

XTRA: Cross-Lingual Topic Modeling with Topic and Representation Alignments

Tien Phat Nguyen^{1*}, Vu Minh Ngo^{1*}, Tung Nguyen¹, Linh Van Ngo^{1†},
Duc Anh Nguyen¹, Sang Dinh¹, Trung Le²,

¹ Hanoi University of Science and Technology, Vietnam,

² University of Monash, Australia,

Abstract

Cross-lingual topic modeling aims to uncover shared semantic themes across languages. Several methods have been proposed to address this problem, leveraging both traditional and neural approaches. While previous methods have achieved some improvements in topic diversity, they often struggle to ensure high topic coherence and consistent alignment across languages. We propose **XTRA** (**Cross-Lingual Topic Modeling with Topic and Representation Alignments**), a novel framework that unifies Bag-of-Words modeling with multilingual embeddings. XTRA introduces two core components: (1) representation alignment, aligning document-topic distributions via contrastive learning in a shared semantic space; and (2) topic alignment, projecting topic-word distributions into the same space to enforce cross-lingual consistency. This dual mechanism enables XTRA to learn topics that are interpretable (coherent and diverse) and well-aligned across languages. Experiments on multilingual corpora confirm that XTRA significantly outperforms strong baselines in topic coherence, diversity, and alignment quality. Code and reproducible scripts are available at <https://github.com/tienphat140205/XTRA>.

1 Introduction

Identifying latent thematic structures within large text corpora is a central goal in computational linguistics, with topic modeling (TM) serving as a foundational technique (Hofmann, 1999; Blei et al., 2003). Extending this capability to the multilingual setting has led to the development of Cross-Lingual Topic Modeling (CLTM), which aims to uncover shared latent topics across languages despite lexical and structural differences (Ni et al., 2009; Mimno et al., 2009; Yuan et al., 2018; Wu et al., 2020, 2023a). CLTM models are essential for bridging

[†]Equally contributed.

[†]Corresponding author: linhvn@soict.hust.edu.vn

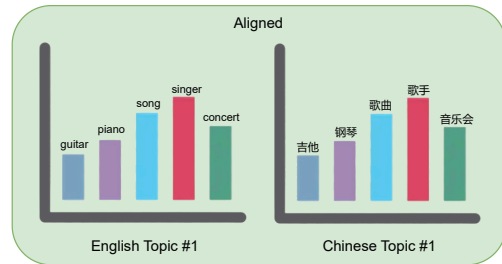


Figure 1: An example of cross-lingual topic alignment: both English and Chinese word clusters describe the shared theme of music.

cultural and linguistic divides, providing a means to understand and compare shared narratives as well as distinct global perspectives. As illustrated in Figure 1, cross-lingual topic models aim to uncover semantically similar themes across languages, such as English and Chinese word clusters that both describe the concept of music. This illustrates how cross-lingual topic models can discover thematic overlap despite significant differences in surface forms.

Despite significant progress, developing robust and broadly applicable Cross-Lingual Topic Models (CLTMs) continues to pose substantial challenges. Dictionary-based approaches remain attractive among the most common strategies due to their simplicity and interpretability. However, these methods are fundamentally limited by the coverage and quality of bilingual lexical resources, which are often insufficient, especially when dealing with low-resource languages or domain-specific vocabularies. This low-coverage problem has been consistently observed across multiple prior works (Shi et al., 2016; Yuan et al., 2018; Wu et al., 2020), where the lack of comprehensive term mappings hampers the alignment of topics across languages and reduces the overall semantic fidelity of the model (Wu et al., 2023a). In addition to lexical

En Topic#1:	shampoo	lotion	fragrance	clay	powder
ZH Topic#1:	婴儿	墨水	涂	糊	干
Translation#1:	baby	ink	smear	paste	dry
En Topic#2:	violin	orchestra	rhythm	performance	classical
ZH Topic#2:	按摩	枕头	鞋	滑	脖子
Translation#2:	massage	pillow	shoe	slippery	neck
En Topic#3:	translation	translator	original	poem	English
ZH Topic#3:	译文	文言文	原文	思录	改动
Translation#3:	T. Trans	C. Chinese	Ori. Text	reflection	revision

Table 1: Example of misaligned topics generated by InfoCTM across English and Chinese. The words grouped under each topic differ in semantic coherence between the two languages.

sparsity, these CLTMs often suffer from repetitive topic collapse, where multiple topics converge toward semantically similar word distributions (Wu et al., 2023a). This redundancy undermines topic diversity and interpretability, particularly in multilingual settings where fine-grained distinctions are critical.

Advanced neural topic models (e.g., InfoCTM (Wu et al., 2023a)) address redundancy and topic collapse by applying contrastive objectives on the decoder’s topic–word distributions (β). These methods increase topic diversity and yield more coherent topics within each language. Yet two limitations remain. First, they often rely on language-specific encoders or independent parameterization, which restricts scalability and leaves β only loosely aligned across languages. More critically, most cross-lingual topic models (Wu et al., 2023a, 2020) neglect the refinement of document–topic proportions (θ) across languages. This omission is consequential: without cross-lingual consistency at the θ level, topic mixtures cannot be compared reliably, weakening semantic alignment across languages. As observed in Table 1, English–Chinese topics may each appear internally coherent yet still fail to correspond semantically, highlighting the limitations of current approaches.

To address these challenges, we propose **XTRA** (Cross-Lingual Topic Modeling with Topic and Representation Alignments), a unified framework designed for robust cross-lingual topic discovery that moves beyond complex language-specific encoders. Instead, XTRA utilizes lightweight MLPs to project Bag-of-Words (BoW) inputs into a shared space, where a common encoder effectively captures both semantic structure and crucial cross-lingual alignment signals. One of the key ideas of our model is a clustering-guided contrastive learning objective that refines document–topic distributions. By clustering documents from different languages into semantically coherent groups, XTRA identifies latent cross-lingual themes. A contrastive

loss then pulls together documents with similar thematic content irrespective of language while pushing apart those with dissimilar themes, fostering topic distributions that are both discriminative and consistent across languages.

In parallel, XTRA introduces a novel approach to align topic–word distributions across languages. Instead of relying on vocabulary matching or dictionary lookups, we learn to project topic–word distributions into a shared semantic space using trainable transformation layers. In this space, a second contrastive objective brings semantically equivalent topics across languages closer together. This dual-contrastive design allows XTRA to learn both high-quality document–topic structures and semantically aligned topic–word meanings without any parallel supervision. Our main contributions are:

- We propose **XTRA**, a contrastive cross-lingual topic modeling framework that directly leverages powerful multilingual contextual embeddings via a shared encoder.
- We design a clustering-based contrastive learning strategy that improves document–topic distributions by aligning documents sharing similar semantics across languages.
- We develop a semantic-space alignment technique for topic–word distributions using projection and contrastive loss, enabling robust cross-lingual topic equivalence.
- We conduct extensive experiments on multilingual datasets, showing that XTRA surpasses state-of-the-art CLTM models in topic coherence, topic diversity, and cross-lingual alignment.

