SentenceLDA: Discriminative and Robust Document Representation with Sentence Level Topic Model

Taehun Cha and **Donghun Lee***
Department of Mathematics
Korea University
{cth127, holy}@korea.ac.kr

Abstract

A subtle difference in context results in totally different nuances even for lexically identical words. On the other hand, two words can convey similar meanings given a homogeneous context. As a result, considering only word spelling information is not sufficient to obtain quality text representation. We propose SentenceLDA, a sentence-level topic model. We combine modern SentenceBERT and classical LDA to extend the semantic unit from word to sentence. By extending the semantic unit, we verify that SentenceLDA returns more discriminative document representation than other topic models, while maintaining LDA's elegant probabilistic interpretability. We also verify the robustness of SentenceLDA by comparing the inference results on original and paraphrased texts. Additionally, we implement one possible application of SentenceLDA on corpus-level key opinion mining by applying SentenceLDA on an argumentative corpus, DebateSum.[1]

1 Introduction

What a word conveys can vary significantly in different contexts. It is not just a matter of polysemy. The word 'gas' in the energy-related context conveys different sentiments and implications compared to that in the environment-related context, though lexically identical. As a result, lexical property cannot fully explain the difference between different topics. To discriminate the differences, extending the semantic unit from word to sentence or paragraph seems natural and necessary.

On the other hand, different words can convey similar meanings given similar contexts. Synonym is a particular example. Identifying similar words helps NLP models be robust to semantic information, not being biased to word spelling. To achieve this property, extending the semantic unit seems also natural and necessary.

SentenceLDA
1. The frequency of wars between states has diminished since **World** War II (...).
2. The ability of countries to cooperate and resolve their differences through (...) direction taken by the **worlds** leading economy.
3. The **worlds** oceans have been shown to be less able to absorb and store carbon dioxide and other greenhouse gases (...).
LDA
1. states, war, **world**, power, china
2. fish, ocean, species, global, warming
3. would, one, even, years, **world**, said

Table 1: Difference between sentence-level topic (SentenceLDA) and word-level topic (LDA), extracted from DebateSum dataset. SentenceLDA captures various aspects of the word 'world', e.g. historical (1), political, economic (2), and environmental (3) aspects. While LDA only captures a political aspect (1) and separates it from an environment-related topic (2). LDA sometimes returns uninterpretable topics (3).

Multiple researchers suggested contextualized word representation by pre-training neural language models (PLM, Peters et al., 2018, Radford et al., 2018 and Devlin et al., 2019) to overcome these shortcomings. PLM considers the contextual information by extending the semantic unit to the whole input text, not a separate word. PLM showed great improvement in various NLP tasks for the last five years.

Meanwhile, topic models are useful tools for corpus-level analysis. Latent Dirichlet Allocation (LDA, Blei et al., 2003) is one of the most successful topic models which combines topic modeling and a probabilistic graphical model. LDA proposed an elegant generative process by utilizing

*corresponding author
[1]Source codes are available on https://github.com/cth127/sentencelda

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 521–538

Bayesian statistics to capture corpus-wide topics, as presented on Table 1. But LDA and most of its derivatives are still word-level and depend on exchangeability assumption between words. It makes LDA-like models less discriminative between similar topics that share a similar word distribution. It also makes them less robust to semantic information.

In this paper, we propose SentenceLDA, a sentence-level topic model. We successfully combine LDA and SentenceBERT (Reimers and Gurevych, 2019) while maintaining its probabilistic interpretation. Our main research hypothesis is "Does the semantic unit extension of a topic model from word to sentence return more discriminative and robust document representation?" We test the discriminativeness with a text classification task by comparing SentenceLDA with various topic models. We also test the robustness by comparing the representation of lexical and syntactic paraphrases. As a potential application, we apply SentenceLDA on the argumentative corpus to obtain corpus-level key opinions, as exemplified on Table 1. The results show SentenceLDA returns a more discriminative and robust document representation than other topic models.

2 Related Works

Various derivatives were proposed after Blei et al. (2003) proposed the LDA. One stream of the research was to exploit continuous word embedding by exchanging the multinomial word distribution in the LDA with Gaussian distribution (GaussianLDA, Das et al., 2015), utilizing linear classifier (Nguyen et al., 2015 and Dieng et al., 2020), or both (Li et al., 2016). After the PLMs were introduced, e.g. ELMO (Peters et al., 2018) and BERT (Devlin et al., 2019), researchers have explored the possibility of combining the PLMs with classic topic models. They used the contextual information as a direct input (Bianchi et al., 2021a and Bianchi et al., 2021b), knowledge distillation (Hoyle et al., 2020), or direct clustering (Thompson and Mimno, 2020, Sia et al., 2020 and Zhang et al., 2022). But this line of research focuses on improving the word-level topic model, not considering sentence-level information.

Another line of research explores the sentence-level topic models. One way is to assign a topic to each sentence and sample words from the sentence topic distribution, not document topic distribution, (Wang et al., 2009, Balikas et al., 2016, Li et al., 2017 and Jiang et al., 2019). But these models still assume the exchangeability between words, which makes a model less effective in discriminating sentences sharing the same words but having different word orders.

To bypass the exchangeability issue, some researchers applied a variational auto-encoder scheme based on RNN or LSTM decoder (Tian et al., 2016, Nallapati et al., 2017, Wang et al., 2019 and Rezaee and Ferraro, 2020). As an extension of this stream, researchers applied PLMs to obtain more useful sentence representation. But in the document-generating process, they usually depend on clustering (Kozbagarov et al., 2021 and Sastre Martinez et al., 2022) or similarity metric between sentence and topic embedding (Yang et al., 2015 and Schneider, 2023). It is difficult to probabilistically interpret the topic embeddings, which are merely the center of each cluster.

To the best of our knowledge, there is no sentence-level topic model, extending the LDA while utilizing contextual information and maintaining its probabilistic interpretation. Moreover, there is no work exploring the benefit of the sentence-level topic model from the view of discriminativeness and robustness.

3 SentenceLDA: Sentence-Level Topic Model

To capture a subtle nuance of a word, semantic unit extension from word to sentence is necessary. This extension would make a topic model more discriminative between documents sharing similar word distribution while containing different topics. On the other hand, this extension would make a topic model to be more robust to the semantics of a document, not being biased toward word spellings.

To achieve these goals, we present our sentence-level topic model SentenceLDA. Table 2 explains how SentenceLDA has evolved from LDA and GaussianLDA. It is truly a simple modification from GaussianLDA. Only the unit of process in 2-(b) is changed from 'word' to 'sentence'.

But thanks to the simple modification, SentenceLDA fully utilizes a Bayesian probabilistic framework. It is the key difference between SentenceLDA and other sentence-level topic models, which utilize clustering and similarity metrics. As a result, we can fully interpret the resulting topic distribution from a probabilistic perspective.

