

Neural Multimodal Topic Modeling: A Comprehensive Evaluation

Felipe González-Pizarro, Giuseppe Carenini

Department of Computer Science, University of British Columbia
Vancouver, BC, Canada
{felipegp, carenini}@cs.ubc.ca

Abstract

Neural topic models can successfully find coherent and diverse topics in textual data. However, they are limited in dealing with multimodal datasets (e.g., images and text). This paper presents the first systematic and comprehensive evaluation of multimodal topic modeling of documents containing both text and images. In the process, we propose two novel topic modeling solutions and two novel evaluation metrics. Overall, our evaluation on an unprecedented rich and diverse collection of datasets indicates that both of our models generate coherent and diverse topics. Nevertheless, the extent to which one method outperforms the other depends on the metrics and dataset combinations, which suggests further exploration of hybrid solutions in the future. Notably, our succinct human evaluation aligns with the outcomes determined by our proposed metrics. This alignment not only reinforces the credibility of our metrics but also highlights the potential for their application in guiding future multimodal topic modeling endeavors.

Keywords: multimodal topic modeling, neural topic model, topic model evaluation

1. Introduction

The vast amount of text that is constantly generated has led to the development of several algorithms designed to interpret and summarize large sets of documents (Peter et al., 2015). A well-known automatic mechanism is topic modeling, a robust approach for extracting core themes from large text corpora. In practice, when topic modeling is applied to a corpus (e.g., news articles), the results will include a set of topics. Usually, each topic is represented by a list of related terms (e.g., *tropical, storm, hurricane, cyclone, weather, rain*) (Zhao et al., 2021). Domain experts (e.g., journalists, physicians, and marketers) can use topic modeling to analyze large document collections without reading every document (Boyd-Graber et al., 2017; Ge et al., 2019).

Since most topic modeling algorithms have been designed specifically to process textual data, their performance is limited in corpora containing information in other modalities (e.g., images and videos). Some work has attempted to address this limitation by expanding well-known probabilistic (Bian et al., 2013; Zhang et al., 2022) or neural topic models (Zosa and Pivovarova, 2022a) to multimodal settings, especially considering images in the documents. Yet, it remains unclear which method works better for which dataset, and what evaluation metrics are more appropriate in this new multimodal scenario. This paper addresses this gap by conducting the first systematic and comprehensive evaluation of multimodal topic modeling applied to documents containing both text and images. In this process, we make several contributions with respect to neural topic modeling algorithms, evaluation metrics, and datasets.

Neural topic modeling algorithms: We have developed two novel neural multimodal topic modeling algorithms by adapting SOTA solutions¹. First, we extend ZeroShotTM (Bianchi et al., 2021b), a neural topic model that only uses pre-trained textual embeddings, to the multimodal Multimodal-ZeroShotTM, which additionally embeds images. In particular, both the Bag-Of-Words (BOW) and the Image Features associated with each document are reconstructed during decoding. Secondly, we present Multimodal-Contrast, derived from M3L-Contrast (Zosa and Pivovarova, 2022a), a recent multimodal multilingual neural topic model that uses Contrastive Learning to map texts from multiple languages and images into a shared topic space. Our Multimodal-Contrast simply omits the encoder and inference networks associated with a second language.

Metrics: The quality of a given set of topics can be automatically assessed primarily based on their coherence and diversity. For topic modeling methods that only process textual data, coherence metrics such as NPML (Lau et al., 2014), C_v (Röder et al., 2015), and WE (Fang et al., 2016) evaluate the semantic relatedness of the topic keywords; while diversity metrics like TD (Dieng et al., 2020) and I-RBO (Bianchi et al., 2021a) measure the lexical overlap between the descriptors of different topics. However, in a multimodal setting, each topic is represented not only by a set of keywords, but also by a set of images. Yet, there are no automatic metrics to assess the coherence and segregation of the images representing a topic. In this paper, to fill this gap, we propose two new metrics, namely Image Embedding-based Coherence (IEC) and Im-

¹Our code is available at : https://github.com/gonzalezf/multimodal_neural_topic_modeling/

age Embedding-based Pairwise Similarity (IEPS), which appear to align with human judgment in our preliminary user study.

Datasets: While previous work on multimodal topic modeling has been tested only on a few rather homogeneous datasets, we are the first to propose and leverage six diverse datasets that vary substantially in terms of the document size (ranging from 6 to 2,425 words per document on average), the source of the documents (e.g., Flickr, Twitter, Wikipedia), the underlying task/domain (e.g., Object recognition, Visual Storytelling) as well as in the way the gold-standard data was collected (e.g., crowd-sourcing, automatic classification).

Armed with a comprehensive set of metrics and a diverse collection of datasets, we perform the first systematic evaluation of multimodal neural topic modeling methods, comparing our novel proposals, `Multimodal-ZeroShotTM` and `Multimodal-Contrast`, among themselves and against only textual SOTA topic modeling methods.

2. Related Work

With the recent developments of deep neural networks, several *Neural Topic Models* (NTMs) have been proposed. For instance, [Srivastava and Sutton \(2017\)](#) proposed *Product-of-Experts LDA* (ProdLDA), a topic modeling algorithm that, like the original (LDA) ([Blei et al., 2003a](#)), still uses a BOW representation of documents, but leverages it in a more sophisticated way by combining a variational autoencoder (VAE) ([Blei et al., 2017](#)) with a Product of Experts (PoE) approach ([Hinton, 2002](#)). As a result, ProdLDA not only consistently identifies more coherent and diverse topics than LDA ([Srivastava and Sutton, 2017](#); [Sridhar et al., 2022](#)), but it also process data more efficiently ([Srivastava and Sutton, 2017](#)). Despite progress, both LDA and ProdLDA are still limited to BOW document representations. By ignoring critical syntactic and semantic relationships among words, they sometimes fail to identify high-quality topics (see ([Bianchi et al., 2021a](#)) and ([Burkhardt and Kramer, 2019](#))). In order to incorporate semantic relationships into topic models, [Dieng et al. \(2020\)](#) proposed *Embedding Topic Models* (ETM), a generative probabilistic model that relies on static word embeddings ([Mikolov et al., 2013](#)) to identify interpretable topics.