2 Preliminaries

2.1 Notations

We model a multilingual corpus comprising D documents in two languages, denoted L_1 and L_2 , with the objective of discovering K shared topics. The corpus is represented as a collection $X = \{x_d\}_{d=1}^D$ of Bag-of-Words (BoW) representations, where each document x_d belongs to either L_1 or L_2 . The vocabulary for L_1 is V_1 of size $|V_1|$, and for L_2 is V_2 of size $|V_2|$. The BoW representation of document d is $x_d \in \mathbb{R}^{|V_\ell|}$, where $\ell \in \{1, 2\}$ indicates

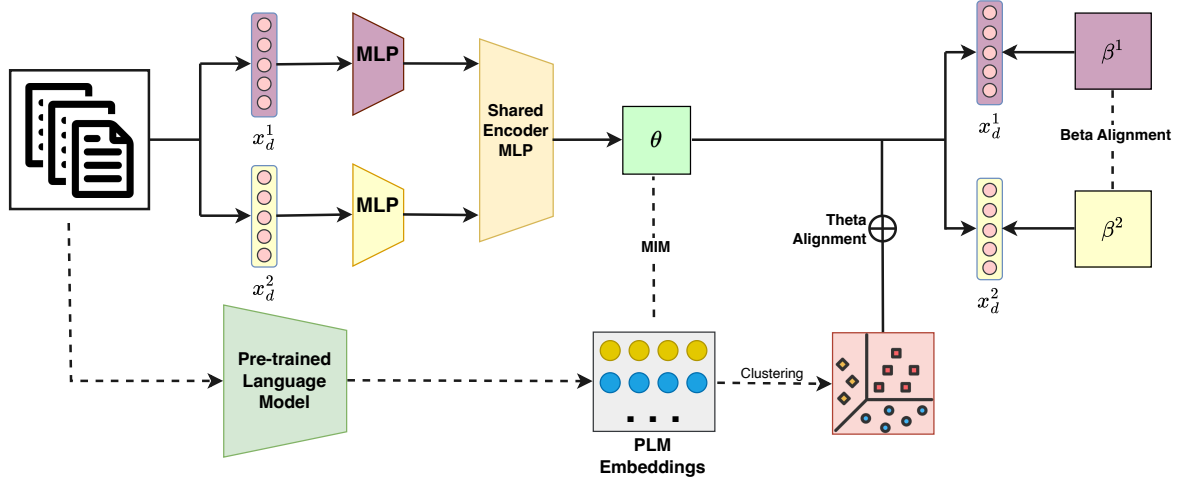


Figure 2: The overall architecture of XTRA. Our method processes language-specific Bag-of-Words inputs (x_d^1 , x_d^2) through dedicated MLPs into a Shared Encoder to estimate document-topic distributions (θ). XTRA proposes a novel dual-alignment approach using Pre-trained Language Model (PLM) embeddings: aligning document-topic distributions (θ) via MIM and clustering (Theta Alignment), and aligning topic-word distributions (β^1 , β^2) semantically (Beta Alignment) to achieve cross-lingual consistency.

the document’s language. A pre-trained multilingual language model (MLM) provides embeddings $x_{d\text{PLM}} \in \mathbb{R}^M$ for each document d , where M is the MLM embedding dimension. Applying a clustering algorithm to $\{x_{d\text{PLM}}\}_{d=1}^D$ produces K clusters, grouping documents by semantic similarity across languages.

For each language L_ℓ where $\ell \in \{1, 2\}$, the topic-word distribution is denoted by $\beta^{(\ell)} \in \mathbb{R}^{|V_\ell| \times K} = (\beta_1^{(\ell)}, \dots, \beta_K^{(\ell)})$, where each $\beta_k^{(\ell)} \in \mathbb{R}^{|V_\ell|}$ represents the word distribution for topic k over the vocabulary V_ℓ , satisfying $\sum_{v \in V_\ell} \beta_{v,k}^{(\ell)} = 1$. Each document x_d is associated with a topic proportion vector $\theta_d \in \mathbb{R}^K$ such that $\sum_{k=1}^K \theta_{d,k} = 1$.

2.2 VAE-based Topic Model

Our foundational structure employs a Variational Autoencoder (VAE). The encoder maps a document’s Bag-of-Words (BoW) representation x_d to parameters (μ, Σ) of a posterior $q(z|x_d) = \mathcal{N}(z|\mu, \Sigma)$. A latent variable z is sampled (via reparameterization (Kingma and Welling, 2013)) from this posterior, with $p(z) = \mathcal{N}(z|\mu_0, \Sigma_0)$ as the prior. The topic proportion is then $\theta_d = \text{softmax}(z)$. The decoder models topic-word distributions $\beta \in \mathbb{R}^{V \times K}$ (e.g., inferred via optimization (Srivastava and Sutton, 2017) or learned embeddings (Dieng et al., 2020; Wu et al., 2023b)) and reconstructs x_d by sampling from

Multinomial($\text{softmax}(\beta\theta_d)$). The training objective is:

$$\mathcal{L}_{\text{TM}} = \frac{1}{D} \sum_{d=1}^D \left[- (x_d)^\top \log(\text{softmax}(\beta\theta_d)) + \text{KL}(q(z|x_d)||p(z)) \right]$$

3 Methodology

We propose XTRA, a novel cross-lingual topic modeling framework whose overall architecture is illustrated in Figure 2. XTRA introduces an innovative inference mechanism for multilingual documents and enhances topic quality through contrastive alignment between topic-word distributions and document-topic distributions. Additionally, we leverage contextualized multilingual embeddings to capture nuanced semantic representations and gain a deeper understanding of the corpus content.

3.1 Enhanced Cross-Lingual Encoder

The encoder transforms document representations into latent topic distributions θ . Prior cross-lingual topic models (e.g., (Wu et al., 2023a)) often use separate encoders for each language, increasing the number of parameters and potentially hindering inherent cross-lingual θ alignment, thus requiring complex post-processing.

We propose an enhanced encoder that uses a single, shared core to process language-specific inputs. Language-specific Bag-of-Words (BoW)

inputs (x_d^ℓ) are first processed by language-specific Multi-Layer Perceptrons (MLPs), which project BoW vectors into a common, language-agnostic space before entering the shared encoder. This reduces parameters and encourages the shared encoder to learn a latent space where θ exhibits better intrinsic cross-lingual alignment.

To further improve θ 's quality and cross-lingual consistency, we leverage large multilingual embeddings (Chen et al., 2024). Inspired by Pham et al. (2024), a contrastive objective aligns θ with this shared semantic space.

This guidance uses an InfoNCE loss (van den Oord et al., 2019):

$$I(\mathbf{X}_{\text{PLM}}; \Theta) \geq \log B + \mathcal{L}_{\text{InfoNCE}}$$

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{D} \sum_{i=1}^D \log \frac{\exp(f(\theta_i, x_{i\text{PLM}}))}{\sum_{\theta' \in B_i} \exp(f(\theta', x_{i\text{PLM}}))}$$

B_i holds sampled topic proportions for document i (positive/negative for $x_{i\text{PLM}}$), drawn from the same batch as x_i with constant size B . $f(\theta, x_{\text{PLM}}) = \frac{\langle \phi_\theta(\theta), x_{\text{PLM}} \rangle}{\|\phi_\theta(\theta)\| \|x_{\text{PLM}}\|}$ (using learnable ϕ_θ) measures similarity between θ and x_{PLM} . Minimizing this loss improves θ 's semantic relevance and cross-lingual alignment by encouraging similarity with corresponding embeddings.