LDA	GaussianLDA	SentenceLDA
1. for $k = 1$ to K (a) Choose topic $\boldsymbol{\beta}_k \sim Dir(\boldsymbol{\eta})$ 2. for each document d in corpus D (a) Choose a topic distribution $\boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha})$ (b) for each word index n from 1 to N_d i. Choose a topic $z_{d,n} \sim Cat(\boldsymbol{\theta}_d)$ ii. Choose word $w_n \sim Cat(\boldsymbol{\beta}_{z_{d,n}})$	1. for $k = 1$ to K (a) Draw topic covariance $\boldsymbol{\Sigma}_k \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$ (b) Draw topic mean $\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{\kappa}\boldsymbol{\Sigma}_k)$ 2. for each document d in corpus D (a) Draw a topic distribution $\boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha})$ (b) for each word index n from 1 to N_d i. Draw a topic $z_{d,n} \sim Cat(\boldsymbol{\theta}_d)$ ii. Draw word $\boldsymbol{v}_{d,n} \sim \mathcal{N}(\boldsymbol{\mu}_{z_{d,n}}, \boldsymbol{\Sigma}_{z_{d,n}})$	1. for $k = 1$ to K (a) Draw topic covariance $\boldsymbol{\Sigma}_k \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$ (b) Draw topic mean $\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{\kappa}\boldsymbol{\Sigma}_k)$ 2. for each document d in corpus D (a) Draw a topic distribution $\boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha})$ (b) for each sentence index n from 1 to N_d i. Draw a topic $z_{d,n} \sim Cat(\boldsymbol{\theta}_d)$ ii. Draw sentence $\boldsymbol{v}_{d,n} \sim \mathcal{N}(\boldsymbol{\mu}_{z_{d,n}}, \boldsymbol{\Sigma}_{z_{d,n}})$

Table 2: Hypothetical data-generating process of each algorithm. K is the number of topics, N_d is the number of words in document d, Dir is a Dirichlet distribution and Cat is a categorical distribution. $\mathcal{W}^{-1}$ is an Inverse-Wishart distribution and $\mathcal{N}$ is a normal distribution.

For example, we cannot probabilistically interpret the center of each topic cluster returned by other sentence-level topic models. In this perspective, each sentence is just 'a point which is close to the center of a topic cluster'. But SentenceLDA's topic embedding is the mean and mode of each Gaussian topic distribution, and each sentence is a random sample from the topic distribution. We can also obtain the variance of each topic distribution, unlike other sentence-level topic models. As a result, SentenceLDA succeeds LDA's elegant document-generative process while considering sentence-level information.

Moreover, SentenceLDA can fully utilize sentence-level information. By utilizing the sentence-level information, SentenceLDA is not biased toward word spelling. Consequently, SentenceLDA can discriminate documents having different topics while sharing similar lexical distribution. Moreover, since SentenceLDA does not assume exchangeability between words, it can capture the difference between two sentences sharing the same words but having different word order. On the other hand, SentenceLDA can robustly capture the semantics of two documents sharing the same meaning but having different lexical distributions or syntactic structures. As a result, SentenceLDA returns a more discriminative and robust document representation by considering sentence-level information. We evaluate these properties on Section 4.

We assume the existence of a feasible encoder f_{enc} to encode each sentence into embedding space. Also, we need a decoder f_{dec} to decode the sentence embedding to a natural language sentence. We can improve the SentenceLDA by exchanging the encoder and decoder thanks to its modular structure. For implementation, we use SentenceBERT (Reimers and Gurevych, 2019) (*all-mpnet-base-v2*) as an f_{enc} and fine-tuned GPT2-XL as f_{dec}. For the topic model parameter inference, we utilize the collapsed Gibbs sampling with Cholesky decomposition as GaussianLDA, which is presented in Das et al. (2015).

4 Experiments

4.1 Setting

Compared Methods: We compare two word-level topic models, one sentence-level topic model, and one hybrid topic model with our SentenceLDA (**SLDA**).

- **LDA** (Blei et al., 2003): The original LDA introduced in Table 2. We utilize the gensim library.
- GaussianLDA (**GLDA**, Das et al., 2015): Word-level GaussianLDA introduced in Table 2. We utilize Python implementation[2] with 300-dimension Word2Vec embedding (Mikolov et al., 2013) pretrained on news corpus (*word2vec-google-news-300*).
- ContextualTM (**CTM**, Bianchi et al., 2021a): Extended version of ProdLDA (Srivastava and

[2]https://github.com/markgw/gaussianlda.

Sutton, 2017), a neural topic model, employing PLM. It takes a Bag-of-Words and contextual embedding vector as input. We use SentenceBERT embedding as ours since our goal is to compare how well the model handles the contextual information. We use the official implementation.[3]

- **SenClu** (Schneider, 2023): Sentence-level topic model which directly uses SentenceBERT embedding of each sentence in the generating process. Unlike SentenceLDA, SenClu computes the likelihood of a sentence given a topic with a similarity metric between sentence and topic embeddings. We also use SentenceBERT like other models and use the official implementation.[4]

For word-level topic models, we lowercase documents and remove non-alphanumeric characters and stopwords as standard topic modeling practice. But for SenClu, SentenceLDA, and the contextual part of ContextualTM, we only apply minimal preprocessing to preserve the contextual information in documents. For more details, see Appendix A.

Datasets: We evaluate our framework on two datasets.

- The 20 Newsgroups (**20News**, Lang, 1995): The **20News** dataset consists of 18,846 documents with 6 coarse-grained classes and 20 fine-grained classes. We split the dataset into training (60%) and test (40%) sets and remove metadata-related information, as recommended in *scikit-learn*.[5] After removing non-alphanumeric characters, we obtain 18,327 non-empty documents.

- The New York Times (**NYT**, Mekala et al., 2021): The **NYT** dataset consists of 11,744 documents with 5 coarse-grained classes and 26 fine-grained classes. We split the dataset into training (80%) and test (20%) sets, maintaining label weights, and apply the same preprocessing as 20News.

20News is a conventionally used dataset in topic modeling literature, and both **20News** and **NYT** are also widely used in text classification literature (Mekala et al., 2021, Wang et al., 2021 and Mekala and Shang, 2020). Summary statistics are on Table 3.

Dataset	Split	Doc	Sent/Doc	Tok/Doc
20News	Train	11.0k	12.0	201.5
	Test	7.3k	11.0	189.9
NYT	Train	9.3k	34.8	688.7
	Test	2.3k	35.3	692.2

Table 3: Summary statistics for each dataset. Doc is the number of documents, Sent/Doc is the number of sentences per document, and Tok/Doc is the number of tokens per document.

To divide each document into sentences, we apply pySBD (Sadvilkar and Neumann, 2020), State-of-the-Art rule-based sentence boundary detector.[6] To check which category is classified into which coarse-grained class, see Appendix B.

4.2 Discriminativeness of Document Representation

As Li et al. (2016) and Liu et al. (2015), we conduct text classification to investigate the discriminativeness of document representation. To check the discriminativeness between topics sharing similar lexical distribution, we divide each dataset into coarse-grained classes. For **20News**, we divide it into 6 coarse-grained classes, *computer, ride, sports, science, religion* and *politics*. For **NYT**, we also divide it into 5 coarse-grained classes, *arts, business, politics, science* and *sports*.

For each training set, we train each topic model to find 10/20 topics and infer the topic distribution of each document on the training and test sets. Then we train a classifier on the training document's topic distribution to predict class. Finally, we check the predictive performance on test documents. We iterate this procedure 5 times for each algorithm and compute the mean and standard deviation of accuracy. Since categorizing 20-26 categories with 10 topics is infeasible, we only present 20 topics results for 'All' category.

SentenceLDA shows more discriminative power in most cases than other topic models, in both logistic regression (Table 4) and random forest predictors scenarios (Appendix C). The same pattern appears for the F1 score (Appendix D). Especially, SentenceLDA's performance dominates CTM and

[3]https://github.com/MilaNLProc/contextualized-topic-models.

[4]https://github.com/johntailor/bertsenclu. Since the author didn't provide topic inference code for test documents, we trained the topic model by merging train and test corpora, which may give an advantage to SenClu.

[5]https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html.

[6]https://github.com/nipunsadvilkar/pySBD.

