An Empirical Analysis of Topic Models: Uncovering the Relationships between Hyperparameters, Document Length and Performance Measures

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

Abstract

Neural Topic Models are recent neural models that aim at extracting the main themes from a collection of documents. The comparison of these models is usually limited because the hyperparameters are held fixed. In this paper, we present an empirical analysis and comparison of Neural Topic Models by finding the optimal hyperparameters of each model for four different performance measures adopting a single-objective Bayesian optimization. This allows us to determine the robustness of a topic model for several evaluation metrics. We also empirically show the effect of the length of the documents on different optimized metrics and discover which evaluation metrics are in conflict or agreement with each other.

1 Introduction

Topic models (Blei, 2012) are probabilistic generative models that aim at identifying the underlying themes, or topics, in a collection of documents. Although they are used in a vast range of applications, from text exploratory purposes to information retrieval tasks (Boyd-Graber et al., 2017), most of the investigations disregard the elements that influence the results generated by the models and, in particular, what is the effect on their performance.

Several works explore topic modeling over a range of different models, topics, and measures, but usually focus on classical topic models (Greene et al., 2014; Stevens et al., 2012), e.g. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) or Non-negative Matrix Factorization (NMF) (Lee and Seung, 2000), and solely on a single evaluation measure (Stevens et al., 2012; O'Callaghan et al., 2015).

Doan and Hoang recently made an effort to benchmark neural topic models, however, the authors seem to disregard the importance of the hyperparameter selection. In fact, the evaluations of topic models are usually limited to the comparison of models whose hyperparameters are fixed. Yet, the hyperparameters that control the models can have a great impact on their performance. Therefore, fixing them prevents the researchers from discovering the best topic model on a given dataset. In the latest years, Neural Topic Models (NTM) (Zhao et al., 2021; Dieng et al., 2020; Bianchi et al., 2021a,b) have gained popularity, due to their flexibility and scalability. The problem of finding the best hyperparameter configuration has become even more compelling, since topic models based on neural networks are usually controlled by a high number of hyperparameters.

In this paper, we perform an empirical analysis of recent NTMs by optimizing the hyperparameters of the models with respect to different metrics. We aim to investigate if there exists a potential relationship between hyperparameters, document length and performance measures, to finally understand under which conditions we can exploit at best the potentiality of each model. In particular, the following research questions have been addressed:

RQ1: To what extent are Neural Topic Models robust across different evaluation metrics?

RQ2: Does the document length affect the Neural Topic Models on different performance measures?

RQ3: Does the optimization of a model's hyperparameters for a given performance metric imply good performance on other measures?

To this purpose, we use Bayesian Optimization (BO) (Archetti and Candelieri, 2019), a well-known and efficient strategy for hyperparameter tuning, to determine the optimal hyperparameter settings for four different evaluation metrics of five state-of-the-art NTMs. The hyperparameter optimization allows us to guarantee a fair comparison

Proceedings of Recent Advances in Natural Language Processing, pages 1408–1416
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_157

between the models and investigate their behavior with different hyperparameter settings.

2 Methodology

In this work, we conduct an empirical comparison of different state-of-the-art NTMs. We adopt a single-objective Bayesian Optimization approach, using the comparative framework topic modeling OCTIS (Terragni et al., 2021a), to optimize the hyperparameters of five different topic models with respect to four different evaluation metrics. Each metric investigates a different aspect of a model. Bayesian Optimization (Archetti and Candelieri, 2019; Kandasamy et al., 2020) is a Sequential Model-Based Optimization (SMBO) strategy that allows us to optimize all the hyperparameters by treating the topic model as a black-box function. The model is in fact viewed just in terms of its input (the hyperparameters) and its output (the distribution of the topics over the vocabulary and the topic distribution for each document).

BO uses the model's configurations evaluated so far to approximate the value of the performance metrics with respect to the model's hyperparameters and then selects a new promising configuration to evaluate. The two key components are the *probabilistic surrogate model* aimed at approximating the performance metrics to optimize, and the *acquisition function* that uses the mean of the surrogate model and the confidence (i.e. its standard deviation) to select the next configuration.

