.598** / **82.3**	0.626 / **80.8**	**0.636** / **78.5**

Table 2: Topic Coherence and Topic Uniqueness results for 25/50/100/200 topics. The best result in each column and for each dataset is highlighted in **bold**.

5 Experimental Results

We report the results in terms of topic coherence/uniqueness, sentiment classification and topic disentanglement rate. We also perform ablation studies to gain more insights into our model.

5.1 Topic Coherence and Uniqueness

We conduct thorough experimental evaluations to assess the quality and disentanglement rate of extracted topics. To assess the quality of topics, we compute their topic coherence (Röder et al., 2015) coupled with their topic uniqueness. We evaluate the topic coherence using the C_V metric, a slightly refined Normalized Pointwise Mutual Information (NPMI) score using a boolean sliding window to determine the words' context (Röder et al., 2015).

Additionally, we monitor the topic uniqueness (TU) to measure word redundancy across topics. Following Nan et al. (2019), we use $cnt(l, k)$ to denote the total number of times the top word l in topic k appears among the top words across all topics, then $TU(k) = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{cnt(l,k)}$. TU is inversely proportional to the number of times each word appears in topics; a higher TU score implies that the top words are rarely repeated and, therefore, more diverse and unique topics.

In Table 2, we report the topic coherence and the topic uniqueness values. The supervised document label information was incorporated into sLDA, JST, SCHOLAR and DIATOM. Other models are purely unsupervised. We can observe that among conventional LDA-based models, JST performs significantly better compared to both LDA and sLDA for different topic settings and across all datasets. Neural topic models give mixed results. In terms of topic coherence, the trend is SCHOLAR $>$ PRODLDA $>$ NTM $>$ GSM $>$ NDVM. However, when we examine the topic uniqueness values, we can see that higher topic coherence values do not necessarily lead to higher topic uniqueness values. This shows that the topic coherence value could be misleading sometimes since a high topic coherence could be due to the redundancy of words across topics. We also notice that models with supervised document label information (except sLDA) generally outperform the unsupervised ones. This shows that the document label information can indeed help to extract more meaningful topics. When compared our proposed DIATOM with the baselines, we can observe that it achieves better coherence and topic uniqueness

Models	IMDB	GoodReads	Amazon
SVM			
+ TFIDF	0.672 ± 0.02	0.711 ± 0.01	0.661 ± 0.02
+ TFIDF + Lexicon	0.683 ± 0.02	**0.719** ± 0.02	0.667 ± 0.02
+ LDA	0.615 ± 0.02	0.659 ± 0.02	0.594 ± 0.01
sLDA	0.637 ± 0.01	0.652 ± 0.01	0.579 ± 0.01
JST	0.639 ± 0.01	0.518 ± 0.01	0.538 ± 0.01
SCHOLAR	0.645 ± 0.02	0.673 ± 0.03	0.613 ± 0.02
DIATOM	0.726 ± 0.03	0.704 ± 0.02	**0.686** ± 0.02
– w/o Plot Network	**0.734** ± 0.03	0.695 ± 0.03	0.603 ± 0.02

Table 3: Sentiment classification accuracy with 50 topics over the test set.

values most of the time, showing the benefit of separating opinion-bearing topics from plot topics by adversarial learning.

5.2 Sentiment Classification

In this section, we compare DIATOM with other supervised topic models for sentiment classification. The purpose of this evaluation is to highlight the discriminative power of the generated representations for the labels of interest while having attractive and unique properties as topic models, rather than confronting them with current state-of-the-art for text classification. We additionally report some baseline results using a Support Vector Machine (SVM) which has been widely employed on these task (Pang and Lee, 2004b) providing an understanding of the relative differences in performance of different approaches.

Table 3 shows the sentiment classification accuracy. In JST, the supervised document label information is only incorporated as prior to the model, while both sLDA and SCHOLAR treat the class label of each document as a response variable and jointly model both documents and their responses. We can observe that the latter is more effective in incorporating supervised information since both sLDA and SCHOLAR outperform JST in general. But DIATOM gives significantly better results all over the baselines with the improvement over the best baseline model, SCHOLAR, by 3-8%. In our models, features used for sentiment classification are opinion-bearing topics. This shows that separating opinion topics from plot/neutral topics is beneficial for sentiment classification. We also observe that the contribution of the plot network to sentiment classification is dataset-dependent. The use of plot network largely boosts the sentiment classification accuracy by over 8% on the Amazon dataset. But its effect is negligible on the other two datasets.