Nevertheless, a remaining key shortcoming of ETM is that by relying on static embeddings, it does not consider contextual relations among words. This limitation was recently addressed by [Bianchi et al. \(2021a\)](#) and [Bianchi et al. \(2021b\)](#), which proposed *Contextualized Topic Models* (CTM). CTM is a family of neural topic models based on a variational autoen-

coder (VAE) (i.e., `CombinedTM` ([Bianchi et al., 2021a](#)), `ZeroShotTM` ([Bianchi et al., 2021b](#))), that relies instead on contextual embeddings (e.g., SBERT ([Reimers and Gurevych, 2019a](#))), obtaining higher quality topics than all previous approaches. One of the two neural topic models we propose in this paper, `Multimodal-ZeroShotTM`, is based on this recent work.

In contrast to the aforementioned neural topic models, `BERTopic` ([Grootendorst, 2022](#)), takes a distinct approach by employing a clustering methodology. Unlike the neural topic models that infer a mixture of topics within documents, `BERTopic` assumes each document correlates with a single topic. This model incorporates heuristic techniques to manage this limitation, yet the effectiveness of these strategies remains an area for further exploration. Our investigation concentrates on neural topic models, which inherently consider the presence of multiple topics in documents, providing a broader understanding of the data’s thematic structure.

Only a few topic modeling algorithms have been proposed to process more than textual data. A notable exception is `M3L-Contrast` ([Zosa and Pivovarova, 2022a](#)), a neural topic model that maps texts from multiple languages and images into a shared topic space by using pre-trained image (CLIP ([Radford et al., 2021b](#))), and text embeddings (SBERT ([Reimers and Gurevych, 2019b](#))) to abstract the complexities between different languages and modalities. The second neural topic model we propose here, `Multimodal-Contrast`, is based on this recent work.

Given the recent success of decoder-only GPT-like systems ([Achiam et al., 2023](#)) in so many NLP tasks, it may seem surprising that they have not been applied yet to the topic modeling task. However, the plain reason is that they are still severely limited in their input size, currently in the 10,000s of tokens ([Bubeck et al., 2023](#)), and therefore cannot process the large corpora for which topic modeling is actually needed. Very recent work (e.g., ([Yu et al., 2023](#))) may inspire ideas on how to overcome this limitation in the future. Tellingly, the evaluation framework and baselines presented in this paper will be critical in assessing these new solutions.

3. Our New Multimodal Algorithms

3.1. Multimodal-ZeroShotTM

We propose `Multimodal-ZeroShotTM`, a novel multimodal topic modeling algorithm based on `ZeroShotTM` ([Bianchi et al., 2021b](#)). Figure 1 shows the architecture of our model. Given a document with a textual and visual component (e.g., an image and its caption), we encode each element using

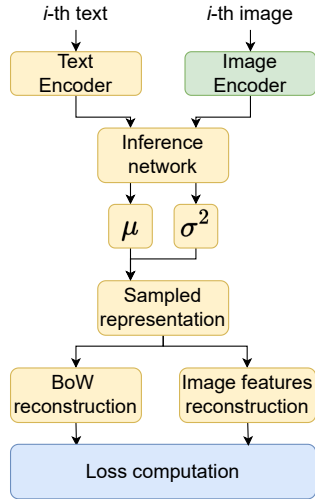


Figure 1: High-level schema of the architecture for Multimodal-ZeroShotTM. The loss function is detailed in Equation 1.

a modality-specific encoder (e.g., by using CLIP image and text encoders). Then, we concatenate those embeddings and pass them to an inference network as input. After that, the model samples a latent representation from a Gaussian distribution parameterized by μ and σ^2 , as is done in related works (Bianchi et al., 2021b; Zosa and Pivovarova, 2022a; Srivastava and Sutton, 2017). The crucial difference between our model and ZeroShotTM is that our decoder network reconstructs both the BoW and the Image Features associated with each document (bottom right of Figure 1). Advantageously, by embedding images and reconstructing their features, the model can capture complementary information not present in the textual parts of the document, thus plausibly identifying topics that are more representative of the multimodal corpus. Our per document loss function \mathcal{L} now includes three components:

$$\mathcal{L} = \mathbb{E}_q [\mathbf{w}^\top \log(\text{softmax}(\beta\theta))] - \mathbb{KL}(Q(\theta | \mathbf{x}) \| P(\theta)) + \lambda(1 - \text{cos}(\mathbf{x}_{img}, \gamma\theta)) \quad (1)$$

where the first term measures the loss associated with the BoW vector reconstruction of the document (see (Srivastava and Sutton, 2017) for more details). The second term corresponds to the sum of the Kullback-Leibler divergence (KL) loss between the posterior and prior distributions for the document embedding \mathbf{x} (i.e., the concatenation of the text and image embeddings). The mean μ and variance σ^2 of the posterior distributions are estimated in each inference network, and θ is the sampled topic distribution per document embedding. Finally, the third term corresponds to the loss associated with the reconstruction of Image Fea-

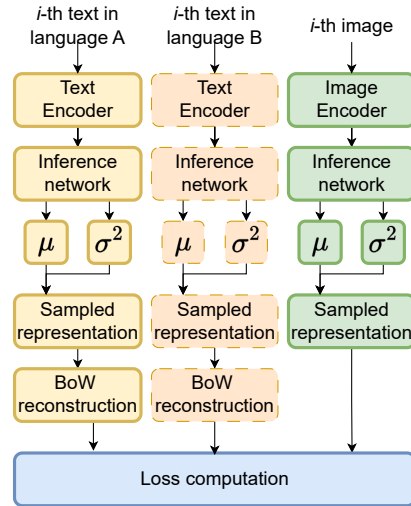


Figure 2: M3L-Contrast topic model architecture. It includes language-specific and modality-specific encoders and inference networks. We highlight with a dashed line (- -) the components that we removed during our adaptation.

tures, as the cosine embedding loss² between the estimated image features $\gamma\theta$ and the actual value \mathbf{x}_{img} . This loss measures the similarity between the two vectors and is often used for learning non-linear embeddings. λ is a parameter to explore the trade-off between textual and image losses but has been kept equal to 1 in the main experiments.