3.2 Topic Distribution Clustering Contrastive Alignment

Aligning document-topic distributions (θ) across languages is crucial but has often been overlooked. We introduce Topic Distribution Clustering Contrastive Alignment, a novel mechanism that explicitly shapes the θ space through cross-lingual clustering and contrastive objectives.

This method achieves cross-lingual θ alignment using document relationships from prior clustering (Figure 3). Documents are clustered via multilingual embeddings (Chen et al., 2024) (see Appendix C for details), identifying related documents across languages without requiring direct translations or parallel data. For a given document's topic distribution, those from the same cluster act as positive examples, while those from different clusters serve as negative ones. The contrastive objective encourages the document's representation to be close to the positive examples and distant from the negative ones.

This direct contrastive pressure on θ steers the encoder towards a latent θ space where proximity

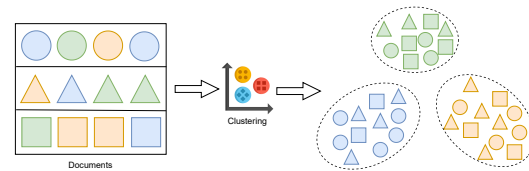


Figure 3: Our clustering-based contrastive alignment illustration. We group similar documents across languages into clusters using multilingual embeddings. Each document is aligned with its cluster via contrastive learning on topic distributions (θ), encouraging cross-lingual consistency in the topic space.

reflects cluster-defined cross-lingual similarity, ensuring similar-themed documents across languages map to similar topic distribution points. This yields more robust, coherent, and aligned document representations. This θ alignment also improves topic interpretability (similar documents load onto similar topic profiles) and aids learning sharper, coherent topic-word distributions (β) via a cleaner decoder signal.

This direct θ alignment is formalized using the multi-positive InfoNCE loss (van den Oord et al., 2019):

$$\mathcal{L}_{\text{Cluster}} = -\frac{1}{D} \sum_{i=1}^D \sum_{\theta_j \in B_i^+} \log S_{ij}$$

Which $S_{ij} = \frac{\exp(g(\theta_i, \theta_j))}{\sum_{\theta' \in B_i} \exp(g(\theta_i, \theta'))}$, B_i denote the set of sampled topic proportions for document i , $B_i^+ \subset B_i$ represent the subset of positive samples θ_i^+ that are semantically related to a given topic proportion θ , and $g(\cdot, \cdot)$ is a similarity function (e.g., cosine similarity). Minimizing $\mathcal{L}_{\text{Cluster}}$ directly enhances the cross-lingual alignment and relevance of the learned document-topic distributions θ .

3.3 Topic-Word Distribution Semantic Alignment

Ensuring semantic equivalence of learned topics across languages, by aligning their topic-word distributions (β), is vital for cross-lingual interpretability. However, directly aligning these high-dimensional topic distributions poses significant challenges, as it requires effective transformations to handle the complexity and discrepancies between languages.

We propose transferring vocabulary item representations from high-dimensional vocabulary space to a lower-dimensional semantic space for alignment. For each language $\ell \in \{1, 2\}$, a learnable

language-specific projection P_ℓ (typically an MLP) maps a topic’s word distribution representation $\beta_k^{(\ell)}$ to a shared D_{sem} -dimensional semantic space. Given $\beta^{(1)}, \beta^{(2)}$, these projections yield K semantic vectors per language, where $y_k^{(\ell)}$ is topic k ’s projected semantic profile in language ℓ .

Alignment is enforced in this shared semantic space via an InfoNCE-based contrastive objective (van den Oord et al., 2019) on projected topic vectors. This maximizes similarity between corresponding topic vectors $(y_k^{(1)}, y_k^{(2)})$ relative to non-corresponding ones $(y_{k'}^{(2)})$, directly aligning their semantic representations. This contrastive learning on projected topic-word distributions also fosters greater topic distinction by pushing non-corresponding topic vectors apart, potentially reducing overlap and improving Topic Uniqueness (TU).

The contrastive loss for β alignment is formalized as:

$$\begin{aligned}\mathcal{L}^{(1 \rightarrow 2)} &= -\frac{1}{K} \sum_{k=1}^K \log \frac{\exp(g(y_k^{(1)}, y_k^{(2)}))}{\sum_{k'=1}^K \exp(g(y_k^{(1)}, y_{k'}^{(2)}))} \\ \mathcal{L}^{(2 \rightarrow 1)} &= -\frac{1}{K} \sum_{k=1}^K \log \frac{\exp(g(y_k^{(2)}, y_k^{(1)}))}{\sum_{k'=1}^K \exp(g(y_k^{(2)}, y_{k'}^{(1)}))} \\ \mathcal{L}_\beta &= \frac{1}{2} (\mathcal{L}^{(1 \rightarrow 2)} + \mathcal{L}^{(2 \rightarrow 1)})\end{aligned}$$

where $y_k^{(\ell)}$ is the projected semantic vector for topic k in language ℓ , and $g(\cdot, \cdot)$ is a similarity function in the projected space. This loss directly compels the projection functions and the underlying topic representations to align semantically corresponding topics across languages in the shared D_{sem} -dimensional space, facilitating interpretable cross-lingual topics.

3.4 Overall Objective

Inspired by prior work (Nan et al., 2019; Joo et al., 2020) suggesting prior modification for better semantic structure capture, we adopt a concise cluster-based Gaussian prior for latent variable z (from which θ is derived). For T clusters ($K = T$), document counts n_k per cluster, and concentration parameters $\alpha_k = n_k + \epsilon$, the prior $p(z)$, following Srivastava and Sutton (2017), is defined with mean and variance:

$$\mu_{\text{prior},k} = \log \alpha_k - \frac{1}{T} \sum_{j=1}^T \log \alpha_j$$

$$\sigma_{\text{prior},k}^2 = \frac{1}{\alpha_k} \left(1 - \frac{2}{T}\right) + \frac{1}{T^2} \sum_{j=1}^T \frac{1}{\alpha_j}$$

This $p(z)$ replaces the standard unit Gaussian prior in the KL divergence term of \mathcal{L}_{TM} .

The overall objective of XTRA is:

$$\mathcal{L} = \mathcal{L}_{\text{TM}} + \lambda_1 \mathcal{L}_{\text{InfoNCE}} + \lambda_2 \mathcal{L}_{\text{Cluster}} + \lambda_3 \mathcal{L}_\beta$$

where $\lambda_1, \lambda_2, \lambda_3$ balance terms. Minimizing \mathcal{L} ensures interpretable topics and cross-lingual alignment, leveraging the cluster’s ability to effectively reflect corpus content. Further details of the algorithm can be found in Algorithm 1.

4 Experiments and Results

Datasets

Our experimental evaluations utilized three benchmark datasets: **EC News** (Wu et al., 2020), a collection of English and Chinese news articles spanning six categories (business, education, entertainment, sports, technology, fashion); **Amazon Review** (Yuan et al., 2018), comprising English and Chinese Amazon reviews adapted for a binary classification task where five-star ratings are labeled “1” and others “0”; and **Rakuten Amazon**, consisting of Japanese reviews from Rakuten and English reviews from Amazon (Yuan et al., 2018), similarly formulated as a binary task based on ratings.