Optimizing a model's hyperparameters not only allows us to investigate the robustness of a model over different evaluation metrics (RQ1), but we can also investigate the performance of the optimized evaluation metric on datasets with different features (RQ2) and the relationship between the optimized evaluation metric and the other metrics (RQ3).

2.1 Topic Models

In our investigation, we focus on the following recent state-of-the-art topic models based on a neural variational frameworks. We consider Neural LDA (Srivastava and Sutton, 2017, NeurLDA), Product-of-experts LDA (Srivastava and Sutton, 2017, ProdLDA), the Embedded Topic Models (Dieng et al., 2020, ETM), and finally we use a variant of the family of Contextualized Topic Models, namely the Zero-shot Contextualized Topic Model (Bianchi et al., 2021b, CTM).

All these neural models are based on the Variational Autoencoder (VAE) presented in Miao et al.. The neural variational framework trains an inference network to map the bag-of-words (BoW) document representation into a continuous latent representation. A decoder network reconstructs the BoW by generating its words from the document representation. This document representation is K-dimensional, where K is the number of topics.

NeurLDA and ProdLDA (Srivastava and Sutton, 2017) explicitly approximate the Dirichlet prior using Gaussian distributions, instead of using a Gaussian prior. In addition, ProdLDA replaces the word-topic distribution in LDA with a product of experts (Hinton, 2002).

CTM (Bianchi et al., 2021b) extends ProdLDA by replacing the BoW document representation of the input with the corresponding pre-trained contextualized representations of the documents. These representations derive from contextualized language models, e.g. BERT (Devlin et al., 2019).[1]

Concerning ETM (Dieng et al., 2020), words and topics are represented in the same embedding space. The word-topic distribution is proportional to the exponentiated inner product of the topic embedding and each word embedding. ETM can automatically learn the word embedding representations or use pre-trained word embeddings. Following the original paper, we will refer to the former version of the model as ETM, while the one that uses pre-trained word embeddings (PWE) will be referred to as ETM-PWE.

We also consider the well-known Latent Dirichlet Allocation (Blei et al., 2003, LDA) as a baseline. LDA is a probabilistic model that describes a corpus through K topics, seen as distributions of words over a vocabulary W. A document is assumed composed of a mixture of topics following a Dirichlet distribution, where a topic drawn from the mixture is assigned to each word of the documents.

2.2 Evaluation Metrics

We consider four evaluation metrics that investigate different aspects of a topic model.

F1 refers to the Micro-F1 measure, the weighted average of the F1 measure for each class. We train a linear support vector machine (SVM) that predicts the document's class using the topic distribution θ of each document (given by each topic model) as

[1] This model has been designed for addressing a task of cross-lingual topic modeling, however, it also outperforms several monolingual neural topic models.

its feature representation (Terragni et al., 2020a).

IRBO (Bianchi et al., 2021a; Terragni et al., 2021b) is a measure of topic diversity (0 for identical topics and 1 for completely different topics). It is based on the Ranked-Biased Overlap measure (Webber et al., 2010). Topics with common words at different rankings are penalized less than topics sharing the same words at the highest ranks.

NPMI (Lau et al., 2014) measures Normalized Pointwise Mutual Information of each pair of words (w_i, w_j) in the 10-top words of each topic. It is a topic coherence measure, that evaluates how much the words in a topic are related to each other.

KL–B (AlSumait et al., 2009; Terragni et al., 2020b) is the Kullback-Leibler distance of a topic to a "background" topic, a topic found equally probable in all the documents. Meaningful topics appear in a small subset of the data, thus higher values are preferred.

3 Experimental Setting

3.1 Datasets and Preprocessing

To analyze the impact of the length of the documents with respect to several models and performance measures, we consider two different datasets: 20Newsgroup[2] (20NG), where each document is characterized by a long text, and M10 (Lim and Buntine, 2014), which is composed of titles of scientific papers, and therefore it represents a case study of short texts.

