Cross-lingual Contextualized Topic Models with Zero-shot Learning

Federico Bianchi
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
f.bianchi@unibocconi.it

Silvia Terragni
University of Milano-Bicocca
Viale Sarca 336, 20126
Milan, Italy
s.terragni4@campus.unimib.it

Dirk Hovy
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
dirk.hovy@unibocconi.it

Debora Nozza
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
debora.nozza@unibocconi.it

Elisabetta Fersini
University of Milano-Bicocca
Viale Sarca 336, 20126
Milan, Italy
elisabetta.fersini@unimib.it

Abstract

Many data sets (e.g., reviews, forums, news, etc.) exist parallelly in multiple languages. They all cover the same content, but the linguistic differences make it impossible to use traditional, bag-of-word-based topic models. Models have to be either single-language or suffer from a huge, but extremely sparse vocabulary. Both issues can be addressed by transfer learning. In this paper, we introduce a zero-shot cross-lingual topic model. Our model learns topics on one language (here, English), and predicts them for unseen documents in different languages (here, Italian, French, German, and Portuguese). We evaluate the quality of the topic predictions for the same document in different languages. Our results show that the transferred topics are coherent and stable across languages, which suggests exciting future research directions.

1 Introduction

Topic models (Blei et al., 2003; Blei, 2012) allow us to find the main themes and overarching tropes in textual data. However, traditional methods are language-specific and cannot be used in a *transferable manner*. They rely on a fixed vocabulary specific to the training language.

Therefore, currently available topic models suffer from two limitations: (i) they cannot handle unknown words by default, and (ii) they cannot easily be applied to other languages - except the one in the training data - since the vocabulary would not match. Training on several languages together, though, results in a vocabulary so vast that it creates problems with parameter size, search, and overfitting (Boyd-Graber et al., 2014). Traditional topic modeling provides methods to extract meaningful word distributions from "unstructured" text but requires language-specific bag-of-words (BoW) representations (Boyd-Graber and Blei, 2009; Jagarlamudi and Daumé, 2010).

A cross-lingual setup proves ideal for *transfer learning*: provided that the gist of topics is the same across languages, we can learn this gist on texts in one language and then apply it to others. This setup is *zero-shot learning*: we train a model on one language and test it on several other languages to which the model had no access during training.

To this end, we need to leverage external information to support the topic modeling task. Indeed, topic models have often gained significant advantages from introducing external knowledge, e.g., document relationships (Yang et al., 2015; Wang et al., 2020; Terragni et al., 2020a,b) and word embeddings (Nozza et al., 2016; Li et al., 2016; Zhao et al., 2017; Dieng et al., 2020). Recently, pre-trained contextualized embeddings, e.g., BERT (Devlin et al., 2019) embeddings, have enabled exciting new results in several NLP tasks (Rogers et al., 2020; Nozza et al., 2020). More importantly, there do exist contextualized embeddings that are also multilingual.

This paper introduces a novel neural topic modeling architecture in which we *replace* the input BoW document representations with multilingual *contextualized* embeddings. Neural topic models take in input the document BoW representations, which provide valuable symbolic information; however, this information's structure is lost after the first hidden layer in any neural architecture. We, therefore, hypothesize that contextual information can replace the BoW representation.

We use a neural encoding layer for the pre-trained document representations from a contextu-

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1676–1683
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

alized embedding model input (e.g., BERT) before the neural topic model's sampling process. This change allows us to address the two limitations mentioned above jointly: (i) our approach solves the problem of dealing with unseen words at test time since we do not need them to have a BoW representation; moreover, (ii) the model infers topics on unseen documents in languages other than the one in the training data. The inferred topics consist of tokens from the training language and can be applied to any supported test language. We show the high quality of the resulting topics for four test languages both quantitatively and qualitatively.

To the best of our knowledge, there is no prior work on zero-shot cross-lingual topic modeling. Our model can be applied to new languages after training is complete and does not require external resources, alignment, or other conditions. Nonetheless, the flexibility of the input means our model will benefit from any future improvement of language modeling techniques.