Dataset	Topics	Class	LDA	GLDA	CTM	SenClu	SLDA (Ours)
20News	10	Computer (5)	44.99% (2.68)	24.03% (1.47)	36.34% (4.05)	**45.66% (3.35)**	42.25% (0.72)
		Ride (2)	64.05% (5.13)	51.92% (0.74)	75.40% (4.91)	73.13% (3.50)	**82.19% (0.84)**
		Sports (2)	76.61% (7.09)	63.29% (2.02)	84.29% (3.12)	77.33% (13.75)	**88.70% (1.53)**
		Science (4)	65.03% (2.32)	30.56% (1.75)	64.11% (3.16)	76.01% (2.01)	**78.38% (0.85)**
		Religion (3)	49.30% (3.05)	40.89% (0.08)	47.10% (2.60)	54.45% (3.55)	**58.64% (0.94)**
		Politics (3)	60.90% (3.21)	37.94% (1.31)	60.80% (3.23)	62.57% (4.14)	**68.28% (0.67)**
	20	Computer (5)	43.73% (1.98)	26.65% (1.39)	37.40% (3.76)	47.69% (3.60)	**52.20% (2.25)**
		Ride (2)	64.39% (2.56)	55.13% (0.80)	78.42% (3.95)	74.34% (2.50)	**80.66% (1.04)**
		Sports (2)	72.57% (5.19)	63.26% (2.34)	87.40% (2.28)	83.86% (2.61)	**88.43% (0.55)**
		Science (4)	66.67% (2.12)	34.75% (2.92)	69.06% (2.57)	76.35% (1.99)	**78.64% (0.76)**
		Religion (3)	46.89% (1.22)	40.85% (0.00)	51.74% (1.76)	57.02% (1.93)	**59.96% (1.15)**
		Politics (3)	63.06% (2.91)	39.57% (1.93)	60.91% (5.09)	68.14% (1.71)	**71.22% (0.96)**
		All (20)	37.99% (2.39)	8.72% (0.35)	34.55% (1.66)	40.91% (2.51)	**42.73% (3.51)**
NYT	10	Arts (4)	65.24% (5.41)	39.52% (0.00)	73.90% (4.65)	78.47% (6.97)	**93.14% (0.83)**
		Business (4)	72.63% (2.72)	46.97% (0.00)	74.24% (3.03)	62.83% (5.93)	**74.85% (2.22)**
		Politics (9)	60.20% (2.23)	41.79% (0.00)	61.09% (2.30)	60.40% (5.94)	**66.86% (0.67)**
		Science (2)	85.26% (6.14)	55.79% (2.58)	78.95% (7.44)	90.52% (2.11)	**91.58% (2.58)**
		Sports (7)	**91.91% (3.77)**	25.72% (0.00)	75.04% (3.75)	71.64% (6.33)	69.33% (3.99)
	20	Arts (4)	71.05% (4.22)	39.52% (0.00)	75.71% (6.10)	85.91% (2.82)	**95.81% (0.76)**
		Business (4)	72.42% (5.19)	46.97% (0.00)	77.48% (3.52)	75.96% (4.31)	**78.48% (1.45)**
		Politics (9)	65.67% (2.16)	41.79% (0.00)	63.48% (1.85)	69.55% (3.14)	**73.13% (1.75)**
		Science (2)	78.95% (11.04)	54.73% (2.58)	66.31% (5.37)	80.00% (2.10)	**89.47% (3.33)**
		Sports (7)	**96.59% (0.35)**	25.86% (0.20)	86.24% (1.12)	85.84% (5.88)	89.50% (2.03)
		All (26)	**82.98% (2.84)**	19.48% (0.25)	70.80% (1.78)	64.86% (6.24)	65.65% (1.12)

Table 4: Mean and standard deviation of the accuracy of a linear logistic classifier trained on each topic distribution. The number in the parenthesis next to the Class is the number of categories in each Class, and next to each accuracy is the standard deviation. The highest score is marked as bold, and the second highest score is marked with underline.

SenClu as well in most of the tested classes, though they utilize the same SentenceBERT. The result suggests that SentenceLDA is a promising way to combine modern NLP techniques with classic topic models.

One notable point is that topic models considering contextual information, CTM and SenClu, outperform other word-level topic models. It shows the importance of sentence-level information to obtain discriminative document representation.

The weak classification performance of GaussianLDA was previously reported (Li et al., 2016). We verify that the GaussianLDA returns nearly identical distribution to any document (See Appendix E). Though GaussianLDA and SentenceLDA share the nearly same generative and inference algorithm, significant improvement is achieved by modifying the model's unit from word to sentence. It is encouraging since only a few modifications to the GaussianLDA source code are applied to implement SentenceLDA.

One outlier case is **NYT**-Sports, where the SenClu and SentenceLDA significantly underperform other word-level topic models. Since 73.5% of **NYT** dataset belongs to Sports-related categories, they show worse performance on 'All' category, consequently. For the error analysis, we compare it with the LDA result and find out that extracted topics are biased toward the baseball category. We present the error analysis on Appendix F.

To check the relationship between a sentence embedding and SentenceLDA, we implement ablation studies on Appendix G and Appendix H. We find out that SentenceLDA returns a more generalizable document representation than a sentence embedding itself.

4.3 Robustness to Paraphrasing

Paraphrasing is to express the meaning of a text using different words or expressions while maintaining its original semantics. To robustly process the contextual information, a topic model should maintain its original topical inference on paraphrases. Here we introduce a novel task that measures the semantic robustness of a topic model with paraphrasing. First, we paraphrase the **20News** and **NYT** in two ways.

- Lexical: Substitute words with synonyms cap-

			Lexical				Syntactic			
Metric	Corpus	Topics	LDA	GLDA*	CTM	SLDA	LDA	GLDA*	CTM	SLDA
D_{sum}	**20News**	10	0.1868	0.0109	**0.1475**	0.1689	0.0765	0.0096	0.1319	**0.0743**
		20	0.2214	0.0210	**0.1636**	0.2223	**0.0909**	0.0156	0.1471	0.1020
	NYT	10	0.1814	0.0122	0.1645	**0.0808**	**0.0342**	0.0048	0.1222	0.0346
		20	0.1823	0.0133	0.1747	**0.1352**	**0.0434**	0.0090	0.1386	0.0594
τ	**20News**	10	0.7460	0.9360	0.5487	**0.8286**	0.8971	0.9500	0.5872	**0.9259**
		20	0.7319	0.9014	0.5765	**0.7748**	0.8875	0.9329	0.6096	**0.8973**
	NYT	10	0.7626	0.7790	0.5506	**0.9145**	0.9237	0.9548	0.5973	**0.9587**
		20	0.7647	0.8757	0.5149	**0.8624**	0.9194	0.9406	0.5454	**0.9326**

Table 5: Robustness to paraphrasing. The lower D_{sum} and higher τ represent better robustness.

tured by the WordNet (Miller, 1992). We only utilize the word with Wu & Palmer similarity (Guessoum et al., 2016) with the original word higher than 0.5.

- Syntactic: Parrot paraphraser (Damodaran, 2021) is a T5 (Raffel et al., 2020) based paraphraser trained on various corpus. It tends to mainly modify the syntactic form of a sentence while maintaining the original vocabulary.

We use two metrics to measure the topic distribution change before and after the paraphrasing. The first one is a naive summation of the absolute difference between two topic distributions, i.e. $D_{sum}(P,Q) = \frac{1}{2}\sum_{i=1}^{K}|P_i - Q_i|$. If two distributions P and Q are totally different and do not have anything in common, $D_{sum}(P,Q) = 1$, while if two distributions are identical, $D_{sum}(P,Q) = 0$. The second one is Kendall's rank correlation τ (Kendall, 1938). It ranges from -1 to 1 showing the consistency of rank between two distributions. We exclude SenClu since the source code does not support inference on unseen paraphrased text. GaussianLDA is marked as a star sign since it returns nearly identical topic distribution for any document as mentioned on Section 4.2.

The result shows SentenceLDA is robust to both lexical and syntactic changes if the semantics of the text are preserved as presented on Table 5.

For D_{sum} - Lexical case, ContextualTM and SentenceLDA outperform the word-level topic model, LDA. It is a reasonable result since LDA utilizes only word spelling information, as a result, changes in words highly affect the LDA. ContextualTM shows robustness on **20News** while SentenceLDA works better on **NYT**. Since they utilize contextual information and sentence-level embedding, they are not relatively affected by lexical changes.