We adopt a common preprocessing procedure[3]: punctuation removal, lemmatization, removal of English stop-words and unfrequent words, removal of documents with less than 3 words (for M10) or 5 words (for 20NG). The stopwords list is the one provided by MALLET[4]. Each dataset is split into training (70%), validation (15%) and test set (15%). Table 1 shows the datasets statistics.

Datasets	# docs	average # words	# unique words	# classes
20NG	16309	48	1612	20
M10	8355	6	1696	10

Table 1: Statistics of the preprocessed datasets.

[2]http://qwone.com/ jason/20Newsgroups/

[3]The preprocessed datasets are already provided by the OCTIS library: https://github.com/mind-Lab/octis.

[4]http://mallet.cs.umass.edu/

3.2 Bayesian Optimization and Model Settings

We optimize the models' hyperparameters using BO for each evaluation metric. We trained each model 30 times and considered the median as the evaluation of the function to be optimized. The initial configurations are randomly sampled via Latin Hypercube Sampling and equal to the number of the hyperparameters to optimize plus 2 (to provide enough configurations for the initial surrogate model). The total number of BO iterations is 30 for LDA and 120 for the other models. We use Random Forests as the surrogate model and the Upper Confidence Bound (UCB) as the acquisition function.

We report the models' hyperparameters and their corresponding ranges in Table 2.

Model	Hyperparameter	Range
LDA	α prior	$[10^{-4}, 10]$
	β prior	$[10^{-4}, 10]$
NeurLDA/ ProdLDA/ CTM	Dropout	[0, 1 - 10^{-6}]
	Learning rate	$[10^{-6}, 10^{-1}]$
	Momentum	[0, 1]
	Activation function	elu, leakyrelu, relu, rrelu, selu, sigmoid, softplus
	Optimizer	adadelta, adagrad, adam, rmsprop, sgd
	# Neurons	100, 200, . . ., 1000
	# Layers	1, 2, 3, 4, 5
	Learn priors	true, false
ETM/ ETM-PWE	Dropout	[0, 1 - 10^{-6}]
	Learning rate	$[10^{-6}, 10^{-1}]$
	Weight decay	$[10^{-6}, 10^{-1}]$
	Activation function	elu, leakyrelu, relu, rrelu, selu, softplus, tanh
	Optimizer	adadelta, adagrad, adam, asgd, rmsprop
	# Neurons	100, 200, . . ., 1000
ETM	Rho size	100, 200, 300

Table 2: Hyperparameters and ranges.

Regarding LDA, we optimize the hyperparameters α and β priors that the sparsity of the topics in the documents and sparsity of the words in the topic distributions respectively. These hyperparameters are set to range between 10^{-4} and 10 on a logarithmic scale.

The hyperparameters of the neural models are mainly related to the architecture of the network. For all the neural models, we optimize the *dropout*

	20NG				M10			
	F1Score*	**IRBO***	**NPMI***	**KL-B***	**F1Score***	**IRBO***	**NPMI***	**KL-B***
LDA	0.469	0.963	0.064	**2.299**	0.472	0.944	-0.089	**2.343**
NeurLDA	0.339	**1.000**	0.067	0.907	0.420	**1.000**	-0.131	0.904
ProdLDA	0.373	0.998	0.107	0.992	0.539	**1.000**	0.044	1.652
CTM	0.361	0.998	**0.118**	1.019	0.563	**1.000**	**0.055**	0.937
ETM	0.453	0.996	0.080	0.370	0.534	0.997	-0.028	0.532
ETM-PWE	**0.471**	0.986	0.089	0.424	**0.585**	0.997	-0.070	0.201

Table 3: Median of each performance metric (columns) for each single-objective optimization (rows).

(ranging between 0 and $1 - 10^{-6}$) and the *momentum* (ranging between 0 and 1). We optimize the *learning rate*, that is set to range between 10^{-4} and 10^{-1}, on a logarithm scale. We also consider different variants of *activation functions* and *optimizers*.