Contributions We release a novel neural topic model that relies on language-independent representations to generate topic distributions. We show that this input can replace the standard input BoW without loss of quality. We show that its multilingual representations enable zero-shot cross-lingual tasks. The solution we propose is straightforward and does not require high computational resources since it can efficiently run on common laptops (see Appendix). We have implemented the tool as a documented python package available at `https://github.com/MilaNLProc/contextualized-topic-models`.

2 Contextualized Neural Topic Models

We extend Neural-ProdLDA (Srivastava and Sutton, 2017), one of the most recent and promising approaches of neural topic modeling, based on the Variational AutoEncoder (VAE) (Kingma and Welling, 2014). The neural variational framework trains an inference network, i.e., a neural network that directly maps the BoW representation of a document onto a continuous latent representation. A decoder network then reconstructs the BoW by generating its words from the latent document representation. This latent representation is sampled from a Gaussian distribution parameterized by μ and σ^2 that are part of the variational inference framework (Kingma and Welling, 2014) — see (Srivastava and Sutton, 2017) for more details.

We replace the input BoW in Neural-ProdLDA with pre-trained multilingual representations from SBERT (Reimers and Gurevych, 2019), a recent and effective model for contextualized representations. In Figure 1, we sketch the architecture of our contextualized neural topic model. The final *reconstructed BoW* layer is still a component of our model: the BoW representation is necessary for the model's training to obtain the topic indicators (i.e., the most likely words representing a topic), but it becomes useless during testing.

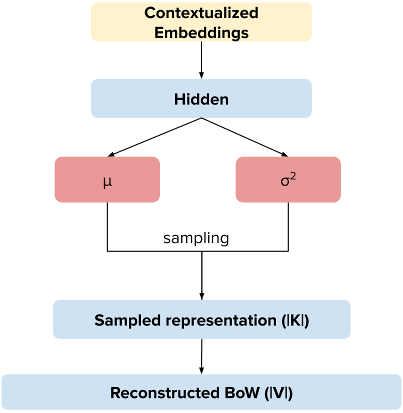

Figure 1: High-level schema of the architecture for the proposed contextualized neural topic model.

Our proposed model, **Zero-Shot Topic Model** (ZeroShotTM), is trained with input document representations that account for word-order and contextual information, overcoming one of the central limitations of BoW models. Moreover, the use of language-independent document representations allows us to do zero-shot topic modeling for unseen languages. This property is essential in low-resource settings in which there is little data available for the new languages. Because multilingual contextualized representations exist for multiple languages, it allows zero-shot modeling in a *cross-lingual* scenario. Indeed, ZeroShotTM is language-independent: given a contextualized representation of a new language as input,[1] it can predict the topic distribution of the document. The predicted topic descriptors, though, will be from the training language. Let us also notice that our method is agnostic about the choice of the neural topic model architecture (here, Neural-ProdLDA), as long as it extends a Variational Autoencoder.

[1] As long as a multilingual model - like multilingual BERT - covers it.

3 Experiments

Our experiments evaluate two main hypotheses: (i) *we can define a topic model that does not rely on the BoW input but instead uses contextual information;* (ii) *the model can tackle zero-shot cross-lingual topic modeling*. The Appendix contains more details about the experiments (e.g., code, data, runtime, replication details).

Datasets We use datasets collected from English Wikipedia abstracts from DBpedia.[2] The first dataset (W1) contains 20,000 randomly sampled abstracts. The second dataset (W2) contains 100,000 English documents. We use 99,700 documents as training and consider the remaining 300 documents as the test set. We collect the 300 respective instances in Portuguese, Italian, French, and German. This collection creates a test set of comparable documents, i.e., documents that refer to the same entity in Wikipedia, but in different languages.

We extract only the first 200 tokens of each abstract to reduce the length limit's effects in the tokenization process. In particular, we use the efficient and effective SBERT (Reimers and Gurevych, 2019),[3] using the multilingual model,[4] on this unpreprocessed text. We then remove stopwords and use the most frequent remaining 2,000 words to create the English vocabulary for BoW model comparisons.