Baseline Models

We evaluated our proposed method against several established baselines: **MCTA** (Shi et al., 2016), a probabilistic framework for cross-lingual topic modeling (CLTM) designed to identify cultural variations; **MTAnchor** (Yuan et al., 2018), which utilizes multilingual anchor words to establish cross-language connections; **NMTM** (Wu et al., 2020), a neural approach aligning multilingual topic representations within a common vocabulary space; and **InfoCTM** (Wu et al., 2023a), which employs mutual information maximization, often via contrastive objectives, to enhance cross-lingual topic representation alignment and address topic repetition; as well as two clustering-based refinement baselines, **u-SVD** and **SVD-LR** (Chang et al., 2024).

Evaluation Metrics

To comprehensively assess generated topic quality and utility, we employed a multi-faceted strategy. For intrinsic quality, we measured cross-lingual

Model	EC News			Amazon Review			Rakuten Amazon		
	CNPMI	TU	TQ	CNPMI	TU	TQ	CNPMI	TU	TQ
MCTA [†]	0.025	0.489	0.012	0.028	0.319	0.009	0.021	0.272	0.006
MTAnchor [†]	-0.013	0.192	0.000	0.028	0.323	0.009	-0.001	0.214	0.000
NMTM [†]	0.031	0.784	0.024	0.042	0.732	0.031	0.009	0.679	0.006
InfoCTM [†]	0.048	0.913	0.044	0.043	0.923	0.040	0.034	0.870	0.030
u-SVD	0.083	0.830	0.069	0.054	0.638	0.034	0.025	0.584	0.015
SVD-LR	0.081	0.827	0.067	0.053	0.631	0.033	0.026	0.567	0.015
XTRA	0.076	0.993	0.075	0.055	0.980	0.054	0.035	0.975	0.034

Table 2: Comparison of CNPMI, TU, and TQ across datasets, where $TQ = \max(0, CNPMI) \times TU$; the best value in each column is in **bold**, and [†] denotes results reported in (Wu et al., 2023a).

topic coherence using **CNPMI** (Cross-lingual Normalized Pointwise Mutual Information (Hao et al., 2018)), an NPMI (Chang et al., 2009) extension for alignment assessment, and topic diversity using **TU** (Topic Uniqueness (Nan et al., 2019)) for redundancy evaluation; both used the top 15 words per topic. Following (Chang et al., 2024), we report **TQ (Topic Quality)**, which combines cross-lingual coherence with uniqueness while clipping negative coherence to zero. To evaluate practical utility and feature transferability, document-topic distributions were used as features for **SVM-based classification** in intra-lingual (-I) and cross-lingual (-C) settings, following common practice (e.g., Yuan et al. (2018); Wu et al. (2023a)). Additionally, inspired by Stambach et al. (2023), **LLM-based ratings** (1-3 scale) were utilized to assess intra-lingual coherence and cross-lingual topic alignment.

4.1 Topic quality analysis

We evaluate cross-lingual topic quality via coherence (CNPMI), diversity (Topic Uniqueness, TU), and a composite Topic Quality (TQ). For comparison, we include both older baselines (numbers from Wu et al. (2023a)) and more modern baselines, u-SVD and SVD-LR, which we re-tune under the same embedding setup for fairness (Table 2). On EC News, XTRA records a CNPMI of 0.076 versus 0.083 for u-SVD, yet it leads on TU 0.993 against 0.830 and on TQ 0.075 against 0.069. On Amazon Review, CNPMI reaches 0.055 for XTRA, with u-SVD second at 0.054 and SVD-LR third at 0.053; TU and TQ also favor XTRA at 0.980 and 0.054 versus the strongest alternative at 0.923 and 0.040. On Rakuten Amazon, CNPMI is 0.035 for XTRA, followed by InfoCTM at 0.034 and SVD-LR at 0.026; TU and TQ again lead at 0.975 and 0.034 over the next best at 0.870 and 0.030. Taken together, XTRA consistently outperforms the base-

lines across datasets and metrics; even when clustering models occasionally yield higher CNPMI, they lack an explicit topic space and cannot infer topic distributions for unseen documents.

4.2 Classification Performance Within and Across Languages

To evaluate practical utility and cross-lingual transferability, XTRA’s document–topic distributions θ are used as features for downstream classification with Support Vector Machines (SVMs), following Wu et al. (2023a). Performance is assessed in both intralingual (-I) and crosslingual (-C) settings. Methods from Chang et al. (2024) (u-SVD and SVD-LR) do not produce document–topic posteriors and therefore cannot be applied to these downstream tasks, so they are omitted here. As shown in Figure 4, XTRA consistently and significantly outperforms baseline models across all datasets and evaluation scenarios. For instance, on the EC News dataset, XTRA demonstrates markedly superior accuracies in both intralingual (EN-I, ZH-I) and crosslingual tasks (EN-C, ZH-C) when compared to prominent baselines. This trend of clear outperformance is consistently replicated across the Amazon Review and Rakuten Amazon datasets, where XTRA again exhibits visibly higher classification accuracies in all tested intralingual and crosslingual settings depicted. This robust and generalizable classification advantage directly stems from the superior θ representations learned by XTRA, which are effectively shaped by its integrated loss functions (L_{InfoNCE} , L_{Cluster} , L_{β}) to ensure robust semantic alignment and high discriminative capability for the SVMs.

4.3 Ablation Study

An ablation study on ECNews with 50 topics (Table 3) confirmed the distinct contributions of XTRA’s core loss functions; the full model deliv-

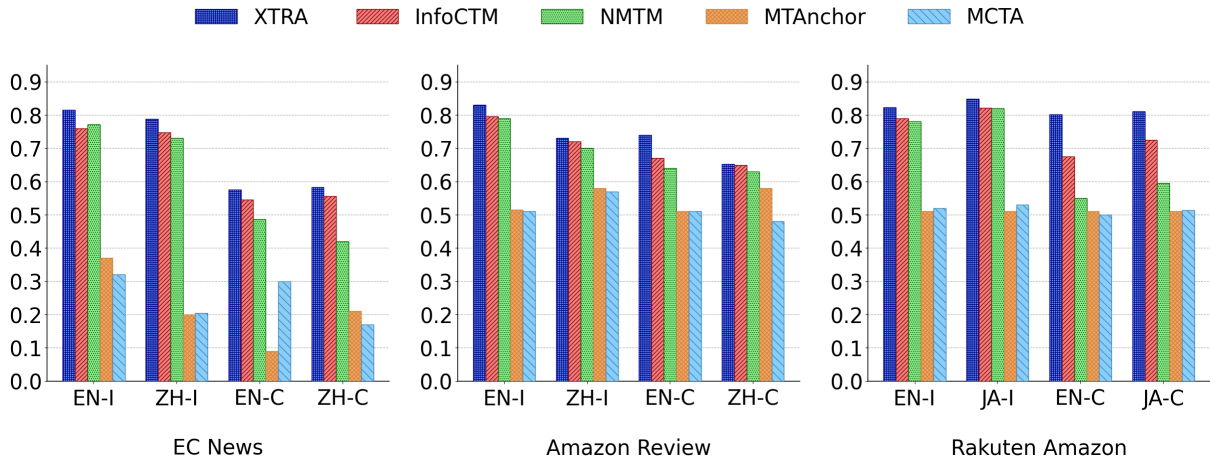


Figure 4: Overall caption for the three figures showing classification results on different datasets.