Regarding NeurLDA, ProdLDA, and CTM in particular, we optimize the *number of layers* (ranging from 1 to 5), and the *number of neurons* (ranging from 100 to 1000). For simplicity, each layer has the same number of neurons. Finally, we also consider the hyperparameter *learn priors* that controls if the priors are learnable parameters.

Following (Bianchi et al., 2021a), we use the contextualized document representations derived from SentenceBERT (Reimers and Gurevych, 2019). We use the pre-trained BERT model fine-tuned on the natural language inference (NLI) task.[5]

Considering ETM and ETM-PWE, in addition to the hyperparameters mentioned above, we only optimize the *number of neurons* (ranging from 100 to 1000). We follow the original implementation, for which the number of hidden layers is set to 1. For ETM-PWE, we use pre-trained word2vec word embeddings (Mikolov et al., 2013), trained on the Google News corpus (3 million 300-dimension English word vectors).

For the neural models, we set the batch size to 200 and we adopted an early stopping criterion for determining the convergence of each model. We set the remaining model parameters to their default values. To have a fair comparison, we set the number of topics to be discovered equals to the number of classes available in each dataset, i.e. 10 for M10 and 20 for 20NG. For running the experiments, we use the open-source library OCTIS (Terragni et al., 2021a), which already integrates the implementations of the considered models and metrics. It is available at the following link: https://github.com/mind-Lab/octis.