3.1 To Contextualize or Not To Contextualize

First, we want to check if **ZeroShotTM** maintains comparable performance to other topic models; if this is true, we can then explore its performance in a cross-lingual setting. Since we use only English text, in this setting we use English representations.[5]

Model	τ (50)	τ (100)
ZeroShotTM	0.1632	0.1381
Combined TM	0.1644	**0.1409***
Neural-ProdLDA	**0.1658**	0.1285
LDA	-0.0246	-0.0757

Table 1: NPMI Coherences on W1 data set. * denotes the statistically significant results (t-test).

We compare **ZeroShotTM** on W1 with: (i) Combined TM (Bianchi et al., 2020), an extension of Neural-ProdLDA that concatenates *both* BoWs and SBERT representations (transformed to the same dimension of the BoWs) as inputs to the model, (ii) Neural-ProdLDA (Srivastava and Sutton, 2017), and (iii) LDA (Blei et al., 2003).

We compute the topic coherence (Lau et al., 2014) via NPMI (τ) for 50 and 100 topics averaging models' results over 30 runs. We report the results in Table 1. ZeroShotTM obtains comparable results to Combined TM and Neural-ProdLDA in this setting. Contextualized embeddings *can* replace BoW input representations without loss of coherence.

3.2 Zero-shot Cross-Lingual Topic Modeling

ZeroShotTM can be used for zero-shot cross-lingual topic modeling. We evaluate multilingual topic predictions on the multilingual abstracts in W2. We use SBERT [6] to generate multilingual embeddings as the input of the model.

3.2.1 Quantitative Evaluation

Since the predicted document-topic distribution is subject to a stochastic sampling process, we average it over 100 samples to obtain a better estimate.

Metrics We expect the topic distributions over a set of comparable documents (e.g., in English and Portuguese) to be similar to each other. We compare the topic distributions of each abstract in a test language with the topic distribution of the respective abstract in English, which is the training language. Note that the English test document is also unseen, i.e., the training data does not include it. We evaluate our model on three different metrics. The first metric is **matches**, i.e., the percentage of times the predicted topic for the non-English test document is the same as for the respective test document in English. The higher the scores, the better.

To also account for similar but not exactly equal topic predictions, we compute the **centroid embeddings** of the five words describing the predicted topic for both English and non-English documents. Then we compute the cosine similarity between those two centroids (CD).

Finally, to capture the **distributional similarity**, we also compute the KL divergence between the

[2] https://wiki.dbpedia.org/downloads-2016-10

[3] https://github.com/UKPLab/sentence-transformers

[4] We use the *distiluse-base-multilingual-cased* embeddings for this experiment available on the authors' repository.

[5] We use the *bert-base-nli-mean-tokens* model.

[6] https://github.com/UKPLab/sentence-transformers

Lang	Mat25↑	KL25↓	CD25↑	Mat50↑	KL50↓	CD50↑
IT	75.67	0.16	0.84	62.00	0.21	0.75
FR	79.00	0.14	0.86	63.33	0.19	0.77
PT	78.00	0.14	0.85	68.00	0.19	0.79
DE	79.33	0.15	0.85	64.33	0.20	0.77
ZeroShotTM Avg	**78.00**	**0.15**	**0.85**	64.41	0.20	0.77
Ori Avg	76.00	**0.15**	0.84	**69.00**	**0.19**	**0.79**
Uni	4.00	0.75	—	2.00	0.85	—

Table 2: Match, KL, and centroid similarity for 25 and 50 topics on various languages on W2.

predicted topic distribution on the test document and the same test document in English. Here, lower scores are better, indicating that the distributions do not differ by much.

Automatic Evaluation We use two baselines: the first one (Ori) consists of performing topic modeling on documents translated into English via DeepL.[7] While this is an easily accessible baseline, automatic translation is costly and may introduce bias in the representations (as shown by Hovy et al. (2020)). We compare the predicted topics of each translated document to the ones predicted for the original English document (as done above). The second baseline is a uniform distribution (Uni): we compute all the metrics over a uniform distribution (this baseline gives a lower bound).