When compared with traditional sentiment classification models such as SVM, we found that DIATOM outperforms SVM trained with various features on both IMDB and Amazon. But it performs slightly worse than SVM trained with TFIDF features with or without an additional incorporation of sentiment lexicon features. Nevertheless, DIATOM gives superior performance compared to SVM trained on LDA topic features in the range of 5-11%, showing the effectiveness of using opinion topics for sentiment classification.

5.3 Topic Disentanglement

None of the aforementioned measures can, however, capture how opinion and plot topics are distributed. To this aim, we use topic labeling to assign a proxy label (`Positive`, `Negative`, `Plot`, `None`) to each topic and then measure the topic-disentanglement rate ρ as the proportion of opinion-bearing topics with respect to the overall set of topics, complementary to the proportion of plot/neutral topics: $\rho = \frac{S}{S+K}$, with S being the number of opinion topics and K the number of plot/neutral topics.

For each topic, we first calculate its embedding by taking the normalized weighted average of the vectors of its top N words: $\overrightarrow{t_z} = \frac{1}{N}\sum_{i=1}^{N} \alpha_i \times \overrightarrow{w_i}$, where α_i is the normalized distribution of word w_i in topic z. We then retrieve the top 10 most similar sentences from the human-annotated sentence set measured by the cosine similarity between the topic embedding and each sentence embedding. The sentence embedding is computed using the Sentence-BERT encoder (Reimers and Gurevych, 2019). The most frequent label among the retrieved sentences is adopted as the topic's label.

To highlight the disentanglement capability of DIATOM, in Figure 3, we analyse how the proportion of opinion-bearing topics varies across standard and sentiment topic models. We notice that

Plot/Neutral Topics
Amazon - Topic 1 *Dent, Gotham, City, Gordon, Bruce, Wayne, Harvey, Joker, Criminal, Nolan*
1. Being imprisoned Batman has enough time to paint a gigantic flaming bat on a bridge while people are literally being executed on the hour. 2. Batman gets with Catwoman... after how hard she sold him out?
Amazon - Topic 2 *Gandalf, Frodo, Jackson, Tolkien, Dwarf, Fellowship, Peter, Orc, Ring, Hobbit*
1. [...] the myriad inhabitants of Middle-earth, the legendary Rings of Power, and the fellowship of hobbits, elves, dwarfs, and humans–led by the wizard Gandalf (Ian McKellen) and the brave hobbit Frodo. 2. This is the beginning of a trilogy; soon to be finalized.
Opinion-Bearing Topics
Amazon - Topic 1 *Expectation, Quality, Definitely, Great, Good, Worth, Graphic, Predictable, Compare, Decent*
1. Action is good. 2. Rachel Weisz was “mostly" good.
Amazon - Topic 2 *Price, Shame, Service , Normally, Purchase, Connection, Greed, Stream, Watch, Frustrate*
1. This experience leaves me skeptical of the Amazon Prime video service. 2. Look closely before purchasing.

Table 4: Example topics extracted by DIATOM from Amazon reviews and their associated most similar sentences.

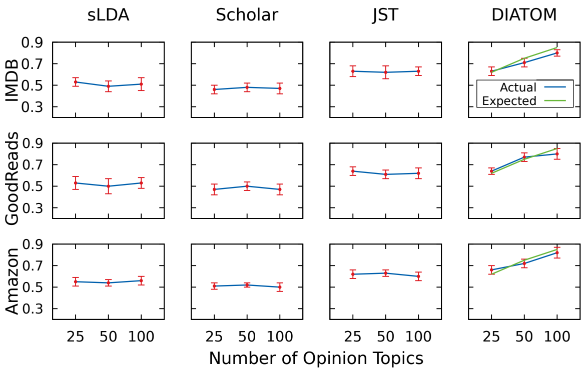

Figure 3: Disentangling rate (%) of topic models across different number of topics.

despite the signal from the document labels, sLDA and SCHOLAR tend to produce topics rather balanced in terms of neutral and opinion-bearing topics. JST has a more skewed distribution towards opinion topics. DIATOM instead generates an actual proportion of opinion topics approaching the expected proportion set up by the model, demonstrating the capability to control the generation of plot and opinion-bearing topics.