[5] https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens

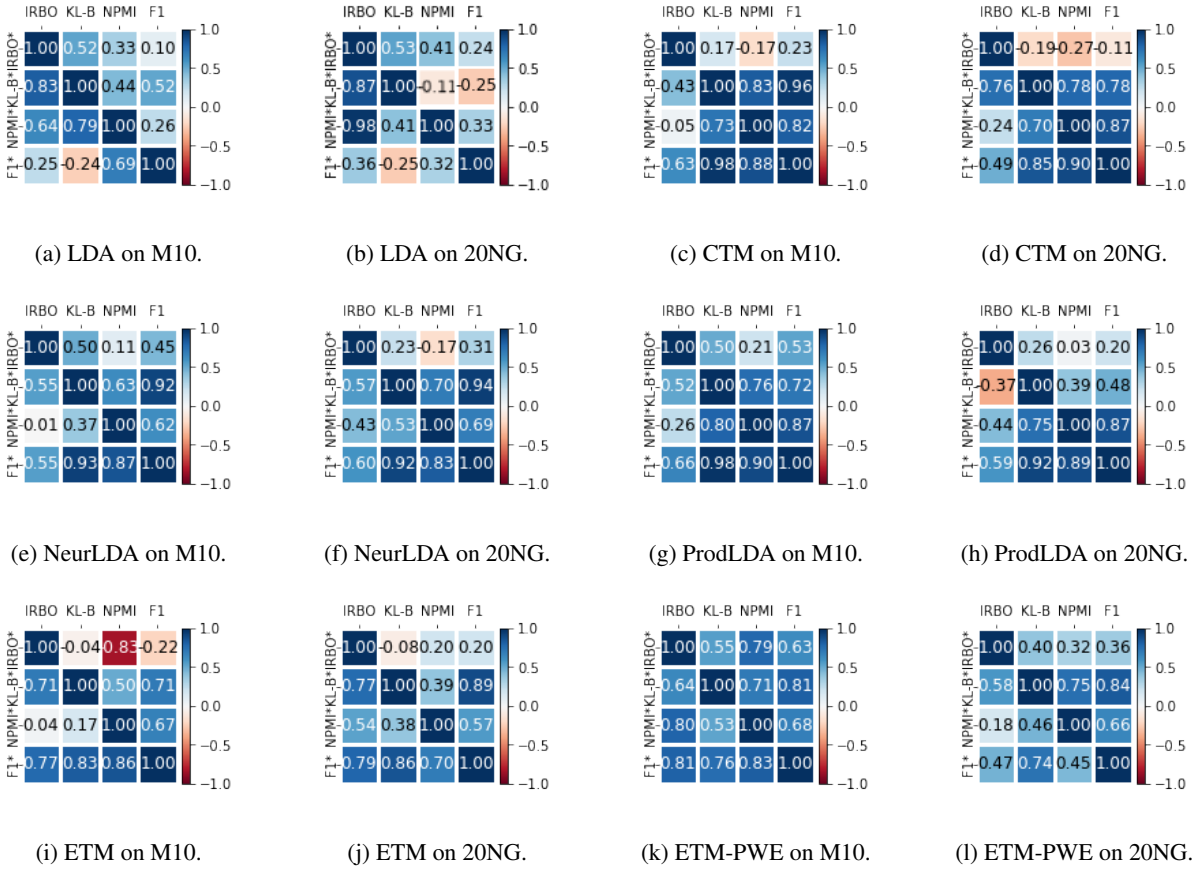
4 Empirical Analysis and Discussion

4.1 Robustness of Neural Topic Models (RQ1)

In table 3 we report the median of the four evaluation metrics for each topic model obtained by the best hyperparameter configuration. Rows represent the optimized metric (marked as *metric**), while columns denote the median of the evaluated metric. The overall best values for each metric and dataset are reported in bold. First of all, we can observe that there is not a model that outperforms the others for all the considered metrics. In fact, it seems that each topic model works better for a specific metric.

In particular, LDA is the topic model that obtains the best performance in terms of KL-B*, thus obtaining topics that are significant rather than background topics. While, the topic models based on the neural variational framework defined in (Srivastava and Sutton, 2017), i.e. NeurLDA, ProdLDA, and CTM, are the ones that obtain the highest diversity. Regarding the topic coherence, CTM obtains the best topic coherence for both datasets. In fact, it improves the performance of ProdLDA (second-best model for the topic coherence) through the incorporation of the contextualized pre-trained representations of the documents. Finally, ETM-PWE outperforms the other models in terms of F1*, probably due to the contribution of the pre-trained word embeddings.

Provided that each topic model seems to reach the best performance only in a specific metric, it follows that they cannot simultaneously guarantee optimal performance for the other metrics. We will further investigate the trade-off between different metrics in Sections 4.2 and 4.3. A complete overview of the best configuration of hyperparameters discovered by BO for all the models and for all the considered evaluation measures is reported in Tables 4, 5, 6, 7, 8 and 9. This would allow a user to choose a promising hyperparameter configuration for the evaluation metric of their interest.



(a) LDA on M10. (b) LDA on 20NG. (c) CTM on M10. (d) CTM on 20NG.

(e) NeurLDA on M10. (f) NeurLDA on 20NG. (g) ProdLDA on M10. (h) ProdLDA on 20NG.

(i) ETM on M10. (j) ETM on 20NG. (k) ETM-PWE on M10. (l) ETM-PWE on 20NG.

Figure 1: Metrics-metrics correlations.

4.2 Impact of the Document Length (RQ2)

We can derive other insights by analyzing Table 3 and comparing the two datasets. In particular, we highlight that for LDA the document length seems to be an invariant when optimizing on the KL-B* metric. This insight can be grasped by considering the KL-B* of LDA (i.e. 2.343 for M10 and 2.299 for 20NG) that, not only are the best performance when compared to the other models, but they suggest that LDA performs well independently on the document length and therefore it guarantees optimal KL-B* both on short and long documents.

Another important insight is about the F1 measures obtained by LDA (0.472 and 0.469), ETM (0.534 and 0.453), and ETM-PWE (0.585 and 0.471), which seem to be not affected by the length of the documents. On the other hand, the results for the F1 measure for NeurLDA, ProdLDA, and CTM (which are based on the same architecture) are affected by the documents' length, obtaining the best performance for short texts. In these cases, when the models achieve a high F1 on short documents (0.420 by NeurLDA, 0.539 by ProdLDA, and 0.563 by CTM), the performance on short documents is lower (0.339 by NeurLDA, 0.373 by ProdLDA, and 0.361 by CTM).