For D_{sum} - Syntactic case, LDA outperforms other topic models. It is also reasonable since syntactic paraphraser tends to maintain original vocabulary, mainly modifying word order. (98.72% of words in **20News** paraphrase is observed in the original text, and 97.75% for **NYT**.) As a result, the word-level topic model LDA with exchangeability assumption is the most robust for this task. But SentenceLDA shows comparable robustness to LDA though it does not assume the exchangeability. Meanwhile, ContextualTM shows the worst robustness though it also considers contextual information like SentenceLDA.

For τ case, SentenceLDA shows the highest rank correlation in all cases. ContextualTM significantly deteriorates than D_{sum} case. We presume that ContextualTM usually predicts a relatively flat distribution. As a result, the mean distribution change D_{sum} for ContextualTM is modest while the topic rank dynamically fluctuates. To verify this, we measure the entropy of each topic distribution and ContextualTM shows consistently higher entropy than other topic models (See Appendix I). Meanwhile, as observed in D_{sum} case, LDA shows a high correlation for syntactic paraphrasing while a lower correlation for the lexical case.

4.4 Qualitative Evaluation

To check the extracted topics of the SentenceLDA, we train GPT2-XL (Radford et al., 2019) in an auto-encoding scheme. In other words, we train GPT2-XL to reconstruct a sentence from the Sentence-BERT embedding of the sentence. We use Huggingface Transformers implementation of GPT2-XL[7], on WikiText-103 corpus (Merity et al., 2016),

[7]https://huggingface.co/gpt2-xl.

Dataset	Model	Category	Extracted Topics
NYT-*politics*	SLDA	*immig.*	Many immigrants in the United States are concerned about the lack of legal access to health care and other social services in their communities, and the recent push for a comprehensive immigration reform bill has only increased these concerns.
	LDA	*immig.* *ACA* *budget*	said, immigrants, immigration, border, major, hasan health, care, insurance, people, coverage, medicaid said, alexis, navy, bill, like, states, north, year
NYT-*arts*	SLDA	*music*	In the meantime, the New York Philharmonic Orchestra has staged a series of concerts with the Vienna State Opera under the baton of the celebrated composer, with music by Bartk, and with a libretto by Richard D Oyly Carte.
	LDA	*television* *dance* *music*	like, show, one, said, shows, television, series new, orchestra, britten, two, festival, one said, new, ballet, music, opera, ms, dance

Table 6: Extracted topics from **NYT** dataset by SentenceLDA and LDA. We highlight overlapped keywords with different colors for different topics from LDA. *immig.* is the immigration category, *ACA* is the Affordable Care Act category, and *budget* is Federal budget category in **NYT** dataset.

which consists of 28.5k Wikipedia articles and 100 million tokens.[8] Though we can use a corpus-wise decoder trained on **20News** and **NYT**, we choose to use a Wiki-based corpus to obtain a decoder that can handle general information. Here we concentrate on LDA and SentenceLDA since LDA shows the best performance on classification tasks among word-level topic models.

We arrange the extracted words or sentence-level topics which are decided to be important for classification. Since we use a simple linear logistic regression model, it is easy to extract important feature topics, just by sorting the weight matrix. We present selected results for **NYT** dataset on Table 6.

From the table, we observe a topic sentence extracted by SentenceLDA includes words from various topics extracted by LDA. Especially, for **NYT**-*politics*, though the extracted sentence belongs to *immig.* category, it contains the word 'health care', which seems valid to be included in *ACA* (The Affordable Care Act) category by itself. But with the given context, 'the lack of legal access to health care' for 'immigrant' is more appropriate to be included in *immig.* category, since it is more about the 'immigrant', not the Affordable Care Act.

Likewise, for **NYT**-*arts*, LDA classifies the word 'series' into a television-related category, maybe because of the word 'TV Series'. But as shown in the extracted topic sentence from SentenceLDA, the word become totally different meaning with the given context, 'series of concerts'. It shows that considering only word-level information is insufficient and easy to be biased to the spelling of the word itself, while sentence-level information can consider rich contextual meaning in each word. In conclusion, SentenceLDA can discriminate the same words in different contexts.

5 Application: Key Opinion Mining from an Argumentative Corpus

Utilizing the discriminative power and robustness of SentenceLDA, we can robustly discriminate subtle nuances of a word in different contexts. One domain where capturing subtle nuances of words is important is argumentative corpus. In an argumentative corpus, word senses are complexly entangled. For example, the word 'Korea' may contain totally different senses when the given context is economic development, democracy, human rights, or nuclear weapons.

Many researchers applied topic modeling to opinion mining-related tasks, e.g. sentiment classification (Vamshi et al., 2018), social community detection (Chen et al., 2017), and opinion summarization (Isonuma et al., 2021). However, the majority of these researches are limited to social media texts or product review datasets, which are relatively short and straightforward, unlike argumentative corpus.

Meanwhile, in the argument mining domain, researchers have utilized topic modeling techniques to access external knowledge through knowledge graph (Li et al., 2021) or to classify each sentence into argumentative unit categories (e.g. claims, and premises) in supervised (Habernal and Gurevych, 2015) and unsupervised manner (Ferrara et al.,

[8] https://blog.salesforceairesearch.com/the-wikitext-long-term-dependency-language-modeling-dataset/.

Model	Extracted Topics
SLDA	1. With the war in Ukraine, Russia has not been able to count on the United States and Europe to keep Moscows feet firmly to the fire, much less to revive the stalled SinoRussian economic cooperation. 6. As China grows more powerful, it is increasingly at odds with Japan, which has a strong economic stake in the success of SinoAmerican relations and is understandably nervous about Beijings intentions in the South China Sea. 10. The worlds oceans have been shown to be less able to absorb and store carbon dioxide and other greenhouse gases, and the number of species known to be experiencing reduced populations has been rising since the 1950s.
LDA	1. nuclear, would, weapons, iran, war, states, could, us, one, attack 3. fish, ocean, one, species, global, change, said, also, data, warming 5. energy, oil, gas, us, said, prices, also, years, new, industry 6. china, us, military, russia, trade, said, japan, security, new, would 8. states, war, world, power, china, conflict, economic, political, united, global

Table 7: Extracted topics from DebateSum dataset - 'Impact Defense Core' topic. We highlight overlapped keywords with different colors for different topics from SLDA. Numbering is arbitrary.

2017). In both ways, topic information is used as one of features. But, by the nature of topic modeling, we can extract key topic words composing an argumentative corpus with word-level topic models. By using the SentenceLDA, we can extract key topic sentences that can be regarded as key opinions of the corpus. To the best of our knowledge, there is no research extracting key opinions from an argumentative corpus using a topic model.

To explore the possibility, we apply the SentenceLDA to DebateSum dataset (Roush and Balaji, 2020). DebateSum dataset consists of 187,386 debate documents extracted from the National Speech and Debate Association over a 7-year period. Since it includes multiple heterogeneous debate topics from national defense to LGBT, we filter it with the keyword 'Impact Defense Core'. As a result, we obtain 762 debate documents with 12,957 sentences. We apply SentenceLDA and LDA with 10 topics, and the result is on Table 7. To decode the sentence embedding, we train another GPT2-XL on the DebateSum corpus.

Though the LDA captures 'war' (1), 'russia' (6), 'economic' (8) as separate topics, the first topic from the SentenceLDA captures a more nuanced topic related to the Ukraine war and its economic consequences related to western countries. Likewise, the LDA captures 'ocean (sea)' (3), 'china' and 'japan' (6), and 'economic' (8) separately, while the sixth topic sentence from the SentenceLDA captures the complicated relationship between the words from economic history to current dispute on the South China Sea between China and Japan. The LDA classifies 'gas' (5) into an energy-related topic and separates it from the environment-related topic (3), but the 10th topic sentence from the SentenceLDA shows that 'gas' can be used in the context of the earth's environment.