3.2. Multimodal-Contrast

We propose Multimodal-Contrast, an adaptation specifically derived from M3L-Contrast (Zosa and Pivovarova, 2022a). While M3L-Contrast is a neural topic modeling technique tailored for analyzing datasets that are both multilingual and multimodal, Multimodal-Contrast shifts the focus to solely multimodal data. In M3L-Contrast, each document must contain an image and textual content in two languages (e.g., English and German), with the model’s architecture including three encoders and inference networks (see Figure 2), each one processing either text in one of the two languages or an image. In our adaptation, we essentially removed from the architecture the encoder and inference network for one of the two languages (dashed in Figure 2, i.e., the components for language B).

The main difference between Multimodal-Contrast and Multimodal-ZeroShotTM is the third component of the loss function \mathcal{L} . In particular, while Multimodal-ZeroShotTM considers the loss associated with the reconstruction of Image Features, Multimodal-Contrast uses the InfoNCE Contrastive Learning loss (Oord et al., 2018) to align

²Using MSE as this loss delivers similar performance.

topic distributions sampled from the different modalities of the document (i.e., its textual and visual parts). This Contrastive Learning loss maps similar instances close to each other and keeps non-related instances apart. Overall, \mathcal{L} combines three components:

$$\begin{aligned} \mathcal{L} = & \mathbb{E}_q [\mathbf{w}^\top \log(\text{softmax}(\beta\theta_{txt}))] - \\ & \sum_{l=1}^k \text{KL}(Q(\theta^l | \mathbf{x}^l) \| P(\theta^l)) - \\ & \omega \sum_{\substack{a,b=1 \\ a \neq b}}^k \log \frac{\exp((\theta^a \cdot \theta^b) / \tau)}{\sum_{j=1}^N \sum_{c,d=1}^k \exp((\theta^c \cdot \theta_j^d) / \tau)} \end{aligned} \quad (2)$$

The first term corresponds to the standard BoW reconstruction loss given the textual component of the document. The second term corresponds to the KL loss between posterior and prior distributions for every component k (i.e., image, text) of a document. So, it is similar but not the same as in Multimodal-ZeroShotTM. Here, the mean μ and variance σ^2 of the posterior distributions are estimated in each inference network; \mathbf{x} can be a textual or visual embedding, and θ is a sampled topic distribution given a textual or visual component of a document. Finally, the third term is the InfoNCE loss, where $(\theta^a \cdot \theta^b)$ are positive pairs and $(\theta^c \cdot \theta^d)$ are negative pairs. Positive terms are aligned components of a document (e.g., an image and its caption). N is the batch size, τ is the temperature, and ω is a parameter (like λ for Multimodal-ZeroShotTM) to explore the trade-off between contrastive and other losses, but has been kept equal to 100 in the main experiments as (Zosa and Pivovarova, 2022a).

4. Automatic Metrics

The quality of topic models is commonly assessed based on their coherence and diversity. Automatic coherence metrics identify the degree of lexical and semantic relatedness between the terms that describe each topic, while automatic diversity metrics measure the lexical overlap between the terms of different topics. In our evaluation, we apply standard metrics to assess the quality of the textual descriptors of the topics, namely NPMI (Lau et al., 2014), C_v (Röder et al., 2015), and WE (Fang et al., 2016) for coherence and TD (Dieng et al., 2020) and I-RBO (Bianchi et al., 2021a) for diversity³.

However, in multimodal topic models, each topic is represented not only by a set of keywords but also by a set of images. Since there are currently no automatic metrics for evaluating the co-

herence and diversity of images representing topics, we propose two new metrics, namely Image Embedding-based Coherence (IEC) and Image Embedding-based Pairwise Similarity (IEPS), to fill this gap.

Image Embedding-based Coherence (IEC) is based on WE (Fang et al., 2016), a metric that has been validated and widely used (Bianchi et al., 2021a; Li et al., 2023). While the WE metric was designed to gauge the semantic relatedness between word embeddings, our IEC evaluates the semantic relatedness between images representing a topic. Formally, let T be the set of topics, and W_t be the set of top- N images in topic $t \in T$. The average pairwise image similarity for topic t is:

$$\text{sim}(W_t) = \frac{1}{\binom{|W_t|}{2}} \sum_{i=1}^{|W_t|} \sum_{j=i+1}^{|W_t|} \text{cosine}(w_i, w_j) \quad (3)$$

where $\text{cosine}(w_i, w_j)$ computes relatedness between any two images w_i and w_j as the cosine similarity between their corresponding embeddings. Finally, our new metric, IEC, is simply the average of the topic-level similarity scores across all topics:

$$\text{IEC} = \frac{1}{|T|} \sum_{t \in T} \text{sim}(W_t) \quad (4)$$

IEC ranges between $[0,1]$, where a higher value suggests more coherent topics.

Image Embedding-based Pairwise Similarity (IEPS) measures the diversity of a topic model by computing the similarity between all the topics and then considering lower scores as a sign of higher diversity. Formally, by adapting the Word Embedding-Based Pairwise Similarity metric (WEPS) (Terragni et al., 2021c) (a validated metric originally designed for assessing the similarity between word embeddings), we first define the similarity between the top- N images that describe any two topics t_i and t_j as follows:

$$\text{IEPS}_{\text{pair}}(t_i, t_j) = \frac{1}{N^2} \sum_{v \in t_i} \sum_{u \in t_j} \text{cosine}(e_v, e_u) \quad (5)$$

where e_v , and e_u denote the image embeddings associated with images v and u respectively.

Next, to estimate the overall IEPS score for a topic model with k topics, we compute the pairwise similarities between all pairs of topics and then aggregate them into a single score. Let S denote the set of all pairs of topics in the model, i.e., $S = (t_i, t_j) \mid 1 \leq i < j \leq k$. Then, we can define the overall IEPS score as:

$$\text{IEPS} = \frac{1}{|S|} \sum_{(t_i, t_j) \in S} \text{IEPS}_{\text{pair}}(t_i, t_j) \quad (6)$$

This similarity metric ranges between $[0,1]$ with a lower score indicating more diverse topics.

³We use the implementations of these metrics provided in the OCTIS library (Terragni et al., 2021a).


MS COCO	VIST	T4SA	MMHS150K	HC-4chan	MEWA
					
there are some children playing a soccer game	Having a good time bonding and talking.	RT @AmBlujay: This is why I won't interfere in people's relationships	carol really said f*ck yall I'm a d*ke and I'm here to save the universe	Yes goyim! Don't fight it! It's inevitable	Flightradar24 is a Swedish internet-based service that...