When optimizing for the IRBO* metric, all the models succeed in obtaining almost completely diverse topics, both for long and short texts. The performance of IRBO* for LDA* is slightly affected when dealing with short texts. Finally, we remark that CTM obtains an excellent topic coherence for both datasets, but, on the other hand, the remaining models seem to be particularly affected when dealing with short texts, assuming NPMI values inferior to 0.

4.3 Metrics-Metrics Correlations (RQ3)

In Figure 1, we report the correlations between the evaluation metrics when a single-objective optimization policy is performed. The rows of the correlation matrices denote the optimized metrics (F1*, IRBO*, KL-B*, and NPMI*), while the columns the non-optimized evaluated measures (F1, IRBO, KL-B, and NPMI). According to these results, we can observe if optimizing a model for a specific metric allows us for an increasing or

		α prior	β prior	Median
20NG	F1*	1.332	1.146	0.472
	IRBO*	0.325	0.004	0.954
	KL-B*	0.006	3.054	2.299
	NPMI*	0.658	0.520	0.066
M10	F1*	0.627	1.870	0.469
	IRBO*	0.349	9.403	0.939
	KL-B*	$2 \cdot 10^{-4}$	9.614	2.343
	NPMI*	0.005	1.531	-0.083

Table 4: Best configuration of hyperparameters discovered by BO for LDA for each evaluation measure.

		Activation	Dropout	Learn Priors	Learning Rate	Momentum	Num Layers	Num Neurons	Optimizer	Median
20NG	F1*	sigmoid	0.0839	1	0.0097	0.789	1	800	adam	0.373
	IRBO*	sigmoid	0.0839	1	0.0097	0.789	1	800	adam	0.998
	KL-B*	sigmoid	0.9481	1	0.0039	0.984	1	1000	sgd	0.992
	NPMI*	selu	0.0381	0	0.0208	0.949	3	600	adam	0.107
M10	F1*	elu	0.0025	1	0.0611	0.742	5	1000	adam	0.539
	IRBO*	sigmoid	0.0839	1	0.0097	0.789	1	800	adam	1.000
	KL-B*	rrelu	0.0198	1	0.0089	0.512	5	100	adam	1.652
	NPMI*	softplus	0.1664	0	0.0006	0.374	1	400	sgd	0.044

Table 5: Best configuration of hyperparameters discovered by BO for ProdLDA for each evaluation measure.

		Activation	Dropout	Learn Priors	Learning Rate	Momentum	Num Layers	Num Neurons	Optimizer	Median
20NG	F1*	sigmoid	0.084	0	0.0314	0.575	1	1000	adam	0.339
	IRBO*	sigmoid	0.062	1	0.0273	0.667	1	400	adam	1.000
	KL-B*	elu	0.0003	0	0.0008	0.891	3	700	adam	0.907
	NPMI*	sigmoid	0.130	0	0.0075	0.797	1	800	rmsprop	0.067
M10	F1*	sigmoid	0.061	0	0.0129	0.756	1	800	rmsprop	0.420
	IRBO*	leakyrelu	0.125	0	0.0019	0.859	2	200	sgd	1.000
	KL-B*	selu	0.0003	0	0.0186	0.269	2	600	adam	0.904
	NPMI*	selu	0.087	1	0.0002	0.754	1	100	sgd	-0.132

Table 6: Best configuration of hyperparameters discovered by BO for NeurLDA for each evaluation measure.

		Activation	Dropout	Learn Priors	Learning Rate	Momentum	Num Layers	Num Neurons	Optimizer	Median
20NG	F1*	sigmoid	0.046	1	0.0018	0.751	1	700	adam	0.361
	IRBO*	leakyrelu	0.145	0	0.0922	0.336	1	800	adam	0.998
	KL-B*	elu	0.013	0	0.0950	0.725	5	300	rmsprop	1.019
	NPMI*	selu	0.064	0	0.0065	0.945	1	1000	rmsprop	0.118
M10	F1*	sigmoid	0.190	1	0.0087	0.091	2	800	adam	0.563
	IRBO*	sigmoid	0.084	1	0.0097	0.789	1	800	adam	1.000
	KL-B*	selu	0.088	1	0.0135	0.964	5	800	adam	0.937
	NPMI*	sigmoid	0.617	0	0.0010	0.308	1	800	sgd	0.055

Table 7: Best configuration of hyperparameters discovered by BO for CTM for each evaluation measure.