In summary, the SentenceLDA enables us to obtain a more nuanced and comprehensive understanding of an argumentative corpus because it considers the relationships between words within the context of the sentence. While LDA can provide insight into the most frequently occurring words and themes in the corpus, it may miss important nuances and variations in meaning that are present in longer text units.

6 Conclusion

We introduce SentenceLDA, a sentence-level topic model combining modern sentence embedding techniques with a classic topic model. With a simple modification from LDA and GaussianLDA, SentenceLDA succeeds LDA's elegant probabilistic interpretability. We demonstrate that the sentence-level topic model can return more discriminative and robust document representation compared to word-level topic models. We evaluate the discriminativeness with a text classification task by comparing topic models on two classification datasets. As a result, SentenceLDA significantly outperforms other models by not being biased toward word spelling but considering word sense in context. By evaluating the robustness of each model with the paraphrased dataset, we observe that SentenceLDA predicts topic distribution for paraphrases more robustly than other topic models. Especially, SentenceLDA returns more discriminative and robust document representation compared to topic mod-

els utilizing the same contextual information, ContextualTM and SenClu. Also, we apply SentenceLDA to an argumentative corpus, DebateSum, and demonstrate the applicability of SentenceLDA to corpus-wide key opinion mining. As a result, we show that SentenceLDA is a powerful tool to obtain discriminative and robust document representation.

Limitations

Though our SentenceLDA is promising, as Li et al. (2016) mentioned, GaussianLDA baseline code is slow in training and inference, as a result, the same sluggishness happens in SentenceLDA. For example, 20 sampling iterations on the **NYT** dataset with 20 topics take nearly a day with a high-performance workstation. The main computational bottleneck occurs in a for loop, which processes each sentence iteratively on the CPU, and it may be improved with batch-wise inference on GPU. Since one of our goals was to implement the SentenceLDA with minimal code modification from the GaussianLDA baseline, we leave it as a future task.

Another limitation of our work is that we cannot verify whether the generated topic sentences are factual or not. Many researchers are trying to solve this anti-factual decoding problem using knowledge-infused decoding (Liu et al., 2022) or knowledge graph (Chaudhuri et al., 2021). We hope development in these factual decoding techniques guides SentenceLDA to return more fact-based results.

The usual assessment tool for a topic model is to measure the coherence and diversity score of each model. But conventional practice for the measurement is fitted to word-level topic models, not sentence-level. As a result, we test SentenceLDA's discriminative performance with a classification task as a surrogate. Additionally, we compute BERTScore (Zhang* et al., 2020) between generated topic sentences to check the coherence of our topic model on Appendix J. We hope development in reasonable and robust measures for sentence-level topic models in future research.

Acknowledgements

This work is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1G1A1102828). This work is supported by the Junior Fellow Research Grant funded by the Korea University.

References

Georgios Balikas, Massih-Reza Amini, and Marianne Clausel. 2016. On a topic model for sentences. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 921–924, 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, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Debanjan Chaudhuri, Md Rashad Al Hasan Rony, and Jens Lehmann. 2021. Grounding dialogue systems via knowledge graph aware decoding with pre-trained transformers. In *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*, pages 323–339. Springer.

Hongxu Chen, Hongzhi Yin, Xue Li, Meng Wang, Weitong Chen, and Tong Chen. 2017. People opinion topic model: Opinion based user clustering in social networks. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 1353–1359, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Prithiviraj Damodaran. 2021. Parrot: Paraphrase generation for nlu.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 795–804, Beijing, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Alfio Ferrara, Stefano Montanelli, and Georgios Petasis. 2017. Unsupervised detection of argumentative units though topic modeling techniques. In *Proceedings of the 4th Workshop on Argument Mining*, pages 97–107, Copenhagen, Denmark. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Djamel Guessoum, Moeiz Miraoui, and Chakib Tadj. 2016. A modification of wu and palmer semantic similarity measure. In *The Tenth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, pages 42–46.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137, Lisbon, Portugal. Association for Computational Linguistics.

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)*, pages 1752–1771, Online. Association for Computational Linguistics.

Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2021. Unsupervised Abstractive Opinion Summarization by Generating Sentences with Tree-Structured Topic Guidance. *Transactions of the Association for Computational Linguistics*, 9:945–961.

Haixin Jiang, Rui Zhou, Limeng Zhang, Hua Wang, and Yanchun Zhang. 2019. Sentence level topic models for associated topics extraction. *World Wide Web*, 22:2545–2560.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Olzhas Kozbagarov, Rustam Mussabayev, and Nenad Mladenovic. 2021. A new sentence-based interpretative topic modeling and automatic topic labeling. *Symmetry*, 13(5).

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339.

Shaohua Li, Tat-Seng Chua, Jun Zhu, and Chunyan Miao. 2016. Generative topic embedding: a continuous representation of documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 666–675, Berlin, Germany. Association for Computational Linguistics.

Shuangyin Li, Yu Zhang, Rong Pan, Mingzhi Mao, and Yang Yang. 2017. Recurrent attentional topic model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Weichen Li, Patrick Abels, Zahra Ahmadi, Sophie Burkhardt, Benjamin Schiller, Iryna Gurevych, and Stefan Kramer. 2021. Topic-guided knowledge graph construction for argument mining. In *2021 IEEE International Conference on Big Knowledge (ICBK)*, pages 315–322.

Ruibo Liu, Guoqing Zheng, Shashank Gupta, Radhika Gaonkar, Chongyang Gao, Soroush Vosoughi, Milad Shokouhi, and Ahmed Hassan Awadallah. 2022. Knowledge infused decoding. *arXiv preprint arXiv:2204.03084*.

Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015. Topical word embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Dheeraj Mekala, Varun Gangal, and Jingbo Shang. 2021. Coarse2Fine: Fine-grained text classification on coarsely-grained annotated data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 583–594, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dheeraj Mekala and Jingbo Shang. 2020. Contextualized weak supervision for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 323–333, Online. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *CoRR*, abs/1609.07843.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Ramesh Nallapati, Igor Melnyk, Abhishek Kumar, and Bowen Zhou. 2017. Sengen: Sentence generating neural variational topic model.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mehdi Rezaee and Francis Ferraro. 2020. A discrete variational recurrent topic model without the reparametrization trick. *Advances in neural information processing systems*, 33:13831–13843.

Allen Roush and Arvind Balaji. 2020. DebateSum: A large-scale argument mining and summarization dataset. In *Proceedings of the 7th Workshop on Argument Mining*, pages 1–7, Online. Association for Computational Linguistics.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Javier Miguel Sastre Martinez, Sean Gorman, Aisling Nugent, and Anandita Pal. 2022. Generating meaningful topic descriptions with sentence embeddings and LDA. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 244–254, Edinburgh, UK. Association for Computational Linguistics.

Johannes Schneider. 2023. Efficient and flexible topic modeling using pretrained embeddings and bag of sentences.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Laure Thompson and David Mimno. 2020. Topic modeling with contextualized word representation clusters. *CoRR*, abs/2010.12626.

Fei Tian, Bin Gao, Di He, and Tie-Yan Liu. 2016. Sentence level recurrent topic model: Letting topics speak for themselves.

Krishna B Vamshi, Ajeet Kumar Pandey, and Kumar A. P. Siva. 2018. Topic model based opinion mining and sentiment analysis. In *2018 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–4.

Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. 2009. Multi-document summarization using sentence-based topic models. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 297–300, Suntec, Singapore. Association for Computational Linguistics.

Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Topic-guided variational auto-encoder for text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 166–177, Minneapolis, Minnesota. Association for Computational Linguistics.

Zihan Wang, Dheeraj Mekala, and Jingbo Shang. 2021. X-class: Text classification with extremely weak supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3043–3053, Online. Association for Computational Linguistics.