Table 1: Dataset document examples. To save space, we only show the first terms of the document from MEWA. **Disclaimer:** HC-4chan and MMHS150K contains hateful textual and graphic elements.

Dataset	# Docs	Vocab size	Avg. Len	Domain	Data collection	Source
MS COCO	30,000	2,000	10.46	Object recognition	Tag-matching & crowdsourcing	Flickr
VIST	30,000	2,000	9.99	Visual storytelling	Crowdsourcing	Flickr
T4SA	30,000	2,000	6.99	Sentiment analysis	Automatic sentiment classification	Twitter
MMHS150K	30,000	2,000	6.03	Hateful content detection	Keywords-matching & crowdsourcing	Twitter
HC-4chan	17,866	1,993	17.43	Hateful content detection	Human annotations & automatic detection	4chan
MEWA	18,592	2,000	2,424.83	Information retrieval	Crowdsourcing	Wikipedia

Table 2: Properties of the datasets used. Vocab size is restricted to 2000 words to speed computation

5. Datasets

We propose a new benchmark for multimodal topic modeling comprising six diverse datasets. For illustration, Table 1 presents a sample document from each dataset, while the key properties of each dataset are shown in Table 2. We want to alert our readers that some datasets include hateful textual and graphic elements. The datasets are:

MS COCO (Microsoft Common Objects in Context) (Lin et al., 2014) is a popular dataset for image captioning, object detection, and image segmentation tasks, consisting of over 200K images labeled with bounding boxes and category labels for more than 80 object categories, as well as 5 captions per image. The images were sourced from Flickr and annotated by crowd workers. We used a randomly selected subset of 30K multimodal documents (image-caption pairs).

VIST (Visual Storytelling Dataset) (Huang et al., 2016) is a dataset of multimodal stories comprising sequences of images with corresponding descriptions. All images come from Flickr, and the textual stories were crowd-sourced. For our experiments, we again randomly selected from VIST 30K multimodal documents, each consisting of an image along with the description of the portion of the story associated with the image.

T4SA (Vadicamo et al., 2017) is a large-scale Twitter dataset designed for **Sentiment Analysis**. It contains textual and multimodal data obtained through the TwitterAPI and each tweet is automatically annotated with its sentiment polarity. In our experiments, we used a random sample of 30K multi-

modal tweets balanced across sentiment classes.

MMHS150K (Gomez et al., 2020) is a hate speech dataset consisting of 150K Twitter multimodal documents. Each document was then labeled by crowdsourcing with the particular community that was attacked, such as racist, sexist, or homophobic. Once again, we selected 30K random documents.

HC-4chan (Hateful content on 4chan) (González-Pizarro and Zannettou, 2023) contains posts, phrases, and images containing hateful and discriminatory content. It consists of 21K images, identified as presenting Antisemitic/Islamophobic content by CLIP. We removed near-duplicate images and documents with no text, resulting in a dataset of about 18K multimodal documents.

MEWA (Multimodal English Wikipedia Articles) (Zosa and Pivovarova, 2022a) comprises English Wikipedia Articles from the Wikipedia Comparable Corpora⁴, aligned with images from the Wikipedia-based Image Text dataset (WIT) (Srinivasan et al., 2021). Each document consists of a complete English Wikipedia Article and its corresponding image. For our analysis, we used the publicly available subset of 18.5K documents⁵.

6. Experiment Setup

Textual Baselines: We consider LDA (Blei et al., 2003b), ZeroShotTM (Bianchi et al., 2021b), and CombinedTM (Bianchi et al., 2021a) as strong textual baselines. For LDA, we use the OCTIS (Ter-

⁴linguatoools.org/tools/corpora/wikipedia-comparable-corpora/

⁵https://github.com/ezosa/M3L-topic-model/tree/master/data

ragni et al., 2021b) implementation. The parameters controlling the document-topic and word-topic distribution for LDA are estimated during training, as in prior work (Bianchi et al., 2021a). For the neural topic models, ZeroShotTM and CombinedTM, we utilize their publicly available implementations⁶. In training these models, we employ the Adam optimizer and apply a 20% dropout rate.

Configurations: We encode text and images with OpenAI’s CLIP (Radford et al., 2021a), which captures content similarity across modalities. We use the text and image encoder of clip-ViT-B-32, which is available in the SBERT’s library⁷ (Reimers and Gurevych, 2019b). For hyperparameter settings, we follow (Bianchi et al., 2021b) and (Zosa and Pivovarova, 2022a). We train models for 100 epochs, computing all the metrics for 25, 50, 75, and 100 topics. Results for each metric are averaged over 5 random seeds. The data is preprocessed following (Bianchi et al., 2021b). We restrict the vocabulary size to the top 2,000 most frequent terms to speed up computation. We remove English stopwords by using the NLTK library⁸. We also remove punctuation and digits as suggested by prior work.

For the development of the Multimodal-ZeroShotTM model, we adapt ZeroShotTM (Bianchi et al., 2021b) and rely on the original implementation. Consistent with the original setup, our inference network structure comprises one fully connected hidden layer followed by a softplus layer with 100 dimensions.

For Multimodal-Contrast model, we adapt M3L-Contrast (Zosa and Pivovarova, 2022a) and base our code on the author’s original implementation⁹. We use a batch size of 32, set the temperature τ to 0.07, and the contrastive weight ω to 100, as in the original model’s configurations.

Topic descriptors: VAE-based neural topic models like our Multimodal-ZeroShotTM and Multimodal-Contrast obtain the representative keywords of each topic from the topic-vocab weight matrix used for reconstructing the BoW. However, Multimodal-Contrast does not reconstruct image features (see Figure 2), so a different approach is needed to obtain the most relevant images per topic. To this end, after model training, we rely on the document-topic distributions associated with the input documents and select the images of the N documents with the highest contribution for each topic. We use this same approach for Multimodal-ZeroShotTM to ensure a fair comparison. All our experiments are run considering the top 10 words and top 10 images per topic. This de-

cision is based on prior work (Bianchi et al., 2021a; Ding et al., 2018; Hoyle et al., 2021; Li et al., 2023; Newman et al., 2010), who identified that the top 10 terms typically account for about 30% of the topic mass, providing sufficient information to determine the subject area and distinguish one topic from another (Newman et al., 2010). We apply IEC and IEPS over 10 images, maintaining consistency with the most respected prior work.