		Activation	BOW norm	Dropout	Learning Rate	Optimizer	Rho size	Hidden size	Weight decay	Median
20NG	F1*	leakyrelu	1	0.315	0.006393	adam	200	800	0.000005	0.453
	IRBO*	sigmoid	0	0.919	0.000176	sgd	200	300	0.000004	0.996
	KL-B*	leakyrelu	1	0.044	0.027539	adagrad	300	300	0.000005	0.370
	NPMI*	leakyrelu	1	0.009	0.004234	adam	200	200	0.000005	0.080
M10	F1*	rrelu	1	0.058	0.006062	adam	100	600	0.000001	0.534
	IRBO*	sigmoid	0	0.206	0.000003	adagrad	200	100	0.007168	0.997
	KL-B*	selu	0	0.602	0.003294	adam	300	1000	0.000155	0.532
	NPMI*	relu	1	0.500	0.005000	adam	300	300	0.000001	-0.028

Table 8: Best configuration of hyperparameters discovered by BO for ETM for each evaluation measure.

		Activation	BOW norm	Dropout	Learning Rate	Optimizer	Hidden size	Weight decay	Median
20NG	F1*	elu	1	0.814	0.000008	adam	700	0.000190	0.471
	IRBO*	relu	0	0.918	0.000002	adam	600	0.001485	0.987
	KL-B*	selu	1	0.157	0.004597	adam	1000	0.000076	0.424
	NPMI*	elu	0	0.121	0.000331	rmsprop	1000	0.000004	0.089
M10	F1*	softplus	0	0.182	0.000042	adam	800	0.000001	0.585
	IRBO*	selu	1	0.406	0.008958	adam	1000	0.002974	0.997
	KL-B*	leakyrelu	1	0.051	0.013990	adam	300	0.000002	0.201
	NPMI*	relu	1	0.500	0.005000	adam	300	0.000001	-0.070

Table 9: Best configuration of hyperparameters discovered by BO for ETM-PWE for each evaluation measure.

decreasing performance of the other metrics. In Figure 1, we report the Spearman correlation coefficients between metrics using all the runs of a given experiment.

Concerning LDA, when the model is optimized for the KL-B*, NPMI*, or F1*, then the IRBO is positively correlated with these metrics. It is then sufficient to optimize one of the other metrics to get also diverse topics. This occurs in particular for the KL-B* and NPMI* on long documents (0.87 and 0.98 respectively). It is also interesting to notice that optimizing for KL-B* does not imply a maximization for the F1 and NPMI on long texts. To achieve better topic coherence and classification, we should consider background topics as well.

Focusing on NeurLDA, ProdLDA, and CTM, we do not observe substantial differences between long and short documents. IRBO* is not strongly correlated with the other metrics, especially for long documents. This can be grasped by observing the coefficients IRBO* vs F1, KL-B, and NPMI reported in Figure (1f), (1h) and (1d). On the contrary, optimizing NeurLDA, ProdLDA, and CTM for F1*, NPMI* or KL-B* guarantees, in most of the cases, a good performance on all the metrics both for short and long documents (Figure (1e), (1f), (1g) and (1h)).