In Table 4, we show a set of topics grouped according to the disentanglement induced by DIATOM from Amazon[8]. For each topic, we report an excerpt of the most similar sentences retrieved. Aside from being overall coherent, we can guess more peculiar plots related for instance to ‘*The Hobbit*’ or ‘*Batman*’ as in the Plot/Neutral topics. The opinion-bearing topics report a collection of commonly appreciated or critic aspects; some of them are made up of mixed terms describing the issues and the associated experience (e.g. Topic 2).

[8]Topic results from IMDB and GoodRead can be found in the Appendix.

5.4 Visualization

Another way to look at the disentangled topics is through the visualization of topic vectors.

As an example, we plot in Figure 4 the 2-dimensional representation of the topic distributions projected by t-SNE for the Amazon dataset. Different colors represent different types of topics generated by DIATOM, namely plot/neutral in blue and opinion in red. We notice how consistently plot/neutral topics tend to cluster together across different number of topics, with the boundary close to polarized topics likely to share common features, as shown in Figure 1 in which the plot topic and the negative topic share a common word ‘*Dwarf*’.

5.5 Ablation Study

We report in Table 5 the results of the ablation study on DIATOM. We observe that removing the orthogonal regularization has a limited effect on sentiment classification, but causes a fluctuation on topic coherence and a clear drop in topic uniqueness. A significant classification performance drop is observed by removing the sentiment classifier, which essentially reduces DIATOM to an unsupervised model. Removing both the orthogonal regularization and the sentiment classifier shows a major negative impact on both accuracy and the topic’s quality. Finally, we assess the influence of the plot network (§3), and while we do not notice any consistent impact across the datasets in terms of sentiment classification, the topics has a notable drop in terms of coherence and diversity.

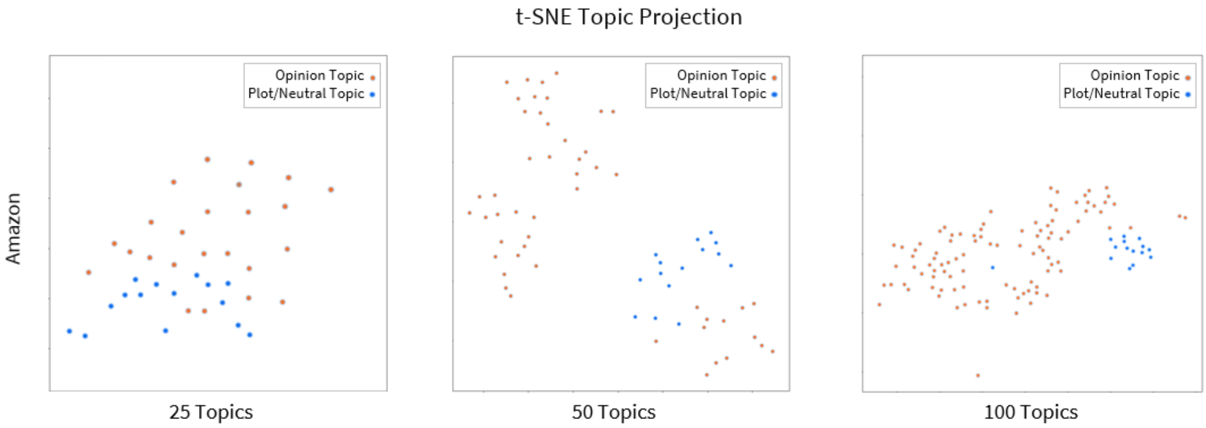

Figure 4: Example of t-SNE projection for the Amazon dataset of the topic distribution for different number of topics. Color are assigned according to plot/neutral and opinion topics.

Datasets	**Models**	**Accuracy**	**TC / TU**
IMDB	DIATOM	0.726 ±0.03	0.639 / 38.1
	– w/o orth reg.	0.723 ±0.01	0.582 / 27.5
	– w/o sent. class.	0.491 ±0.03	0.601 / 35.4
	– w/o both	0.478 ±0.03	0.544 / 25.4
	– w/o Plot Net	0.734 ±0.03	0.603 / 36.7
GoodReads	DIATOM	0.704 ±0.02	0.634 / 52.9
	– w/o orth. reg.	0.681 ±0.02	0.612 / 41.1
	– w/o sent. class.	0.446 ±0.02	0.638 / 47.6
	– w/o both	0.410 ±0.02	0.552 / 39.6
	– w/o Plot Net	0.695 ±0.03	0.615 / 49.3
Amazon	DIATOM	0.686 ±0.02	0.598 / 82.3
	– w/o orth reg.	0.682 ±0.01	0.605 / 55.3
	– w/o sent. class.	0.601 ±0.03	0.573 / 76.9
	– w/o both	0.548 ±0.03	0.567 / 52.1
	– w/o Plot Net	0.603 ±0.02	0.584 / 78.3

Table 5: Ablation study over DIATOM by removing the orthogonal regularization, the sentiment classifier or just the auxiliary Plot Network.