Dataset size: To ensure computational feasibility for our extensive experiments, we restricted the datasets size to 30K documents. Notice that testing on 30K documents considerably outnumbered assessments of previous topic modeling algorithms, such as CombinedTM (Bianchi et al., 2021a) and our baseline, M3L-Contrast (Zosa and Pivovarova, 2022b), which were initially assessed using a sample of only 20K documents.

Topic Models Overlap: Two topic models might exhibit similar coherence and diversity, but actually generate different topics. So, as part of our systematic comparison between Multimodal-ZeroShotTM and Multimodal-Contrast, we also explore the overlap between the topics they generate. Specifically, given a pair of topic models, we construct a topic similarity matrix based on the most relevant keywords using the I-RBO diversity metric (Bianchi et al., 2021a), ranging in [0-1] with higher scores indicating more substantial topic overlap. Then, to align the topics between models, we apply the Hungarian method (Kuhn, 1955). In the results, we report the mean M and standard deviation SD of the topic overlap between models across multiple datasets, numbers of topics, and random seeds.

Visual learned features: In order to reconstruct image embeddings from the input, Multimodal-ZeroShotTM uses a weight topic-image features matrix. After training the model, analyzing such structure can provide valuable insights on the relevance of specific visual attributes to each topic as well as into the neural model’s limitations. In our experiments, we use a CLIP Guided Diffusion model (Dhariwal and Nichol, 2021)¹⁰ to generate an image per topic given the weight topic-image feature matrix.

User Study: We conduct a user study with two core objectives. Firstly, we validate our proposed metrics, IEC and IEPS, ensuring their alignment with human judgments. Secondly, we perform a qualitative analysis to discern variations between our proposed models. Nine computer scientists participated in the study, evaluating the coherence and diversity of keyword and image sets generated by our multimodal solutions.

For coherence evaluation, our approach, inspired by prior studies (Aletras and Stevenson,

⁶<https://github.com/MilaNLPProc/contextualized-topic-models>

⁷<https://www.sbert.net/>

⁸<https://www.nltk.org/>

⁹<https://github.com/ezosa/M3L-topic-model>

¹⁰<https://github.com/nerdyrodent/CLIP-Guided-Diffusion>

2013; Hoyle et al., 2021), involves presenting participants with either 10 keywords or 10 images representing a topic. Participants use a 5-point Likert scale, where higher scores indicate higher similarity, to gauge the relatedness between these topics’ descriptors. For diversity evaluation, participants assess the similarity between two topics at a time, each represented by a set of 10 keywords or 10 images, using the same Likert scale, with higher values indicating greater similarity (i.e., lower diversity). We restrict the evaluation to topics generated solely from the MS COCO dataset due to time constraints, supported by our pilot study’s findings.

The inter-annotator agreement (IAA) is computed as the average Spearman correlation between each respondent’s scores and the averages of scores from other respondents. For coherence, the IAA is 0.71 and 0.62 for keywords and images, respectively, indicating a strong agreement among participants. For diversity, it is 0.75 and 0.84, pointing to a very strong consensus. To assess the alignment between humans and our proposed metrics, IEC, and IEPS, we calculate the mean Spearman correlation between the automatic metrics and human ratings, following the methodology suggested by prior work (Aletras and Stevenson, 2013; Hoyle et al., 2021).

7. Results

As an example, Table 3 presents topics that can be retrieved from MS COCO using our multimodal solutions. We report the models’ performance below.

Topic Coherence and Diversity: Table 4 shows the overall performance of the models in terms of coherence and diversity of the topics’ descriptors.

For the metrics assessing the coherence of the textual descriptors, as expected, LDA performs worse, while both Multimodal-ZeroShotTM and Multimodal-Contrast perform similarly to ZeroShotTM and CombinedTM, indicating that processing the images of the corpus does not influence the coherence of the textual descriptors. Interestingly, we also observe subtle differences in the NPMI and C_v scores between Multimodal-ZeroShotTM and Multimodal-Contrast. As shown in Table 5, depending on the metrics and dataset combinations, one model outperforms the other without a clear winner. For instance, Multimodal-ZeroShotTM can generate more coherent topics than Multimodal-Contrast in the VIST, T4SA, and MMHS150 datasets but not in MS COCO and MEWA. As for image descriptors, the top relevant images of each topic identified by Multimodal-Contrast seem more related (i.e., higher IEC) than those selected

Model	Topics descriptors
M-Z	snow, person, skis, covered, mountain, ski, snowy, slope, hill, skiing 
M-C	snow, covered, skis, slope, hill, snowy, ski, mountain, skiing, snowboard 
M-Z	mirror, toilet, bathroom, sink, wall, tub, shower, window, door, bath 
M-C	bathroom, toilet, mirror, sink, tub, shower, bath, wall, tiled, door 
M-Z	water, boat, body, beach, ocean, river, boats, lake, sandy, shore 
M-C	beach, water, boat, body, ocean, boats, kite, river, kites, sandy 

Table 3: Topics retrieved using Multimodal-ZeroShotTM (M-Z) and Multimodal-Contrast (M-C) on MS COCO

Metrics	Coherence			Diversity		
	NPMI	C_v	WE IEC	TD	I-RBO	IEPS
LDA	-0.14	.39	.15	.84	.97	
CombinedTM	.04	.52	.21	.50	.96	
ZeroShotTM	.03	.51	.22	.60	.98	
Multimodal-Contrast	.04	.50	.22	.67	.47	.94
Multimodal-ZeroShotTM	.03	.51	.22	.56	.60	.98

Table 4: Average topic coherence and diversity scores across all datasets. Top scores are bold.

by Multimodal-ZeroShotTM overall and across datasets.

Table 4 and 6 report the performance of the models in terms of their ability to generate diverse topics. LDA is one of the top performers, but it scored the lowest for coherence by a wide margin (see Table 4 and 5).