Table 2 shows the evaluation results of our model in the zero-shot context. Note that because we trained on English data, the topic descriptors are in English. Topic predictions are significantly better than the uniform baselines: more than 70% of the times, the predicted topic on the test set matches the topic of the same document in English. The CD similarity suggests that even when there is no match, the predicted topic on the unseen language is at least similar to the one on the English testing data. Simultaneously, the predictions for the contextualized model are in line with the ones obtained using the translations (Ori Avg), showing that our model is capable of finding good topics for documents in unseen languages without the need for translation.

Manual Evaluation We rated the predicted topics for 300 test documents in five languages (thus, 1500 docs including English) on an ordinal scale from 0-3. A 0 rate means that the predicted topic is wrong, a 1 rate means the topic is somewhat related, a 2 rate means the topic is good, and a 3 rate means the topic is entirely associated with the considered document. Table 3 shows the results per language. We evaluate the inter-rater reliability using Gwet AC1 with ordinal weighting (Gwet, 2014). The resulting value of 0.88 indicates consistent scoring.

Language	Average Topic Quality
English	2.35
Italian	2.29
French	2.22
Portuguese	2.26
German	2.19
Average	2.26

Table 3: Average topic quality (out of 3).

3.2.2 Qualitative Evaluation

In Table 4, we show some examples of topic predictions on test languages. Our model predicts the main topic for all languages, even though they were unseen during training.

The predicted topic is generally consistent with the text. I.e., the topics are easily interpretable and give the user a coherent impression. In some circumstances, noise biases the results: dates in the abstract tend to make the model predict a topic about time. Another interesting case is the abstract of the artist Joan Brossa, who was both a poet and a graphic designer. In the English and Italian abstract, the model has discovered a topic related to writing. In constrast, in the Portuguese abstract, the model has found a topic related to art, which is still meaningful.

[7] https://www.deepl.com/

Lang	Sentence	Predicted Topic
EN	Blackmore's Night is a British/American traditional folk rock duo [...]	rock, band, bass, formed
IT	I Blackmore's Night sono la band fondatrice del renaissance rock [...]	rock, band, bass, formed
PT	Blackmore's Night é uma banda de folk rock de estilo renascentista [...]	rock, band, bass, formed
EN	Langton's ant is a two-dimensional Turing machine with [...]	mathematics, theory, space, numbers
FR	On nomme fourmi de Langton un automate cellulaire [...]	mathematics, theory, space, numbers
DE	Die Ameise ist eine Turingmaschine mit einem zweidimensionalen [...]	mathematics, theory, space, numbers
EN	The Journal of Organic Chemistry, colloquially known as JOC or [...]	journal, published, articles, editor
IT	Journal of Organic Chemistry è una rivista accademica [...]	journal, published, articles, editor
PT	Journal of Organic Chemistry é uma publicação científica [...]	journal, published, articles, editor
EN	Joan Brossa [...] was a Catalan poet, playwright, graphic designer [...]	book, french, novel, written
IT	Fu l'ispiratore e uno dei fondatori della rivista "Dau al Set"[...]	book, french, novel, written
PT	Joan Brossa i Cuervo [...] foi um poeta, dramaturgo, artista plástico [...]	painting, art, painter, works

Table 4: Examples of zero-shot cross-lingual topic classification in various languages with ZeroShotTM.

4 Related Work

While not in a zero-shot fashion, several researchers have studied multilingual and cross-lingual topic modeling (Ma and Nasukawa, 2017; Gutiérrez et al., 2016; Hao and Paul, 2018; Heyman et al., 2016; Liu et al., 2015; Krstovski et al., 2016).

The first model proposed to process multilingual corpora with LDA is the Polylingual Topic Model by Mimno et al. (2009). It uses LDA to extract language-consistent topics from parallel multilingual corpora, assuming that translations share the same topic distributions. Models that transfer knowledge on the document level have many variants, including (Hao and Paul, 2018; Heyman et al., 2016; Liu et al., 2015; Krstovski et al., 2016). However, existing models require to be trained on multilingual corpora and are always language-dependent: they cannot predict the main topics of a document in an unseen language.