Model	Topic Quality		Classification			
	CNPMI	TU	EN-I	ZH-I	EN-C	ZH-C
NMTM [†]	0.031	0.784	0.771	0.731	0.487	0.420
InfoCTM [†]	0.048	0.913	0.760	0.747	0.545	0.556
w/o L_β	0.064	0.978	0.803	0.791	0.550	0.562
w/o L_{cluster}	0.054	0.991	0.814	0.789	0.577	0.580
w/o L_{InfoNCE}	0.051	0.972	0.785	0.765	0.372	0.453
XTRA	0.076	0.993	0.815	0.788	0.575	0.582

Table 3: Ablation study results on the ECNews dataset (50 topics). Performance is measured by Topic Quality (CNPMI, TU) and Classification Accuracy. The best results for XTRA and its ablated versions are in bold. [†]Results reported in (Wu et al., 2023a).

ered superior topic quality, with CNPMI 0.076 and TU 0.993. While TU remained high (above 0.97) across configurations, optimal CNPMI, crucial for cross-lingual coherence, relied on all components. Removing L_{InfoNCE} reduced CNPMI to 0.051, impacting the intended enhancement of document-topic (θ) alignment via PLM embeddings. Omitting L_{cluster} yielded a CNPMI of 0.054, indicating less effective direct shaping of the θ space for cross-lingual consistency. The absence of L_β resulted in a CNPMI of 0.064, suggesting a weakened semantic correspondence for topic-word (β) distributions. These ablated CNPMI scores, though lower than the full model, generally surpassed baselines like NMTM (0.031) and InfoCTM (0.048). For classification, XTRA was competitive, with an EN-I score of 0.815 and a ZH-C score of 0.582. The essential role of L_{InfoNCE} in fostering transferable representations was clear, as its removal significantly dropped EN-C performance from 0.575 to 0.372. While configurations lacking L_β (ZH-I: 0.791) or L_{cluster} (EN-C: 0.577) showed specific

sub-task strengths, their primary design for topic and document-topic alignment for coherence was validated as key. These findings underscore the complementary roles of each loss in XTRA’s overall performance.

4.4 Qualitative Analysis: Discovered Topic Word Examples

To further assess topic quality, Table 4 presents the top keywords on EC News produced by the autoencoding topic models NMTM, InfoCTM, and XTRA, with noisy or misaligned terms highlighted in red. XTRA consistently generates more coherent and better-aligned cross-lingual topics than the baselines.

For the Social Media and Internet topic, NMTM and InfoCTM include noisy terms (such as NMTM’s “sandy” and “allegedly”, or InfoCTM’s “ram”) and exhibit weak cross-lingual alignment. In stark contrast, XTRA produces a highly consistent topic. It effectively captures core concepts through semantically aligned English keywords (including “followers”, “posts”, “tweets”) and their Chinese counterparts (namely “网民 (netizen)”, “微博 (weibo)”, “转发 (retweet)”), all without the noise seen in other models.

A similar advantage for XTRA is evident in the Finance topic. While NMTM again introduces outliers like “month” and “jackpot”, and InfoCTM includes noisy or off-topic words, such as “访问 (visit)”, “明日 (tomorrow)”, “无尽 (endless)”, XTRA delivers a semantically focused and cross-lingually coherent topic. It covers essential financial concepts, featuring relevant English keywords (such as “banks”, “financial”) and coherent Chinese terms (such as “银行 (bank)”, “存款 (deposit)”,

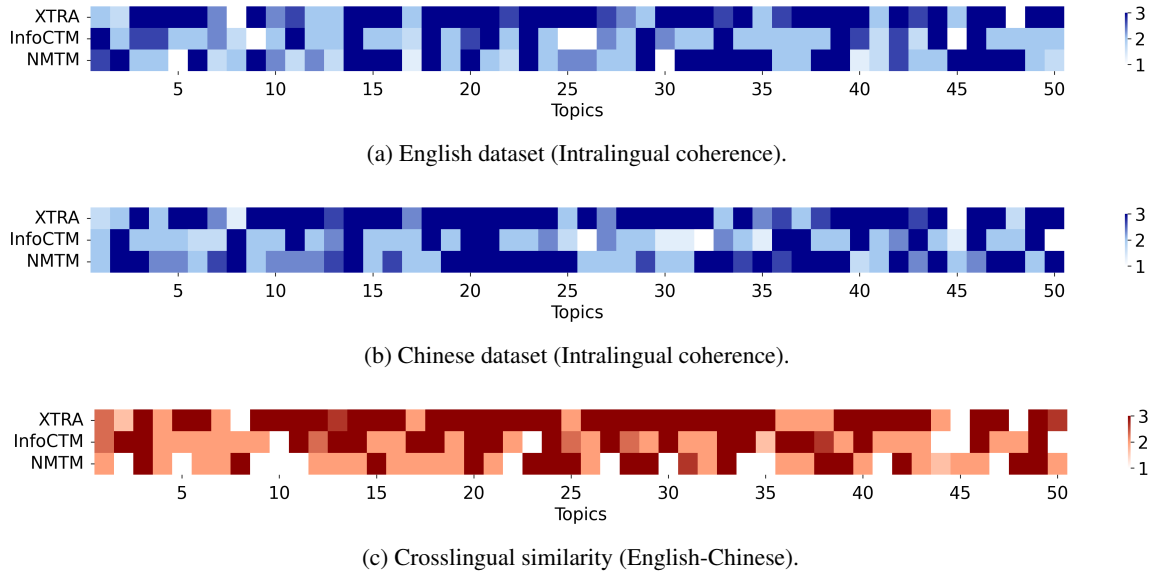


Figure 5: LLM-based topic quality evaluations on the Amazon Review dataset. Darker shades indicate higher scores. The final reported score for each evaluation is a real number, representing the average of three independent LLM assessments.

Topic: Social Media and Internet					
NMTM Topic 45					
EN Topic:	site	online	internet	sandy	allegedly
ZH Topic:	游戏	网民	软件	网页	网游
Translations:	game	netizen	software	webpage	online game
InfoCTM Topic 32					
EN Topic:	offer	data	web	ram	remote
ZH Topic:	网址	管理	销售	客户	互联网
Translations:	website	manage	sales	customer	internet
XTRA Topic 32					
EN Topic:	followers	posts	posted	tweets	social
ZH Topic:	网民	微博	转发	网	网站
Topic Translations:	netizen	weibo	retweet	network	website
Topic: Finance					
NMTM Topic 8					
EN Topic:	stock	fund	market	month	jackpot
ZH Topic:	基金	股型	上证	华夏	净值
Translations:	fund	stock type	shanghai idx	huaxia	net value
InfoCTM Topic 15					
EN Topic:	stock	bank	share	report	due
ZH Topic:	投资者	访问	明日	无尽	股市
Translations:	investor	visit	tomorrow	endless	stock market
XTRA Topic 1					
EN Topic:	lenders	banks	financial	banking	debt
ZH Topic:	信贷	利率	存款	银行	贷款
Translations:	credit	interest rate	deposit	bank	loan

Table 4: Top-5 topic words from EC News for two selected topics ("Social Media and Internet" and "Finance") learned by NMTM, InfoCTM, and XTRA. Red words indicate noisy or disaligned terms with the corresponding cross-lingual topic.