Min Yang, Tianyi Cui, and Wenting Tu. 2015. Ordering-sensitive and semantic-aware topic modeling. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2353–2359. AAAI Press.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. 2022. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

pages 3886–3893, Seattle, United States. Association for Computational Linguistics.

A Experimental Details

We use a machine with AMD Ryzen 9 5900X 12-Core Processor CPU with one NVIDIA RTX 3090 GPU. For each topic model, we use the default setting without any hyperparameter tuning. But for GaussianLDA and SentenceLDA, we find out that default prior to the covariance matrix doesn't work at all. We find out that the default scale term on the diagonal covariance matrix is set too high, and decrease it to 0.1, which works well on both settings. We run 20 training and inference sampling on each dataset for the ContextualTM, GaussiaLDA, and SentenceLDA, and check that the loglikelihood converges. We run 50 iterations for LDA and 40 for SenClu for the same reason.

We use a pre-trained SentenceBERT encoder (*all-mpnet-base-v2*, 109M parameters) without fine-tuning, and fine-tuned GPT2-XL decoder (1.5B parameters) on the corpus we mentioned. Since GPT2-XL is too large to load on the RTX 3090, we use NVIDIA A100 GPU to train it, with batch size 64 and ADAMW optimizer with learning rate 1e-5, for 5 epochs. Since we want the decoder to return only one sentence, we truncate the original sentence to be a maximum of 64 tokens and train it in an auto-encoding scheme. Note that since the embedding dimension of the SenteceBERT (768) is different from the GPT2-XL embedding dimension (1600), we repeat the smaller one until it matches the larger one.

B Explanations on Coarse-grained Categories

The **20News** dataset consists of 20 fine-grained categories while the **NYT** dataset consists of 26 fine-grained categories. We divide them into 6 and 5 coarse-grained categories and use each as a separate dataset. For **20News**, we exclude "misc.forsale" category in the fine-grained setting since it does not belong to any of the other high-level categories (Computer, Ride, Sports, Science, Religion, Politics). How we divide them is presented on the Table 15.

C Classification Results of Random Forest Classifier

We present the classification results with a non-linear random forest classifier. We utilize scikit-learn package with default parameters. The results are on Table 13.

D F1 Scores on Text Classification

Here we present macro F1-score (Table 14) for classification. We can check that the SentenceLDA still outperforms the other models in most cases.

E Mean Deviation of Topic Distribution

To examine the low performance of GaussianLDA on text classification task on Section 4.2, we hypothesize that GaussianLDA returns nearly similar distribution for any document. To verify this, we measure the mean deviation of the topic distribution. Mathematically, we compute $\frac{1}{K}\sum_{i=1}^{K} D_i$ where $D_i = SD[p_i]$ is the standard deviation of i-th topic probability for all documents. The results are on Table 8.

Corpus	Topics	LDA	GLDA	CTM	SLDA
20News	10	0.1195	0.0191	0.0725	0.1755
	20	0.0768	0.0149	0.0395	0.1273
NYT	10	0.1220	0.0064	0.0947	0.1349
	20	0.0708	0.0057	0.0538	0.0913

Table 8: Mean deviation of the topic distribution.

As the table shows, we verify that the GaussianLDA returns nearly the same distribution for any documents. As a result, the GaussianLDA shows the worst performance for text classification on Section 4.2, while showing no distributional change to paraphrases on Section 4.3

F Error Analysis

As both sentence-level topic models, SenClu and SentenceLDA, underperform on 10 Topics-**NYT**-Sports than the other classes, we perform qualitative error analysis for this case. We compare SentenceLDA with the LDA on the Table 16.

We find out that the LDA result contains topics related to all categories on Table 15-**NYT**-Sports, e.g. soccer (2), football (3), basketball (4), baseball (5, 7), hockey (6), golf (8) and tennis (10). But for the SentenceLDA, we find only five categories within the topics, e.g. hockey (1), baseball (2, 3, 9, 10), golf (6), football (7), and basketball (8), while tennis and soccer-related topics are not found.

We suspect the encoder is biased since 62% of the training set used to train the SentenceBERT-*all-mpnet-base-v2* is from the Reddit data between

2015 to 2018[9]. Through the Google Search, we check that 8,480,000 pages are found with the keyword 'soccer' and 3,830,000 pages are found with 'tennis', and 14,500,000 pages are found with 'baseball' from Reddit. Though the SentenceLDA gives contextual meaning to vague words, like game, league, and points, this result shows the model is dependent on the encoder. But since the SentenceLDA is a modular structure, we hope it can be resolved with improved encoder and decoder techniques.

G Ablation Study

Since we utilize SentenceBERT, it is natural to ask about the degree of the contribution of the sentence embedding itself (SBERT). We train another logistic classifier which takes the mean of sentences' embedding in a document. We compare the answer accuracy with our SentenceLDA - 20 topics. The results are on Table 9.

Dataset	Class	SBERT	SLDA
20News	Computer	34.96%	**52.20%**
	Ride	70.79%	**80.66%**
	Sports	80.60%	**88.43%**
	Science	52.07%	**78.64%**
	Religion	57.34%	**59.96%**
	Politics	63.68%	**71.22%**
NYT	Arts	56.19%	**95.81%**
	Business	65.15%	**78.48%**
	Politics	**75.62%**	73.13%
	Science	84.21%	**89.47%**
	Sports	**94.57%**	89.50%

Table 9: Answer accuracy of the linear logistic classifier with SentenceBERT embedding and SentenceLDA topic distribution.

As shown in the table, SentenceLDA outperforms the SentenceBERT embedding in most cases. We observe that the logistic regressor learns more generalizable features avoiding over-fitting since SentenceBERT embedding has 768 dimensions, while a topic distribution from SentenceLDA has only 20 dimensions. We compute generalization error by subtracting the classifier's test accuracy score from the train accuracy score. We compute the mean and standard deviation of the generalization error and present it on the Table 10. From the table, we can see that SentenceLDA topic distribution is a more generalizable feature than sentence embedding itself.

[9]https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

Dataset	SBERT	SLDA
20News	31.18% (7.93)	**2.94% (3.12)**
NYT	24.05% (13.01)	**5.44% (6.05)**

Table 10: Generalization error for each model and dataset. Numbers in the parenthesis are standard deviations.

H Classification Results of Different Sentence Embeddings

We present classification results of SentenceLDA with different sentence embedding models. We compare three sentence embedding models, *multi-qa-mpnet-base-dot-v1* (**QA**), *all-distilroberta-v1* (**Distill**) and *all-mpnet-base-v2* (**SLDA**). The results are on Table 17

I Entropy of Topic Distribution

We present the mean entropy of topic distribution for each model on Table 11. ContextualTM shows the highest entropy, which implies ContextualTM predicts a 'flatter' topic distribution. As a result, the sum of the change in topic distribution D_{sum} is relatively small, while topic rank fluctuates.

Corpus	Topics	LDA	GLDA	CTM	SLDA
20News	10	0.1066	0.0156	0.2061	0.0822
	20	0.0613	0.0141	0.1360	0.0533
NYT	10	0.0958	0.0166	0.1780	0.0666
	20	0.0646	0.0165	0.1178	0.0562

Table 11: Entropy of topic distribution of test corpus. Intuitively, the higher entropy, the flatter the topic distribution is.

J Topic Coherence Analysis

Conventionally, normalized pointwise mutual information (NPMI) is frequently used to evaluate the coherence of word-level topic models. To compute the NPMI for the sentence-level topic model, Wikipedia should contain two exact topic sentences multiple times. However, computing the score with the sentence-level topic model is impossible since it is improbable that two exact sentences appear in multiple documents.

Instead, we sample 3 sentences for each topic and compute BERTScores. We utilize 20News - Total - 10 Topics case and present it on Table 18. Though some topics are hard to interpret (like Topic

1), high BERTScore shows the topic sentences share similar contents, i.e. coherent.