Other models use multilingual dictionaries (Boyd-Graber and Blei, 2009; Jagarlamudi and Daumé, 2010), requiring some predefined mapping. Embeddings, both for words and documents, have been shown to capture a wide range of semantic, syntactic, and social aspects of language (Hovy and Purschke, 2018; Rogers et al., 2020). Our work adds language-independent topics to that list.

5 Conclusions

We propose a novel neural architecture for cross-lingual topic modeling using contextualized document embeddings as input. Our results show that (i) contextualized embeddings can replace the input BoW representations and (ii) using contextualized representations allows us to tackle zero-shot cross-lingual topic modeling. The resulting model can be trained on any one language and applied to any other language for which embeddings are available.

Acknowledgements

Federico Bianchi, Dirk Hovy, and Debora Nozza are members of the Bocconi Institute for Data Science and Analytics (BIDSA) and the Data and Marketing Insights (DMI) unit.

References

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jordan L. Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*, pages 75–82. AUAI Press.

Jordan L. Boyd-Graber, David Mimno, and David Newman. 2014. *Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements*, CRC Handbooks of Modern Statistical Methods. CRC Press.

Stephen Carrow. 2018. Pytorchavitm: Open source avitm implementation in pytorch. Github.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2019, Volume 1 (Long and Short Papers), pages 4171–4186.

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.

E.D. Gutiérrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, and Luca Gilardi. 2016. Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics*, 4:47–60.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Shudong Hao and Michael J. Paul. 2018. Learning multilingual topics from incomparable corpora. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, pages 2595–2609. Association for Computational Linguistics.

Geert Heyman, Ivan Vulic, and Marie-Francine Moens. 2016. C-BiLDA extracting cross-lingual topics from non-parallel texts by distinguishing shared from unshared content. *Data Mining and Knowledge Discovery*, 30(5):1299–1323.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "You sound just like your father" Commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

Jagadeesh Jagarlamudi and Hal Daumé. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Advances in Information Retrieval*, pages 444–456, Berlin, Heidelberg. Springer Berlin Heidelberg.

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

Kriste Krstovski, David A. Smith, and Michael J. Kurtz. 2016. Online multilingual topic models with multi-level hyperpriors. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016*, pages 454–459. Association for Computational Linguistics.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 530–539.

Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 165–174.

Xiaodong Liu, Kevin Duh, and Yuji Matsumoto. 2015. Multilingual topic models for bilingual dictionary extraction. *ACM Transactions on Asian Language Information Processing*, 14(3):11:1–11:22.

Tengfei Ma and Tetsuya Nasukawa. 2017. Inverted bilingual topic models for lexicon extraction from non-parallel data. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 4075–4081. IJCAI.org.

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*, pages 880–889. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [MASK]? Making Sense of Language-Specific BERT Models. *arXiv preprint arXiv:2003.02912*.

Debora Nozza, Elisabetta Fersini, and Enza Messina. 2016. Unsupervised irony detection: a probabilistic model with word embeddings. In *International Conference on Knowledge Discovery and Information Retrieval*, volume 2, pages 68–76. SCITEPRESS.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3973–3983. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *arXiv preprint arXiv:2002.12327*.

Akash Srivastava and Charles A. Sutton. 2017. Autoencoding variational inference for topic models. In *In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net.

Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2020a. Constrained relational topic models. *Information Sciences*, 512:581–594.

Silvia Terragni, Debora Nozza, Elisabetta Fersini, and Messina Enza. 2020b. Which matters most? comparing the impact of concept and document relationships in topic models. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 32–40, Online. Association for Computational Linguistics.

Chaojie Wang, Hao Zhang, Bo Chen, Dongsheng Wang, Zhengjue Wang, and Mingyuan Zhou. 2020. Deep relational topic modeling via graph poisson gamma belief network. *Advances in Neural Information Processing Systems*, 33.