5.6 Further Discussion

Although the adversarial mechanism implemented in DIATOM is rather effective in disentangling opinion and neutral/plot topics, at times the opinion topics could exhibit terms of mixed polarities. An additional adversarial mechanism can be a viable solution at the cost of increasing the model's complexity.

In our current model, the latent plot topics z_a extracted from reviews are encouraged to have a similar discriminative power as the latent topic z_d learned from plots directly for predicting the plots. It is also possible to impose a Gaussian prior centred on z_d for the latent plot topics in reviews instead of using the Gaussian prior of zero mean and unit variance.

While we focus on separating opinion topics from plot or neutral ones in movie and book reviews in this work, our proposed framework can be applicable in other scenarios. For example, in veracity classification of Twitter rumours, we want to disentangle latent factors which are indicative of veracity of tweets from those which are event-related. Our proposed framework provides a potential solution to it.

6 Conclusions

We have described DIATOM, a new neural topic model to generate disentangled topics through the combination of VAE and adversarial learning. The results on the novel MOBO dataset show that DIATOM generates better topics in terms of both topic coherence and topic uniqueness, and can disentangle opinion-bearing topics from plot/neutral ones. Finally, we have identified some existing limitations and provided viable solutions to be explored in the future.

Acknowledgements

Our gratitude goes to all the annotators and to Fabio Miceli and Gianluca Pergola for their relentless effort. This work is funded by the EPSRC (grant no. EP/T017112/1, EP/V048597/1). YH is supported by a Turing AI Fellowship funded by the UK Research and Innovation (UKRI) (grant no. EP/V020579/1).

References

Alessandro Achille and Stefano Soatto. 2018. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research (JMLR)*, 19(1):1947–1980.

Emmanuel Bengio, Valentin Thomas, Joelle Pineau, Doina Precup, and Yoshua Bengio. 2017. Independently controllable features. *CoRR*, abs/1703.07718.

Y. Bengio, A. Courville, and P. Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

David M. Blei, John D. Lafferty, and Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, (ICML), pages 113–120.

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

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*.

Dallas Card, Chenhao Tan, and Noah A. Smith. 2018. Neural Models for Documents with Metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, (ACL)*, pages 2031–2040, Melbourne, Australia.

Tim Cooijmans, Nicolas Ballas, César Laurent, Çağlar Gülçehre, and Aaron Courville. 2017. Recurrent batch normalization. In *Proceedings of the 2017 International Conference for Learning Representations, (ICLR)*, Touloun, France.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL)*. Association for Computational Linguistics.

Adji B. Dieng, Chong Wang, Jianfeng Gao, and John W. Paisley. 2017. Topicrnn: A recurrent neural network with long-range semantic dependency. In *5th International Conference on Learning Representations, (ICLR), Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Ran Ding, Ramesh Nallapati, and Bing Xiang. 2018. Coherence-aware neural topic modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, (EMNLP)*, pages 830–836, Brussels, Belgium.

Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, N Siddharth, Brooks Paige, Dana H. Brooks, Jennifer Dy, and Jan-Willem van de Meent. 2019. Structured disentangled representations. In *In Proceedings of Machine Learning Research (PMLR)*, pages 2525–2534.

Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. 2019. Auto-encoding total correlation explanation. volume 89 of *Proceedings of Machine Learning Research*, pages 1157–1166. (PMLR).

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics, (AISTATS)*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27, (NIPS)*, pages 2672–2680.

Lin Gui, Jia Leng, Gabriele Pergola, yu zhou, Ruifeng Xu, and Yulan He. 2019. Neural topic model with reinforcement learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. (ACL).

I. Higgins, Loïc Matthey, A. Pal, C. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner. 2017. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, Workshop Track Proceedings, (ICLR)*.