K Comparison to BERTopic

We implement BERTopic (Grootendorst, 2022) and compare its discriminative performance with our model. Because of the nature of HDBSCAN, BERTopic returns less than 10/20 topics for some categories. As a result, we could not implement categorization with some categories. We mark it with X.

Topics	Dataset	Class	SLDA	BERTopic
10	**20News**	Computer	**42.25%**	40.03%
		Ride	**82.19%**	74.60%
		Sports	88.70%	**93.27%**
		Science	78.38%	**78.92%**
		Religion	**58.64%**	42.65%
		Politics	68.28%	**74.78%**
	NYT	Arts	**93.14%**	76.19%
		Business	**74.85%**	61.61%
		Politics	**66.86%**	50.74%
		Science	91.58%	X
		Sports	69.33%	**81.84%**
20	**20News**	Computer	52.20%	**54.00%**
		Ride	**80.66%**	73.68%
		Sports	88.43%	**90.55%**
		Science	78.64%	**83.58%**
		Religion	**59.96%**	54.25%
		Politics	71.22%	**72.05%**
	NYT	Arts	95.81%	X
		Business	**78.48%**	65.65%
		Politics	**73.13%**	61.19%
		Science	89.47%	X
		Sports	89.50%	**91.21%**

Table 12: Accuracy score of a linear logistic classifier with SentenceLDA and BERTopic.

Dataset	Topics	Class	LDA	GLDA	CTM	SenClu	SLDA (Ours)
20News	10	Computer (5)	39.68% (1.30)	23.40% (0.29)	35.43% (3.36)	**42.22% (3.22)**	39.11% (1.24)
		Ride (2)	58.79% (3.21)	53.42% (1.47)	73.53% (5.43)	69.76% (4.19)	**78.82% (1.81)**
		Sports (2)	73.87% (7.79)	60.11% (1.67)	85.07% (3.05)	76.48% (13.15)	**88.98% (1.77)**
		Science (4)	61.60% (2.75)	30.09% (1.14)	64.16% (4.18)	74.96% (2.21)	**77.64% (0.36)**
		Religion (3)	46.70% (4.14)	41.15% (0.56)	46.77% (2.76)	51.30% (2.84)	**53.21% (1.05)**
		Politics (3)	57.58% (1.33)	40.39% (2.21)	61.89% (3.69)	61.97% (3.41)	**69.87% (0.91)**
	20	Computer (5)	42.51% (1.84)	26.54% (1.01)	36.98% (5.01)	44.11% (2.90)	**49.51% (2.62)**
		Ride (2)	61.50% (2.99)	58.03% (0.92)	78.18% (4.89)	70.71% (1.61)	**78.60% (0.98)**
		Sports (2)	71.23% (5.03)	65.72% (2.12)	**87.84% (1.89)**	82.35% (2.15)	87.27% (0.40)
		Science (4)	64.75% (2.84)	37.82% (2.51)	69.45% (1.86)	75.32% (1.83)	**79.03% (0.64)**
		Religion (3)	48.55% (3.40)	42.76% (2.14)	51.62% (1.69)	52.25% (1.54)	**56.36% (0.33)**
		Politics (3)	60.68% (3.71)	42.30% (2.06)	62.98% (4.65)	66.84% (1.55)	**70.19% (1.11)**
		All (20)	37.54% (2.40)	9.22% (0.55)	35.69% (2.22)	38.48% (2.59)	**41.31% (3.47)**
NYT	10	Arts (4)	47.90% (6.73)	44.57% (2.55)	78.76% (5.07)	78.47% (7.63)	**94.86% (0.55)**
		Business (4)	71.72% (3.80)	56.67% (1.21)	76.87% (4.08)	60.30% (6.05)	**77.88% (1.29)**
		Politics (9)	60.50% (4.47)	43.18% (3.47)	69.35% (3.06)	53.23% (5.72)	**73.03% (2.60)**
		Science (2)	72.63% (14.66)	**94.74% (0.00)**	83.16% (8.42)	92.63% (4.21)	**94.74% (4.71)**
		Sports (7)	**92.56% (4.09)**	32.72% (2.26)	79.26% (3.19)	69.18% (5.95)	71.69% (3.97)
	20	Arts (4)	65.52% (3.58)	52.95% (2.60)	81.71% (3.52)	84.95% (2.96)	**96.19% (0.30)**
		Business (4)	72.63% (5.04)	63.33% (3.56)	80.30% (2.12)	73.23% (2.69)	**80.60% (1.48)**
		Politics (9)	72.94% (3.72)	54.93% (2.84)	75.42% (1.36)	63.48% (4.09)	**81.09% (2.46)**
		Science (2)	70.53% (5.37)	90.52% (2.11)	76.84% (8.55)	80.00% (2.10)	**93.68% (3.94)**
		Sports (7)	**97.29% (0.21)**	44.32% (3.85)	89.46% (0.35)	84.07% (6.77)	92.07% (2.06)
		All (26)	**85.43% (2.86)**	34.15% (3.07)	75.24% (1.31)	62.27% (6.48)	67.29% (0.47)

Table 13: Mean and standard deviation of accuracy score of a non-linear random forest classifier.

Dataset	Topics	Class	LDA	GLDA	CTM	SenClu	SLDA (Ours)
20News	10	Computer (5)	**44.04% (3.17)**	16.73% (3.33)	34.06% (4.36)	43.42% (4.54)	39.87% (1.00)
		Ride (2)	63.70% (5.22)	41.98% (1.79)	75.34% (4.92)	73.11% (3.51)	**82.16% (0.85)**
		Sports (2)	76.55% (7.14)	59.98% (2.40)	84.26% (3.14)	77.05% (14.26)	**88.67% (1.54)**
		Science (4)	64.59% (2.44)	28.57% (2.65)	64.38% (3.05)	76.07% (1.98)	**78.43% (0.80)**
		Religion (3)	36.55% (2.91)	19.54% (0.34)	35.62% (3.10)	46.79% (3.78)	**49.04% (1.43)**
		Politics (3)	55.77% (5.22)	24.37% (2.12)	57.69% (3.23)	60.53% (4.93)	**68.22% (0.57)**
	20	Computer (5)	42.44% (2.08)	21.86% (1.28)	34.13% (5.07)	47.06% (3.88)	**52.00% (2.16)**
		Ride (2)	64.00% (2.82)	51.32% (1.69)	78.37% (3.92)	74.23% (2.51)	**80.65% (1.04)**
		Sports (2)	72.38% (5.29)	61.41% (2.91)	87.37% (2.30)	83.82% (2.64)	**88.42% (0.55)**
		Science (4)	66.27% (2.32)	33.51% (3.63)	69.34% (2.72)	76.42% (1.86)	**78.65% (0.75)**
		Religion (3)	33.36% (1.93)	19.35% (0.00)	40.38% (1.77)	48.47% (2.94)	**52.34% (2.43)**
		Politics (3)	59.94% (4.11)	27.93% (2.93)	54.59% (6.89)	67.32% (2.26)	**70.42% (1.08)**
		All (20)	34.56% (2.47)	4.72% (0.41)	30.90% (1.66)	36.08% (2.40)	**41.19% (3.87)**
NYT	10	Arts (4)	45.41% (6.43)	14.16% (0.00)	59.53% (7.29)	65.17% (12.70)	**91.87% (0.84)**
		Business (4)	63.97% (8.28)	15.98% (0.00)	66.58% (8.53)	52.30% (11.05)	**69.99% (4.20)**
		Politics (9)	33.87% (3.06)	6.55% (0.00)	38.45% (6.59)	36.78% (6.73)	**43.31% (0.55)**
		Science (2)	84.47% (6.79)	41.22% (5.50)	76.86% (9.09)	90.30% (2.19)	**91.39% (2.68)**
		Sports (7)	**90.07% (5.89)**	5.85% (0.00)	68.93% (5.40)	62.84% (9.07)	67.39% (7.20)
	20	Arts (4)	51.36% (9.10)	14.16% (0.00)	63.30% (11.58)	79.73% (8.60)	**94.84% (0.89)**
		Business (4)	62.84% (11.42)	15.98% (0.00)	72.64% (6.57)	72.55% (7.92)	**76.75% (1.22)**
		Politics (9)	44.78% (2.34)	6.55% (0.00)	38.58% (2.44)	53.33% (3.61)	**57.81% (4.36)**
		Science (2)	76.31% (13.15)	38.97% (5.50)	59.72% (8.56)	78.60% (2.44)	**89.15% (3.54)**
		Sports (7)	**96.67% (0.31)**	6.16% (0.42)	84.75% (2.10)	83.18% (8.42)	89.51% (1.81)
		All (26)	**42.16% (1.08)**	1.59% (0.12)	33.04% (1.25)	29.66% (4.20)	26.25% (1.61)

Table 14: Mean and standard deviation of macro-F1 score of linear logistic classification with a topic distribution.