Dataset	MS COCO				VIST				T4SA				MMHS150K				HC-4chan				MEWA			
Metrics	τ	ϕ	α	γ	τ	ϕ	α	γ	τ	ϕ	α	γ	τ	ϕ	α	γ	τ	ϕ	α	γ	τ	ϕ	α	γ
LDA	-.10	.37	.14		-.23	.34	.13		-.25	.36	.13		-.17	.31	.16		-.21	.40	.13		.10	.57	.20	
CombinedTM	.13	.55	.22		-.01	.42	.21		.02	.45	.22		.00	.45	.22		-.07	.58	.21		.13	.64	.20	
ZeroShotTM	.12	.54	.23		-.02	.39	.23		.02	.44	.23		.00	.44	.22		-.07	.57	.22		.14	.67	.22	
Multimodal-Contrast	.13	.56	.23	.75	-.03	.35	.23	.69	-.02	.38	.20	.59	-.01	.40	.23	.65	-.02	.54	.19	.65	.16	.77	.20	.68
Multimodal-ZeroShotTM	.12	.54	.23	.66	-.01	.39	.23	.56	.01	.44	.23	.47	.00	.45	.22	.56	-.07	.58	.21	.57	.14	.66	.21	.54

Table 5: Topic’s coherence scores per dataset. We used the following abbreviations: NPMI (τ), C_v (ϕ), WE (α), and IEC (γ). Average results over 4 number of topics ($K = 25, 50, 75, 100$), where the results for each K are averaged over 5 random seeds. We bold the highest scores.

Dataset	MS COCO			VIST			T4SA			MMHS150KK			HC-4chan			MEWA		
Metrics	TD	I-RBO	IEPS	TD	I-RBO	IEPS	TD	I-RBO	IEPS	TD	I-RBO	IEPS	TD	I-RBO	IEPS	TD	I-RBO	IEPS
LDA	.75	.98		.93	1.00		.96	1.00		.85	.90		.89	1.00		.65	.96	
CombinedTM	.57	.98		.40	.94		.43	.96		.41	.95		.40	.94		.76	.99	
ZeroShotTM	.67	.99		.60	.99		.56	.98		.50	.97		.50	.96		.77	.99	
Multimodal-Contrast	.65	.99	.46	.56	.98	.45	.52	.97	.28	.29	.87	.41	.23	.87	.50	.58	.98	.37
Multimodal-ZeroShotTM	.68	.99	.47	.60	.99	.48	.56	.98	.34	.50	.97	.45	.50	.96	.53	.77	.99	.38

Table 6: Diversity scores of the top keywords and images. Average results over 4 number of topics ($K = 25, 50, 75, 100$), with results for each K averaged over 5 random seeds. We bold best scores.

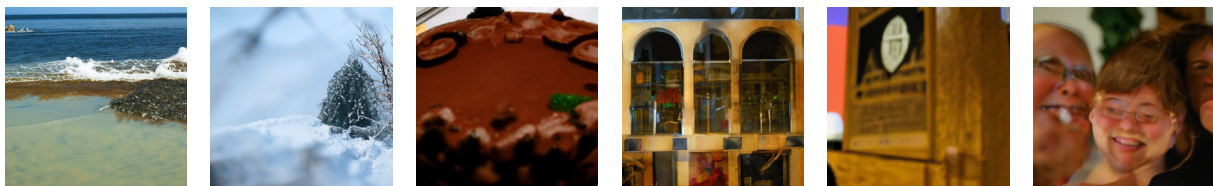


Figure 3: Images from a CLIP-Guided Diffusion Model over the latent space. Topics from left to right are {water, lake, sand, beach}, {snow, tree, Christmas, white}, {cake, made, candles, birthday}, {art, building, glass, amazing, architecture}, {students, graduation, speech, school}, and {family, together, happy, whole}.

We also observe that for all the datasets, the keywords generated by Multimodal-ZeroShotTM are significantly more diverse than those generated by Multimodal-Contrast. Finally, we find that the images representing the topics are more diverse (i.e., lower IEPS) in Multimodal-Contrast than in Multimodal-ZeroShotTM.

Plausibly, the observed superiority of Multimodal-Contrast over Multimodal-ZeroShotTM in terms of image coherence and diversity can be attributed to their different training objectives. Multimodal-Contrast incorporates a Contrastive Learning loss (Chopra et al., 2005) to maximize the similarity between positive pairs (text-image pairs that belong together) while minimizing the similarity between negative pairs. Training the model to explicitly differentiate between related and unrelated pairs may stimulate the model to learn more discriminative image representations (Yu et al., 2022), which can better support similar judgments involved in creating more coherent and diverse topic models. More speculatively, the reason why this works for image but not for text descriptors

may be due to the fact that although pre-trained multimodal representations (e.g., CLIP) map data from different modalities into the same space, embeddings from different modalities are located in separate regions (Liang et al., 2022) and therefore could be influenced differently by Contrastive Learning.

Topic Overlap And Visual Features: Table 7 shows the topic’s overlap between ZeroShotTM and our proposed models.

Models		M	SD
ZeroShotTM	Multimodal-Contrast	.22	.16
ZeroShotTM	Multimodal-ZeroShotTM	.50	.23
Multimodal-Contrast	Multimodal-ZeroShotTM	.21	.16

Table 7: Topic’s overlap between models across all datasets and number of topics.

Remarkably, while ZeroShotTM and Multimodal-ZeroShotTM exhibit similar performance (i.e., for coherence and diversity), the generated topics only partially overlap ($M = .50$; $SD = .23$). In other words, these two

models generate some topics that are similar to each other, but also unique ones. Moreover, Multimodal-Contrast produces topics with significantly less keyword overlap when compared to both ZeroShotTM ($M = .22$; $SD = .16$) and Multimodal-ZeroShotTM ($M = .21$; $SD = .16$).

Figure 3 showcases images generated by the CLIP-Guided diffusion model considering the topic-image feature matrix γ from Multimodal-ZeroShotTM, with topics extracted from VIST. Promisingly, the generated images seem to align well with the topic’s descriptors, capturing abstract and complex concepts (like happy family).

User Study: Table 8 displays human ratings for topics from our models, while Table 9 presents corresponding automatic scores.

Ratings scores	Coherence				Diversity			
	Keywords		Images		Keywords		Images	
	M	SD	M	SD	M	SD	M	SD
Multimodal-ZeroShotTM	3.84	1.05	4.66	0.60	1.28	0.64	1.42	0.95
Multimodal-Contrast	4.05	0.95	4.51	0.79	1.23	0.67	1.43	0.95

Table 8: Mean and standard deviation of rating scores for evaluating coherence (higher values indicate higher coherence) and diversity (lower values indicate higher diversity) of sets of keywords and images generated by our models.