Irina Higgins, David Amos, David Pfau, Sébastien Racanière, Loïc Matthey, Danilo J. Rezende, and Alexander Lerchner. 2018. Towards a definition of disentangled representations. *CoRR*, abs/1812.02230.

Tai Hoang, Huy Le, and Tho Quan. 2019. Towards autoencoding variational inference for aspect-based opinion summary. *Applied Artificial Intelligence (API)*, 33:796–816.

Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining, (WSDM)*, pages 815–824.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, (ACL)*, Florence, Italy.

Sudipta Kar, Suraj Maharjan, A. Pastor López-Monroy, and Thamar Solorio. 2018. MPST: A corpus of movie plot synopses with tags. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, (LREC)*, Miyazaki, Japan.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 2015 International Conference for Learning Representations, (ICLR)*.

Diederik P. Kingma, Max Welling, and Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, (ICLR)*.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. 2017. Variational inference of disentangled latent concepts from unlabeled observations. In *5th International Conference on Learning Representations, (ICLR)*.

C. Lin, Y. He, R. Everson, and S. Ruger. 2012. Weakly supervised joint sentiment-topic detection from text, 2012. *IEEE Transactions on Knowledge and Data Engineering*, (6):1134–1145.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, (CIKM), pages 375–384.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, (ACL)*, pages 142–150.

James Martens. 2010. Deep learning via hessian-free optimization. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, (ICML), pages 735–742.

Tomonari Masada, Atsuhiro Takasu, and Takasu. 2018. Adversarial learning for topic models. In *Advanced Data Mining and Applications, (ADMA)*.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR)*.

Jon D. Mcauliffe, David M. Blei, and Blei. 2008. Supervised topic models. In *Advances in Neural Information Processing Systems 20, (NIPS)*, pages 121–128.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. *In Proceedings of The 34th International Conference on Machine Learning, (ICML)*.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of the 33nd International Conference on Machine Learning, (ICML)*, pages 1727–1736.

Andriy Mnih, Karol Gregor, and Gregor. 2014. Neural variational inference and learning in belief networks. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, (ICML).

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with Wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, (ACL)*, pages 6345–6381, Florence, Italy.

Bo Pang and Lillian Lee. 2004a. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 271–278.

Bo Pang and Lillian Lee. 2004b. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain.

Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. 2019. Domain agnostic learning with disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning, (ICML)*, pages 5102–5112.

Gabriele Pergola, Lin Gui, and Yulan He. 2019. TDAM: A Topic-Dependent Attention Model for Sentiment Analysis. *Information Processing & Management*, 56(6):102084.

Gabriele Pergola, Yulan He, and David Lowe. 2018. Topical phrase extraction from clinical reports by incorporating both local and global context. In *The 2nd AAAI Workshop on Health Intelligence, (AAAI)*, pages 499–506.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, (WSDM)*, pages 399–408.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, (UAI), pages 487–494.

Akash Srivastava, Charles A. Sutton, and A. Sutton. 2017. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, (ICLR)*.

Valentin Thomas, Emmanuel Bengio, William Fedus, Jules Pondard, Philippe Beaudoin, Hugo Larochelle, Joelle Pineau, Doina Precup, and Yoshua Bengio. 2017. Disentangling the independently controllable factors of variation by interacting with the world. pages 2590–2599.

Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, (ACL), Florence, Italy*, pages 2605–2610.

Rui Wang, Deyu Zhou, and Yulan He. 2019. Atm: Adversarial-neural topic model. *Information Processing & Management*, 56(6):102098.

Xuerui Wang, Andrew McCallum, and McCallum. 2006. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (SIGKDD)*, pages 424–433.

Steven Wilson, Rada Mihalcea, Ryan Boyd, and James Pennebaker. 2016. Disentangling topic models: A cross-cultural analysis of personal values through words. In *Proceedings of the First Workshop on NLP and Computational Social Science*.

Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. 2017. Variational autoencoder for semi-supervised text classification. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, (AAAI)*.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. 2017. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of the 34th International Conference on Machine Learning, (ICML)*, pages 3881–3890.

A Appendix

A.1 Training Details

Hyperparameters We tune the models' hyperparameters on the development set via a random search over combinations of learning rate $\lambda \in [0.001, 0.5]$, dropout $\delta \in [0.0, 0.6]$ and topic vector size $\gamma_t \in [25, 50, 100, 200]$. Encoder and decoder are configured following (Srivastava et al., 2017). The hidden representation of documents is set to 100 and sentiment classifier's hidden size to 50. Matrices are randomly initialized with the Xavier and sparse methods (Glorot and Bengio, 2010; Martens, 2010). We employ the Adam optimizer (Kingma and Ba, 2015), set the batch size to 64 and apply batch normalization as additional regularizer (Cooijmans et al., 2017).