Weiwei Yang, Jordan Boyd-Graber, and Philip Resnik. 2015. Birds of a feather linked together: A discriminative topic model using link-based priors. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 261–266.

He Zhao, Lan Du, Wray Buntine, and Gang Liu. 2017. Metalda: A topic model that efficiently incorporates meta information. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 635–644. IEEE.

A Datasets

We used the English DBpedia 2016-10 abstract dump[8] to create our datasets.

W1 We randomly sampled 20,000 documents from the English DBpedia abstract dump to create our first set of documents. We created W1 to provide a quick collection of documents to test if our Contextual TM performance does not decrease significantly.

W2 We collected 100,000 abstracts sampling randomly from those that had at least 200 chars. Given this set, we extracted 300 random English abstracts. Given the random abstracts, we retrieved the respective version in other languages using the DBpedia SPARQL endpoint.[9] We manually evaluated the quality of the 300 abstracts since we looked at each of those during our manual evaluation, finding no mismatch between the abstract and no corrupted text.

A.1 Preprocessing

We followed a standard pre-processing pipeline to generate the preprocessed set of documents. We removed punctuation, digits, and nltk's English stop-words.[10] Following other researchers, we selected 2,000 as the maximum number of words for the BoW, and thus we kept in the abstracts only the 2,000 most frequent words.

B Models and Baselines

B.1 Neural-ProdLDA

We use the implementation made available by Carrow (2018) since it is the most recent and with the most updated packages (e.g., one of the latest versions of PyTorch). The model is trained for 100 epochs. We use ADAM optimizer (with a learning rate equal to 2e-3). The inference network is composed of a single hidden layer and 100-dimension of softplus units. The priors over the topic and document distributions are learnable parameters. Momentum is set to 0.99, the learning rate is set to 0.002, and we apply 20% of drop-out to the hidden document representation. The batch size is equal to 200. More details related to the architecture can be found in the original work (Srivastava and Sutton, 2017).

B.2 ZeroShot TM

The model and the hyper-parameters are the same for Neural-ProdLDA, with the difference that we replace the BoW with SBERT features. The model is trained for 100 epochs. We use ADAM optimizer.

B.3 Combined TM

The model (Bianchi et al., 2020)[11] and the hyper-parameters are the same used for Neural-ProdLDA with the difference that we also use SBERT features in combination with the BoW: we take the SBERT embeddings, apply a (learnable) function/dense layer $R^{512} \rightarrow R^{|V|}$ and concatenate the representation to the BoW. The model is trained for 100 epochs. We use ADAM optimizer.

B.4 LDA

We use Gensim's[12] implementation of this model. The hyper-parameters $alpha$ and $beta$, controlling the document-topic and word-topic distribution respectively, are estimated from the data during training.

[8] https://wiki.dbpedia.org/downloads-2016-10

[9] https://dbpedia.org/sparql

[10] https://www.nltk.org/

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

[12] https://radimrehurek.com/gensim/models/ldamodel.html

C Computing Infrastructure

We ran experiments on two common laptops, equipped with a GeForce GTX 1050 (running CUDA 10). As our experiments show, the models can be easily run with basic hardware (having a GPU is better than just using CPU, but the experiments can also be replicated on CPU). Both laptops have 16GB of RAM.

C.1 Runtime

Our implementation is written in PyTorch and runs on both GPU and CPU. Table 5 shows the runtime for one epoch of both our Combined TM and Neural-ProdLDA for 25 and 50 topics on the GeForce GTX 1050. Neural-ProdLDA is slightly faster than our ZeroShotTM. This is due to the additional representation that cannot be encoded as a sparse matrix. However, we believe that these numbers are comparable and make our model easy to use even with common hardware.

Model	W1 (25)	W1 (50)
ZeroShot TM	1.2s	1.2s
Neural-ProdLDA	0.8s	0.9s

Table 5: Time to complete one epoch on the **W1** dataset with 25 and 50 topics.

D Source Code

D.1 Development

Our software is available as a Python package that a user can easily install.[13]

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