Metrics	NPMI	C_v	WE	IEC	TD	I-RBO	IEPS
Multimodal-ZeroShotTM	.09	.48	.23	.75	1.00	1.00	0.43
Multimodal-Contrast	.06	.44	.25	.70	1.00	1.00	0.44

Table 9: Automatic coherence and diversity scores from the topics used in our user study.

To validate our metrics, IEC and IEPS, we calculated their Spearman correlation with human ratings. The results showed robust and statistically significant positive correlations: IEC ($r(27) = .45$, $p < .001$) and IEPS ($r(27) = .44$, $p < .001$), confirming the reliability of our metrics. According to annotators, the images representing the topics in Multimodal-ZeroShotTM were more coherent and diverse compared to those generated by Multimodal-Contrast. This observation aligns with the IEC and IEPS scores for these topics (see Table 9), where Multimodal-ZeroShotTM emerged as the superior model. This result reinforces the credibility of our proposed metrics and underscores their potential for evaluating multimodal topic models. Human evaluators reported that Multimodal-Contrast generated topics with more coherent and diverse keywords compared to its multimodal counterpart, a trend supported by the WE metric. In contrast, NPMI and C_v metrics, relying on the reference corpus (in this case, MS

COCO), favored Multimodal-ZeroShotTM. Finally, automatic metrics indicated a tie and perfect score for topic keyword diversity, possibly due to the limited topic subsets in the user study’s diversity tasks and the criteria used by TD and I-RBO, which assess diversity based on exact keyword overlap.

8. Conclusions and Future Work

We present the first systematic evaluation of neural multimodal topic modeling, considering multiple non-homogeneous datasets and a comprehensive set of evaluation metrics. In particular, we contribute a repository of corpora that vary in document size, source, and underlying task/domain, along with two novel metrics to assess topic image descriptors’ coherence and diversity, which we validated in a preliminary user study. We apply the resulting evaluation framework to compare two novel multimodal topic modeling methods that we developed by adapting current SOTA architectures. Overall, our results indicate that ensemble and hybrid solutions should be explored in the future, for instance, by either merging the output of different models or by combining different components in more complex loss functions. Leveraging GPT-like systems (Achiam et al., 2023) is also a potential direction for future work, but first, the formidable limitation in their input size (Bubeck et al., 2023) must be addressed. In another short-term direction we plan to assess whether the topic-image feature matrix γ (Eq.1) can benefit multimodal text classification and document similarity.

Limitations

As in any study, ours has limitations that need to be considered. First, we used datasets that are only available in English, which might restrict the generalizability of our findings. Moving forward, we aim to tackle this limitation by including datasets in various languages and exploring multilingual models that handle multiple languages simultaneously. Secondly, we focus our topics’ evaluation on their coherence and diversity. Future work should identify the quality of the results based on other aspects, such as document coverage (i.e., how well documents match their assigned topics) and topic model comprehensiveness (i.e., how thoroughly the model covers the topics appearing in the corpus). These aspects are challenging to assess when ground truth is unavailable. Finally, future work should explore how hyperparameters (e.g., dropout rate, weight for KL divergence loss) impact neural multimodal topic models.

Ethics Statement

Our neural multimodal topic models are intended solely for research purposes. Any use of these models or their derived artifacts outside research contexts should be authorized accordingly. We use datasets that are publicly available. Users must be aware of the potential risks associated with using topic modeling algorithms. Topic models may amplify biases present in the data (e.g., if the dataset contains hateful content, the generated topics can perpetuate those discriminatory practices). Users also need to consider the biases and limitations of text and image encoders (e.g., CLIP). Moreover, neural topic models lack transparency and interpretability, meaning it becomes challenging to understand how the model arrives at particular topics.

Acknowledgements

We would like to express our profound gratitude, first and foremost, to Aditya Chinchure, Sahithya Ravi and Raymond Li for their invaluable feedback and constructive critiques at various stages of this research. We are also thankful to Debora Nozza and Federico Bianchi for providing the initial ideas and engaging in the preliminary discussions that significantly shaped the direction of our study. Additionally, our sincere thanks go to the anonymous reviewers whose meticulous comments and suggestions have greatly enhanced the quality of this paper.

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Nous remercions le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) de son soutien.

This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) and the Digital Research Alliance of Canada (<https://alliancecan.ca>).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Alvenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nikolaos Aletras and Mark Stevenson. 2013. *Evaluating topic coherence using distributional semantics*. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS*

2013) – Long Papers, pages 13–22, Potsdam, Germany. Association for Computational Linguistics.

Jingwen Bian, Yang Yang, and Tat-Seng Chua. 2013. *Multimedia summarization for trending topics in microblogs*. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, CIKM '13*, page 1807–1812, New York, NY, USA. Association for Computing Machinery.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. *Pre-training is a hot topic: Contextualized document embeddings improve topic coherence*. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. *Cross-lingual contextualized topic models with zero-shot learning*. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

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

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003b. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. 2017. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. *Sparks of artificial general intelligence: Early experiments with gpt-4*.

Sophie Burkhardt and Stefan Kramer. 2019. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *J. Mach. Learn. Res.*, 20(131):1–27.

- S. Chopra, R. Hadsell, and Y. LeCun. 2005. [Learning a similarity metric discriminatively, with application to face verification](#). In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1.
- Prafulla Dhariwal and Alexander Nichol. 2021. [Diffusion models beat gans on image synthesis](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc.
- Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.
- 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*, pages 830–836, Brussels, Belgium. Association for Computational Linguistics.
- Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016. [Using word embedding to evaluate the coherence of topics from twitter data](#). In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16*, page 1057–1060, New York, NY, USA. Association for Computing Machinery.
- Amon Ge, Hyeju Jang, Giuseppe Carenini, Kendall Ho, and Young Ji Lee. 2019. Octvis: ontology-based comparison of topic models. In *2019 IEEE Visualization Conference (VIS)*, pages 66–70. IEEE.
- Raul Gomez, Jaume Gibert, Lluís Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
- Felipe González-Pizarro and Savvas Zannettou. 2023. [Understanding and detecting hateful content using contrastive learning](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):257–268.
- Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Geoffrey E Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.
- Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. *Advances in Neural Information Processing Systems*, 34:2018–2033.
- Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. [Visual storytelling](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.
- H. W. Kuhn. 1955. [The hungarian method for the assignment problem](#). *Naval Research Logistics Quarterly*, 2(1-2):83–97.
- Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.
- Raymond Li, Felipe Gonzalez-Pizarro, Linzi Xing, Gabriel Murray, and Giuseppe Carenini. 2023. [Diversity-aware coherence loss for improving neural topic models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1710–1722, Toronto, Canada. Association for Computational Linguistics.
- Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. [Mind the gap: Understanding the modality gap in multimodal contrastive representation learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 17612–17625. Curran Associates, Inc.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