Sequential Unfreezing Instead of simultaneously training all the model components, we unfreeze them sequentially. We first freeze the sentiment classifier and update only the autoencoder. At the e_{th} epoch, we unfreeze the sentiment classifier uniquely on the polarized features to let the classifier training. Finally, at the $(e+n)_{th}$ epoch, we unfreeze the adversarial mechanism to drive the generation of neutral features fooling the classifier. We follow an analogous approach with regards to the plot classifier. The values of e and a are treated as hyperparameters and chosen through the random search. We found that the sequential unfreezing scheme leads to better topic disentanglement.

A.2 Example Topic Results

In Table A1, we show a set of topics grouped according to the disentanglement induced by DIATOM from IMDB and GoodReads. For each topic, we report an excerpt of the most similar sentences retrieved. Aside from being overall coherent, we can guess rather paradigmatic themes as the Topic 1 about peace and war between countries in IMDB-Topic 1. It is worth having a closer look at the IMDB-Topic 2, which despite the *negative* theme of depression and suicide, the model is able to correctly gather those words under the same plot topic. The opinion-bearing topics report a collection of commonly appreciated or critic aspects; some of them are mainly collections of related adjectives with the same polarity (e.g. IMDB-Topic 1).

Plot/Neutral Topics
IMDB - Topic 1 *Government, Country, Peace, Information, Free, Plane, Theory, Anti, Soldier, Hitler*
1. Groundbreaking in the realm of socially relevant drama, it dealt with issues such as abortion, domestic violence, student protest, child neglect, illiteracy, slumlords, the anti-war movement, [...]. 2. This effort by Charlie ultimately evolves into a major portion of the U.S. foreign policy known as the Reagan Doctrine, under which the U.S. expanded assistance beyond just the [...].
IMDB - Topic 2 *Window, Hospital, Apartment, Suicide, Commit, Pitt, Serial, Strange, Killer, Mental*
1. Even re-think why two boys/young men would do what they did - commit mutual suicide via slaughtering their classmates. 2. It’s the patented scene where the killer creeps up behind the victim.
GoodReads - Topic 1 *Cure, Plague, Trial, Betray, Thomas, Secret, Dashner, Ball, Betrayal, Wicked*
1. Blaming Cinder for her daughter’s illness, Cinder’s stepmother volunteers her body for plague research, an "honor" that no one has survived. 2. By age thirteen, she has undergone countless surgeries, transfusions, and shots so that her older sister, Kate, can somehow fight the leukemia that has plagued her since childhood.
GoodReads - Topic 2 *Teenager, Fault, Illness, Mental, Depression, Maddy, Grief, Bully, Topic, Greg*
1. She’s got a lot of mental strength, having been ostracized for most of her life. 2. She went through a divorce, a crushing depression, another failed love, and the eradication of everything she ever thought she was supposed to be.

Opinion-Bearing Topics
IMDB - Topic 1 *Badly, Stock, Remove, Poorly, Hype, Ridiculous, Insult, Disaster, Excuse, Lame*
1. I can’t imagine how anyone could have read this badly written script and given it the greenlight. 2. Although there has obviously been a lot of money spent on them the numbers are badly staged and poorly photographed.
IMDB - Topic 2 *Exceptional, Recommend, Excellent, Craft, Believable , Overlook, Vhs, Solid, Festival, Amaze*
1. Overall, this is a good film and an excellent adaption. 2. It’s great acting, superb cinematography and excellent writing.
GoodReads - Topic 1 *Negative, Judge, Note, Pretend, Embarrass, Quality, Extreme, Guilty, Fake, Borrow*
1. Can you give something negative stars? 2. And while it must be hard reading negative reviews you need to be able to deal with this in a graceful way (no one likes a sore loser).
GoodReads - Topic 2 *Teen, Nice, Normally, Little, Genre, Amuse, Theme, Enjoyment, Blow, Reread*
1. What would have made the book a lot more fun to read was more meatier characters in the other girls. 2. But I feel like that was part of the fun of it.

Table A1: Example topics extracted by DIATOM from IMDB and GoodReads and their associated most similar sentences.