"贷款 (loan)", with no apparent noise, unlike the baselines which struggle with coherence and alignment.

4.5 LLM-based Topic Quality Evaluation

Inspired by recent work using large language models for automated topic model assessment (Stammbach et al., 2023), we incorporated LLM-based evaluations on three VAE-based topic models: NMTM, InfoCTM, and XTRA. We used LLMs for two tasks: assessing intra-lingual coherence (relatedness, 1–3 scale) and evaluating cross-lingual semantic similarity (similarity, 1–3 scale). System prompts are in Appendix D.

A comprehensive overview of LLM ratings for Amazon Review is presented in Figures 5a (English Intralingual), 5b (Chinese Intralingual), and 5c (Crosslingual Similarity), where darker shades indicate higher scores. These visualizations collectively indicate XTRA’s strong performance. In both intralingual coherence evaluations (Figures 5a, 5b), XTRA consistently achieves a high concentration of top scores, performing competitively with or surpassing baselines including InfoCTM (Wu et al., 2023a) and NMTM (Wu et al., 2020). While models like NMTM can exhibit good top word coherence, this may associate with reduced topic diversity (Wu et al., 2023a). XTRA, in contrast, demonstrates strong intralingual coherence while also maintaining high topic diversity, suggesting a more balanced, robust topic generation. XTRA’s benefits become particularly evident in crosslingual similarity assessment (Figure 5c). Here, XTRA maintains strong performance with predominantly

high similarity scores, while other baselines like InfoCTM and NMTM show more varied results. This LLM-based assessment on Amazon Review suggests XTRA not only produces highly coherent topics within each language but also excels at establishing strong semantic alignment across languages, positioning it favorably against current methods.

5 Conclusion

We propose XTRA, a cross-lingual topic modeling framework integrating Bag-of-Words with multilingual embeddings via a dual-alignment mechanism. XTRA aligns document-topic distributions through contrastive learning and projects topic-word distributions into a shared semantic space, enhancing cross-lingual consistency beyond lexical matching. Experiments show XTRA outperforms baselines in topic coherence, diversity, and alignment, demonstrating its efficacy for reliable and interpretable cross-lingual theme discovery.

Limitation

XTRA’s approach to cross-lingual topic modeling faces certain constraints. It depends on predefined numbers of topics and clusters, which limits its adaptability for datasets with ambiguous or changing themes and may result in less effective outcomes. Moreover, since it relies on pre-trained multilingual embeddings and offline clustering, its usefulness in real-time or dynamic environments is restricted, highlighting the need for further research to enhance its flexibility across various multilingual scenarios.

Ethical Considerations

We adhere to the ACL Code of Ethics and the terms of each codebase license. Our method aims to advance the field of topic modeling, and we are confident that, when used properly and with care, it poses no significant social risks.

Acknowledgements

Trung Le was partly supported by the Air Force Office of Scientific Research under award number FA2386-23-1-4044.

References

Dimo Angelov. 2020. [Top2vec: Distributed representations of topics](#). *CoRR*, abs/2008.09470.

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, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 759–766. 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, EACL 2021, Online, April 19 - 23, 2021*, pages 1676–1683. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, pages 993–1022.

Jordan L. Boyd-Graber and David M. Blei. 2012. [Multilingual topic models for unaligned text](#). *CoRR*, abs/1205.2657.

Chia-Hsuan Chang, Tien-Yuan Huang, Yi-Hang Tsai, Chia-Ming Chang, and San-Yih Hwang. 2024. [Refining dimensions for improving clustering-based cross-lingual topic models](#). *CoRR*, abs/2412.12433.

Chia-Hsuan Chang and San-Yih Hwang. 2021. [A word embedding-based approach to cross-lingual topic modeling](#). *Knowl. Inf. Syst.*, 63(6):1529–1555.

Jonathan D. Chang, Jordan L. Boyd-Graber, Sean Gerish, Chong Wang, and David M. Blei. 2009. [Reading tea leaves: How humans interpret topic models](#). In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pages 288–296. Curran Associates, Inc.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [BGE m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). *CoRR*, abs/2402.03216.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. [Infoxlm: An information-theoretic framework for cross-lingual language model pre-training](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3576–3588. Association for Computational Linguistics.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. [Indexing by latent semantic analysis](#). *J. Am. Soc. Inf. Sci.*, 41(6):391–407.

- Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, pages 439–453.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. **Simcse: Simple contrastive learning of sentence embeddings**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.
- Maarten Grootendorst. 2022. **Bertopic: Neural topic modeling with a class-based tf-idf procedure**. Preprint, arXiv:2203.05794.
- Cuong Ha, Van-Dang Tran, Ngo Van Linh, and Khoat Than. 2019. **Eliminating overfitting of probabilistic topic models on short and noisy text: The role of dropout**. *Int. J. Approx. Reason.*, 112:85–104.
- Shudong Hao, Jordan L. Boyd-Graber, and Michael J. Paul. 2018. **Lessons from the bible on modern topics: Low-resource multilingual topic model evaluation**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1090–1100. Association for Computational Linguistics.
- Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.
- Alexander Miserlis Hoyle, Pranav Goel, and Philip Resnik. 2020. **Improving neural topic models using knowledge distillation**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1752–1771. Association for Computational Linguistics.
- Jagadeesh Jagarlamudi and Hal Daumé III. 2010. **Extracting multilingual topics from unaligned comparable corpora**. In *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings*, volume 5993 of *Lecture Notes in Computer Science*, pages 444–456. Springer.
- Weonyoung Joo, Wonsung Lee, Sungrae Park, and Il-Chul Moon. 2020. **Dirichlet variational autoencoder**. *Pattern Recognit.*, 107:107514.
- Diederik P. Kingma and Jimmy Ba. 2015. **Adam: A method for stochastic optimization**. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations*.
- Ngo Van Linh, Tran Xuan Bach, and Khoat Than. 2022. **A graph convolutional topic model for short and noisy text streams**. *Neurocomputing*, 468:345–359.
- James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 5, pages 281–298. University of California press.
- Khai Mai, Sang Mai, Anh Nguyen, Ngo Van Linh, and Khoat Than. 2016. **Enabling hierarchical dirichlet processes to work better for short texts at large scale**. In *Advances in Knowledge Discovery and Data Mining - 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part II*, volume 9652 of *Lecture Notes in Computer Science*, pages 431–442. Springer.
- David M. Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. **Polylingual topic models**. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 880–889. ACL.
- Aaron Mueller and Mark Dredze. 2021. **Fine-tuning encoders for improved monolingual and zero-shot polylingual neural topic modeling**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3054–3068. Association for Computational Linguistics.
- Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. **Topic modeling with wasserstein autoencoders**. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6345–6381. Association for Computational Linguistics.
- Duc Anh Nguyen, Ngo Van Linh, Nguyen Kim Anh, and Khoat Than. 2017. **Keeping priors in streaming bayesian learning**. In *Advances in Knowledge Discovery and Data Mining - 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part II*, volume 10235 of *Lecture Notes in Computer Science*, pages 247–258.
- Ha Nguyen, Hoang Pham, Son Nguyen, Ngo Van Linh, and Khoat Than. 2022a. **Adaptive infinite dropout for noisy and sparse data streams**. *Mach. Learn.*, 111(8):3025–3060.
- Quang Duc Nguyen, Tung Nguyen, Duc Anh Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. 2025a. **Glocom: A short text neural topic model via global clustering context**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume*