Dataset	Class	Category	Doc
20News	Computer	comp.graphics / comp.os.ms-windows.misc / comp.windows.x / comp.sys.ibm.pc.hardware / comp.sys.mac.hardware	4,776
	Ride	rec.autos / rec.motorcycles	1,905
	Sports	rec.sport.baseball / rec.sport.hockey	1,933
	Science	sci.crypt / sci.electronics / sci.med / sci.space	3,835
	Religion	alt.atheism / soc.religion.christian / talk.religion.misc	2,360
	Politics	talk.politics.guns / talk.politics.mideast / talk.politics.misc	2,559
NYT	Arts	dance / music / movies / television	1,043
	Business	economy / energy companies / international business / stocks and bonds	983
	Politics	abortion / federal budget / gay rights / gun control / immigration / law enforcement / military / surveillance / the affordable care act	989
	Science	cosmos / environment	90
	Sports	baseball / basketball / football / golf / hockey / soccer / tennis	8,639

Table 15: Dataset structure. Class is a coarse-grained class where topic models are trained for each, and Category is the fine-grained category that a logistic regression model should predict with a topic distribution. Doc is the number of documents included in the Class.

Model	Extracted Topics
SLDA	1. With the score tied at two goals apiece, Henrik Sedin and the Canucks scored a goal in the third period to take a 4 2 lead over the Rangers. 2. The Reds scored three runs in the top of the 11th inning against the Giants, but the Giants won their second straight game, 5 4, with a walk off single by Buster Posey. 3. I mean, you have a chance to win now with ... 4. With the Giants, he batted.286 with 11 hits, 3 home runs, and 11 RBIs. 5. b c. 6. On the par 5 18th hole, Woods made a birdie on his first shot, but then had to settle for a 72 putting stroke, tying with Paul Oeschger for third place at the event. 7. Against the Tennessee Titans, the rookie quarterback led the way again, completing 11 of his 19 passes for 113 yards and a touchdown, but the Giants were forced to punt after losing three yards on a rush by defensive end Jared Crick. 8. With a game high 23 points, Marcus Paige led the team with 9 for 9 three point shooting. 9. With two outs, Chase Utley singled to lead off the inning and scored on a throwing error by Jorge Posada. 10. He struck out six batters in the first inning, but allowed a run in the third after two outs, ending a streak of nine consecutive scoreless innings.
LDA	1. said, players, would, one, years, n, team, new, like, fans 2. league, cup, club, world, team, season, last, said, united, champions 3. said, yards, game, season, first, two, quarterback, coach, last, touchdown 4. game, said, nets, points, knicks, first, team, season, games, one 5. said, yankees, season, would, last, game, team, hes, going, get 6. game, goal, rangers, first, games, said, goals, two, scored, period 7. first, two, hit, game, runs, innings, inning, three, hits, run 8. said, open, tour, golf, woods, round, first, two, one, last 9. points, state, game, first, scored, lead, half, c, second, big 10. open, said, first, match, set, nadal, wimbledon, williams, tennis, grand

Table 16: Comparison table between SentenceLDA and LDA on 10 Topics-**NYT**-Sports.

Dataset	Topics	Class	QA	Distill	SLDA (Ours)
20News	20	Computer (5)	42.39%	**55.36%**	53.53%
		Ride (2)	79.21%	76.32%	**81.05%**
		Sports (2)	83.70%	**89.00%**	**89.00%**
		Science (4)	71.63%	78.92%	**79.65%**
		Religion (3)	55.64%	58.83%	**59.57%**
		Politics (3)	66.11%	70.89%	**72.44%**
		All (20)	22.27%	**47.32%**	40.49%
NYT	20	Arts (4)	88.10%	**95.71%**	94.76%
		Business (4)	**79.29%**	76.26%	78.28%
		Politics (9)	64.68%	**79.10%**	73.63%
		Science (2)	**89.47%**	**89.47%**	84.21%
		Sports (7)	68.50%	87.46%	**92.14%**
		All (26)	53.94%	65.14%	**66.03%**

Table 17: Accuracy score for each sentence embedding model.

Topic (BERTScore)	Extracted Topic Sentences
Topic 1 (0.9653)	1. This is a 2. This is a 3. This is
Topic 2 (0.8581)	1. With the Penguins leading the series 3 0, the Canadiens had two players Guy Lafleur and Paul Stastny with at least a goal and an assist. 2. With the Penguins leading the series 3 2, the Maple Leafs recalled defenceman Bobby Orr and forward Bernie Federko from the minors the two were the only players on the team to have played in all 82 games. 3. With the playoffs starting, the Canadiens had a record of 22 3 2 with Sutter, the only NHL player on the roster, leading the league in plus minus at 4.
Topic 3 (0.8457)	1. See also the Contact Information. 2. See also the Internet resource at www.cern.caltech.edu 877 895 6456. 3. See also the Internet Archive s Wayback Machine for the electronic version.
Topic 4 (0.8970)	1. The ATI Mobility Radeon HD 5670 supports up to 256 MB of DDR2 ECC RAM. 2. The ATI Mobility Radeon HD 5350 and the ATI Mobility Radeon HD 5670 support up to 512 MB of DDR2 ECC RAM. 3. The Dell PowerEdge 2900 series supports both SDRAM and I O modules.
Topic 5 (0.8812)	1. It is not the case that if you have a gun law, you are going to protect people. 2. If you want to have a gun law, fine, but don t make it a crime to defend against it... 3. If you have a constitutional right, you have a problem with the people who are arming the people.
Topic 6 (0.9071)	1. The Armenians in Palestine are not the victims of some crazy Arabs who want to annihilate the state... 2. The Armenians in Palestine are not only victims of Turkish policy... 3. If the Armenians in the territories occupied by Turkey are not killed, then they are terrorists...
Topic 7 (0.8745)	1. Jesus himself is the fulfillment of this teaching see below . 2. Jesus is not to be understood in the sense of the Bible... 3. Jesus is not to be understood as the Son of God, but as the Savior of mankind.
Topic 8 (0.8953)	1. It is not the case that if I am not a Christian, then I am not capable of understanding the truth of God. 2. If you believe in God, you re going to have a very strong objection to those who don t. 3. It is not the case that if I do not believe in God, I am not trying to prove that I do not believe in him.
Topic 9 (0.8641)	1. I m sorry, but I don t think you re getting a yes from me. 2. You re kidding me. 3. I m not saying, OK, you can t do this anymore.
Topic 10 (0.8952)	1. A version of the program for Microsoft Windows is included with Xgrid. 2. A version of X Window System is available as a free download from the project s website. 3. A free and open source version of Xgrid is available for Microsoft Windows, Mac OS X and Linux.

Table 18: Generated topic sentences from SentenceLDA and corresponding BERTScore to check the topic coherence.