- David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, page 100–108, USA. Association for Computational Linguistics.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Jessica Peter, Steve Szigeti, Ana Jofre, and Sara Diamond. 2015. Topicks: Visualizing complex topic models for user comprehension. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 207–208. IEEE.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021b. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.
- Nils Reimers and Iryna Gurevych. 2019a. [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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019b. 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)*, pages 3982–3992.
- 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 '15, page 399–408, New York, NY, USA. Association for Computing Machinery.
- Dhanya Sridhar, Hal Daumé III, and David Blei. 2022. [Heterogeneous supervised topic models](#). *Transactions of the Association for Computational Linguistics*, 10:732–745.
- Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. [Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA. Association for Computing Machinery.
- Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations*.
- Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021a. Octis: comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270.
- Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021b. [OCTIS: Comparing and optimizing topic models is simple!](#) In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270, Online. Association for Computational Linguistics.
- Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021c. Word embedding-based topic similarity measures. In *International Conference on Applications of Natural Language to Information Systems*, pages 33–45. Springer.
- Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi. 2017. [Cross-media learning for image sentiment analysis in the wild](#). In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 308–317.
- En Yu, Zhuoling Li, and Shoudong Han. 2022. Towards discriminative representation: Multi-view trajectory contrastive learning for online multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8834–8843.

- Lili Yu, Dániel Simig, Colin Flaherty, Armen Aghajanyan, Luke Zettlemoyer, and Mike Lewis. 2023. [Megabyte: Predicting million-byte sequences with multiscale transformers](#).
- Huakui Zhang, Cai Yi, Bingshan Zhu, Haopeng Ren, and Qing Li. 2022. [Multimodal topic modeling by exploring characteristics of short text social media](#). *IEEE Transactions on Multimedia*, pages 1–1.
- He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021. Topic modelling meets deep neural networks: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4713–4720. International Joint Conferences on Artificial Intelligence Organization. Survey Track.
- Elaine Zosa and Lidia Pivovarova. 2022a. Multilingual and multimodal topic modelling with pre-trained embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4037–4048.
- Elaine Zosa and Lidia Pivovarova. 2022b. [Multilingual and multimodal topic modelling with pre-trained embeddings](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4037–4048, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A. Performance of Multimodal-ZeroShotTM across different λ values

Table 10 presents the performance of Multimodal-ZeroShotTM for varying values of λ . This parameter adjusts the balance between the textual and image feature reconstruction losses, as defined in Equation 1. For our main experiments, λ was set to 1.

Our results indicate that higher λ values improve the coherence and diversity of the images representing a topic. However, this improvement is accompanied by a slight decrease in the coherence and diversity of the topics' most relevant keywords.

	Coherence			Diversity			
	NPMI	C_v	WE	IEC	TD	I-RBO	IEPS
$\lambda = 1$.03	.51	.22	.56	.60	.98	.44
$\lambda = 60$.03	.50	.22	.66	.57	.96	.41
$\lambda = 120$.02	.49	.21	.69	.54	.95	.40
$\lambda = 240$.02	.49	.21	.70	.52	.95	.40

Table 10: Quality of topics for different λ values in Multimodal-ZeroShotTM. Averages are calculated across multiple datasets for 4 number of topics ($K = 25, 50, 75, 100$), where the results for each K are averaged over 5 random seeds. Top scores are bold.

B. Using a Different Contextualized Representation

We also compare the performance of neural topic models using a different text encoder. We employ the sentence-transformer model `all-mpnet-base-v2`, available in the SBERT library. Table 11 displays the resulting performance of the neural topic models.

	Coherence			Diversity			
	NPMI	C_v	WE	IEC	TD	I-RBO	IEPS
CombinedTM	.03	.51	.21	.56	.96		
ZeroShotTM	.04	.52	.23	.58	.98		
Multimodal-Contrast	.04	.51	.22	.66	.48	.94	.41
Multimodal-ZeroShotTM	.03	.51	.22	.56	.58	.98	.44

Table 11: Performance of neural topic models using a different contextualized text encoder (i.e., `all-mpnet-base-v2`). Averages are calculated across datasets for 4 number of topics ($K = 25, 50, 75, 100$), with each K averaged over 5 random seeds. Best scores are highlighted in bold.

C. Computing Infrastructure

Our experiments were conducted on an NVIDIA A100 GPU, with 20 GB of memory and 12 cores of an AMD Milan 7413 processor. Although previous studies have shown that neural topic models can be run on less performant hardware, we chose this high-performance computing infrastructure to ensure efficient data processing.

D. Runtime

Previous research has demonstrated that vocabulary size significantly impacts the computational time of VAE-based neural topic models (Bianchi et al., 2021a). Consequently, in line with prior studies, we restrict the maximum number of terms in the Bag-Of-Word reconstructions to 2,000. We select the 2,000 most frequent words in the corpus for this purpose.

To compare the computational efficiency of different models, we report the time taken in seconds to complete one epoch during training. Table 12 presents the required time for each neural topic model to complete one epoch. For LDA, however, we report the average training time. Our findings indicate that multimodal neural topic models require approximately 1.5 seconds more per epoch to complete compared to the unimodal ZeroShotTM model. This additional time can be attributed to the larger input sizes and the simultaneous analysis of two modalities in multimodal models. Such a difference in training time is anticipated and justifiable, considering the increased complexity of these models.

	Training time per epoch
LDA	12.49
CombinedTM	7.38
ZeroShotTM	7.27
Multimodal-Contrast	8.70
Multimodal-ZeroShotTM	8.33

Table 12: Time in seconds required to complete one epoch. Averages are calculated across all datasets for 4 sets of topics ($K = 25, 50, 75, 100$), with results for each K averaged over 5 random seeds. The lowest training time is bold.