- 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 1109–1124. Association for Computational Linguistics.
- Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. In *Advances in Neural Information Processing Systems*, pages 11974–11986.
- Thong Thanh Nguyen, Xiaobao Wu, Xinshuai Dong, Cong-Duy T. Nguyen, See-Kiong Ng, and Anh Tuan Luu. 2024. [Topic modeling as multi-objective contrastive optimization](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Tung Nguyen, Tue Le, Hoang Tran Vuong, Quang Duc Nguyen, Duc Anh Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. 2025b. [Sharpness-aware minimization for topic models with high-quality document representations](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 4507–4524. Association for Computational Linguistics.
- Tung Nguyen, Trung Mai, Nam Nguyen, Linh Ngo Van, and Khoat Than. 2022b. [Balancing stability and plasticity when learning topic models from short and noisy text streams](#). *Neurocomputing*, 505:30–43.
- Tung Nguyen, Linh Van Ngo, Duc Anh Nguyen, and Sang Dinh Viet. 2025c. [A framework for neural topic modeling with mutual information and group regularization](#). *Neurocomputing*, 645:130420.
- Tung Nguyen, Duy-Tung Pham, Quang Duc Nguyen, Linh Ngo Van, Anh Nguyen Duc, and Sang Dinh Viet. 2025d. [Topicot: Neural topic model aligning with pre-trained clustering and optimal transport](#). *Neurocomputing*, page 131268.
- Tung Nguyen, Tung Pham, Ngo Van Linh, Ha-Bang Ban, and Khoat Than. 2025e. [Out-of-vocabulary handling and topic quality control strategies in streaming topic models](#). *Neurocomputing*, 614:128757.
- Van-Son Nguyen, Duc-Tung Nguyen, Ngo Van Linh, and Khoat Than. 2019. [Infinite dropout for training bayesian models from data streams](#). In *2019 IEEE International Conference on Big Data (IEEE BigData), Los Angeles, CA, USA, December 9-12, 2019*, pages 125–134. IEEE.
- Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. [Mining multilingual topics from wikipedia](#). In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 1155–1156. ACM.
- Duy-Tung Pham, Thien Trang Nguyen Vu, Tung Nguyen, Linh Ngo Van, Duc Anh Nguyen, and Thien Huu Nguyen. 2024. [Neuromax: Enhancing neural topic modeling via maximizing mutual information and group topic regularization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*.
- Bei Shi, Wai Lam, Lidong Bing, and Yinqing Xu. 2016. [Detecting common discussion topics across culture from news reader comments](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.
- Akash Srivastava and Charles Sutton. 2017. [Autoencoding variational inference for topic models](#). In *International Conference on Learning Representations*.
- Dominik Stambach, Vilém Zouhar, Alexander Miserlis Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. [Revisiting automated topic model evaluation with large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9348–9357. Association for Computational Linguistics.
- Anh Phan Tuan, Tran Xuan Bach, Thien Huu Nguyen, Ngo Van Linh, and Khoat Than. 2020. [Bag of biterns modeling for short texts](#). *Knowl. Inf. Syst.*, 62(10):4055–4090.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. [Representation learning with contrastive predictive coding](#). *Preprint*, arXiv:1807.03748.
- Tu Vu, Manh Do, Tung Nguyen, Ngo Van Linh, Sang Dinh, and Thien Huu Nguyen. 2025. [Topic modeling for short texts via optimal transport-based clustering](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 7666–7680. Association for Computational Linguistics.
- Hoang Tran Vuong, Tue Le, Tu Vu, Tung Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. 2025. [Hicot: Improving neural topic models via optimal transport and contrastive learning](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 13894–13920. Association for Computational Linguistics.
- Xiaobao Wu, Xinshuai Dong, Thong Nguyen, Chaoqun Liu, Liangming Pan, and Anh Tuan Luu. 2023a. [Infotcm: A mutual information maximization perspective of cross-lingual topic modeling](#). In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13763–13771. AAAI Press.

- Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, and Anh Tuan Luu. 2023b. [Effective neural topic modeling with embedding clustering regularization](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 37335–37357. PMLR.
- Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. [Learning multilingual topics with neural variational inference](#). In *Natural Language Processing and Chinese Computing - 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14-18, 2020, Proceedings, Part I*, volume 12430 of *Lecture Notes in Computer Science*, pages 840–851. Springer.
- Weiwei Yang, Jordan L. Boyd-Graber, and Philip Resnik. 2019. [A multilingual topic model for learning weighted topic links across corpora with low comparability](#). 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 2019, Hong Kong, China, November 3-7, 2019*, pages 1243–1248. Association for Computational Linguistics.
- Michelle Yuan, Benjamin Van Durme, and Jordan L. Ying. 2018. [Multilingual anchoring: Interactive topic modeling and alignment across languages](#). In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8667–8677.
- He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray L. Buntine. 2021. [Neural topic model via optimal transport](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

A Related Work

Topic Models and Cross-lingual topic model

Topic modeling is a core technique for discovering latent semantic structures in text corpora. The classical Latent Dirichlet Allocation (LDA) (Blei et al., 2003) models documents as mixtures of latent topics. Recent research has significantly advanced this field, integrating language representation and deep learning, leading to Neural Topic Models (NTMs). Notable NTMs include VAE-based models like Neural Variational Document Model (NVDM) and ProdLDA (Srivastava and Sutton, 2017). Advancements involve representing topics as embeddings (e.g., ETM using pre-trained embeddings (Dieng et al., 2020), or optimal transport methods (Zhao et al., 2021; Wu et al., 2023b; Vu et al., 2025; Vuong et al., 2025; Nguyen et al., 2025d)). Combining NTMs with pre-trained language models like BERT enhances contextual understanding (Bianchi et al., 2021a,b; Hoyle et al., 2020; Nguyen et al., 2025a,b). Contrastive learning and related regularization frameworks improve NTM training and topic–document distribution refinement (Nguyen and Luu, 2021; Nguyen et al., 2024; Vuong et al., 2025; Nguyen et al., 2025c), while clustering of word or document embeddings from models like Doc2Vec or BERT provides alternative topic discovery paradigms (Angelov, 2020; Grootendorst, 2022; Nguyen et al., 2025a). In addition, a substantial line of work focuses on short and noisy text, including graph convolutional models for text streams (Linh et al., 2022), dropout-regularized probabilistic topic models (Ha et al., 2019), adaptive infinite dropout for noisy and sparse data streams (Nguyen et al., 2022a), infinite dropout for streaming Bayesian models (Nguyen et al., 2019), bag-of-biterms models for short texts (Tuan et al., 2020), global clustering context methods (Nguyen et al., 2025a), continual learning strategies for balancing stability and plasticity (Nguyen et al., 2022b), and streaming-specific mechanisms such as out-of-vocabulary handling and topic quality control (Nguyen et al., 2025e), as well as Bayesian streaming frameworks that preserve prior information (Nguyen et al., 2017) and hierarchical extensions for short texts (Mai et al., 2016).

Cross-lingual topic modeling (CLTM) extends topic modeling to multilingual settings to discover aligned themes across languages. Early methods like (Mimno et al., 2009) relied on parallel corpora, limiting applicability. Later approaches uti-

lized bilingual dictionaries for vocabulary alignment (Jagarlamudi and III, 2010; Boyd-Graber and Blei, 2012), with dictionary-based translation improvements by Shi et al. (2016), Yuan et al. (2018), Yang et al. (2019), Wu et al. (2020) and Wu et al. (2023a). An alternative line uses multilingual word embeddings (Chang and Hwang, 2021), facing issues like isomorphism assumptions. Transformer-based methods (Bianchi et al., 2021b; Mueller and Dredze, 2021) enable zero-shot inference but still struggle with cross-lingual alignment.

Mutual Information Maximization Mutual Information Maximization (MIM) is a principle in unsupervised/self-supervised learning, often approximated by InfoNCE (van den Oord et al., 2019). It has been applied to sentence embedding (Gao et al., 2021) and multilingual representation alignment (Chi et al., 2021). In neural topic modeling, contrastive learning based on InfoNCE has been used for discriminative topic distributions (Nguyen and Luu, 2021), document-topic alignment (Pham et al., 2024), and extended to the cross-lingual setting for aligned topics in InfoCTM (Wu et al., 2023a). These works highlight MIM’s effectiveness in capturing structured information.

B Algorithm

In this section, we present the **XTRA** training procedure:

Algorithm 1 XTRA training procedure

Input: Input corpus $\mathbf{X} = \mathbf{X}^{(1)} \cup \mathbf{X}^{(2)}$, K , N , C , $\lambda_{1,2,3}$.

Output: Optimized parameters $\Theta^* = \{\text{Encoders}^*, \beta^{(1)*}, \beta^{(2)*}, \text{Projectors}^*\}$

- 1: Initialize parameters $\Theta = \{\text{Encoders}, \beta^{(1)}, \beta^{(2)}, \text{Projectors}\}$ and Optimizer.
- 2: **for** epoch from 1 to N **do do**
- 3: Shuffle $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$.
- 4: **for** each balanced mini-batch b sampled from \mathbf{X} **do do**
- 5: Compute document-level components: $\{\theta_d\}$, L_{TM}^b , $L_{InfoNCE}^b$ for all $x_d \in b$;
- 6: Compute batch-level cluster loss $L_{Cluster}^b$ using $\{\theta_d\}$ and C ;
- 7: Compute batch-level beta alignment loss L_{β}^b using $\beta^{(1)}, \beta^{(2)}, P^{1,2}$;
- 8: $\mathcal{L}_{batch} \leftarrow L_{TM}^b + \lambda_1 L_{InfoNCE}^b + \lambda_2 L_{Cluster}^b + \lambda_3 L_{\beta}^b$;
- 9: Update Θ with $\nabla \mathcal{L}_{batch}$;
- 10: **end for**
- 11: **end for**

C Clustering Method

We use an asymmetric clustering approach to ensure cross-lingual alignment. After reducing dimensionality via Singular Value Decomposition (SVD) (Deerwester et al., 1990) and applying L2 normalization, we perform KMeans clustering (MacQueen, 1967) on the pivot language (e.g., English). Documents from the other language are then assigned to the nearest clusters based on cosine similarity. This avoids the issue of monolingual clusters that can arise when clustering both languages jointly, leading to better cross-lingual consistency.

D Detailed Prompts for LLM Evaluation

This appendix provides the detailed system prompts used for the LLM-based evaluation tasks described in the main text. Tables 5 and 6 shows the prompts side-by-side for intralingual coherence and crosslingual similarity assessment across the different datasets.

Dataset	Prompt
EC News	You are a helpful assistant evaluating the top words of a topic model output for a given topic. The dataset is EC News, a collection of English and Chinese news with 6 categories: business, education, entertainment, sports, tech, and fashion. Please rate how related the following words are to each other on a scale from 1 to 3 ("1"=not very related, "2"=moderately related, "3"=very related). Reply with a single number, indicating the overall appropriateness of the topic.
Amazon Review	You are a helpful assistant evaluating the top words of a topic model output for a given topic. The dataset is Amazon Review, which includes English and Chinese reviews from the Amazon website. Please rate how related the following words are to each other on a scale from 1 to 3 ("1"=not very related, "2"=moderately related, "3"=very related). Reply with a single number, indicating the overall appropriateness of the topic.
Rakuten Amazon	You are a helpful assistant evaluating the top words of a topic model output for a given topic. The dataset is Rakuten Amazon, which contains Japanese reviews from Rakuten, and English reviews from Amazon. Please rate how related the following words are to each other on a scale from 1 to 3 ("1"=not very related, "2"=moderately related, "3"=very related). Reply with a single number, indicating the overall appropriateness of the topic.

Table 5: Intralingual Coherence Prompts for LLM-based Evaluation

Dataset	Prompt
EC News	You are a helpful assistant evaluating the similarity of topics derived from topic modeling on parallel news corpora. The dataset is EC News, with English and Chinese news. You will be given two sets of top words, one for an English topic (Language 1) and one for a Chinese topic (Language 2). Please rate how similar the underlying topics represented by these two sets of words are, on a scale from 1 to 3 ("1"=not very similar, "2"=moderately similar, "3"=very similar). Reply with a single number.
Amazon Review	You are a helpful assistant evaluating the similarity of topics derived from topic modeling on parallel review corpora. The dataset is Amazon Review, with English and Chinese reviews. You will be given two sets of top words, one for an English topic (Language 1) and one for a Chinese topic (Language 2). Please rate how similar the underlying topics represented by these two sets of words are, on a scale from 1 to 3 ("1"=not very similar, "2"=moderately similar, "3"=very similar). Reply with a single number.
Rakuten Amazon	You are a helpful assistant evaluating the similarity of topics derived from topic modeling on parallel review corpora. The dataset is Rakuten Amazon, with Japanese reviews (Rakuten - Language 2) and English reviews (Amazon - Language 1). You will be given two sets of top words, one for an English topic and one for a Japanese topic. Please rate how similar the underlying topics represented by these two sets of words are, on a scale from 1 to 3 ("1"=not very similar, "2"=moderately similar, "3"=very similar). Reply with a single number.

Table 6: Crosslingual Similarity Prompts for LLM-based Evaluation

E Implementation Details

Our models were trained on a single NVIDIA P100 GPU (Kaggle). We employed the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 0.002 and trained for 800 epochs. A learning rate scheduler decayed the learning rate by a factor of 0.5 every 250 epochs. Hyperparameters were tuned over:

- λ_1 for $\mathcal{L}_{\text{InfoNCE}}$: {70, 80, 85}
- λ_2 for $\mathcal{L}_{\text{Cluster}}$: {5, 10}
- λ_3 for \mathcal{L}_β : {7, 15}

We report results with the best hyperparameters. On Amazon Reviews, XTRA trained in about 30 minutes on a single P100; InfoCTM took about 35 minutes; NMTM finished in about 7 minutes and 35 seconds. Rakuten Amazon showed comparable wall-clock times under the same hardware and settings. On ECNews, training ran longer: around 60 minutes for XTRA, 70 for InfoCTM, and 14 for NMTM.