Concerning ETM, the difference between long and short documents is clear: the optimization of a given metric can be detrimental to the majority of the other metrics when dealing with short documents. In fact, the optimization of ETM w.r.t. IRBO* and NPMI* originates correlation values with all the other metrics that are close to zero or negative (Figure 1i). On the other hand, F1* and KL-B* seem not to be affected by the difference of the datasets. This suggests that maximizing KL-B* or F1* implies good performance also for other purposes. Focusing on long documents (Figure 1j), the optimization of ETM w.r.t. F1*, KL-B*, and NPMI* originates positive correlation values for all the other metrics. On the other hand, we can highlight that optimizing the topic diversity IRBO* does not allow us to simultaneously obtain good performance on topic coherence (NPMI) on long documents. Regarding ETM-PWE, we do not notice a clear difference between the two datasets. The introduction of the pre-trained word embedding into the training process of the model seems to be beneficial for all the metrics.

To summarize, optimizing the neural models according to the IRBO* is not always convenient and may lead to incoherent topics or poor document classification performance. Another important insight concerns the optimization of F1*, which usually guarantees to maximize IRBO, KL-B, and NPMI, for both short and long documents, except for LDA.

5 Conclusions and Future Work

In this paper, we presented an empirical analysis of Neural Topic Models to determine the relationship between hyperparameters, document length and performance measures. Three main research questions have been addressed for understanding under which conditions such Topic Models could work for guaranteeing their best performance.

The main findings could help both practitioners on tuning the models according to their objectives and researchers to explore the role of hyperparameters and document length with respect to any given performance measure. Regarding the future work, the problem of hyperparameter optimization by considering multi-objective optimization (Horn and Bischl, 2016) will be addressed for understanding to which extent multiple metrics could be optimized according to the length of the documents and the hyperparameters of the models.

References

Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. 2009. Topic significance ranking of LDA generative models. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2009, Proceedings, Part I*, pages 67–82.

Francesco Archetti and Antonio Candelieri. 2019. *Bayesian Optimization and Data Science*. Springer International Publishing.

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

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

David M. Blei. 2012. Probabilistic topic models. *Commun. 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, Yuening Hu, and David M. Mimno. 2017. Applications of topic models. *Found. Trends Inf. Retr.*, 11(2-3):143–296.

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. Association for Computational Linguistics.

Adji Bousso Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Trans. Assoc. Comput. Linguistics*, 8:439–453.

Thanh-Nam Doan and Tuan-Anh Hoang. 2021. Benchmarking neural topic models: An empirical study. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, pages 4363–4368. Association for Computational Linguistics.

Derek Greene, Derek O'Callaghan, and Padraig Cunningham. 2014. How many topics? stability analysis for topic models. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Proceedings, Part I*, volume 8724 of *Lecture Notes in Computer Science*, pages 498–513. Springer.

Geoffrey E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800.

Daniel Horn and Bernd Bischl. 2016. Multi-objective parameter configuration of machine learning algorithms using model-based optimization. In *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*, pages 1–8.

Kirthevasan Kandasamy, Karun Raju Vysyaraju, Willie Neiswanger, Biswajit Paria, Christopher R. Collins, Jeff Schneider, Barnabás Póczos, and Eric P. Xing. 2020. Tuning Hyperparameters without Grad Students: Scalable and Robust Bayesian Optimisation with Dragonfly. *Journal of Machine Learning Research*, 21:81:1–81:27.

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*, pages 530–539.

Daniel D. Lee and H. Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000*, pages 556–562. MIT Press.

Kar Wai Lim and Wray L. Buntine. 2014. Bibliographic analysis with the citation network topic model. In *Proceedings of the Sixth Asian Conference on Machine Learning, ACML 2014*.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1727–1736. JMLR.org.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, pages 3111–3119.

Derek O'Callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. 2015. An analysis of the coherence of descriptors in topic modeling. *Expert Syst. Appl.*, 42(13):5645–5657.

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 3980–3990. 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*. OpenReview.net.

Keith Stevens, W. Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012*, pages 952–961. ACL.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021a. OCTIS: Comparing and Optimizing Topic models is Simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, EACL 2021*, pages 263–270. Association for Computational Linguistics.

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

Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021b. Word embedding-based topic similarity measures. In *Natural Language Processing and Information Systems - 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021*, volume 12801 of *Lecture Notes in Computer Science*, pages 33–45. Springer.

Silvia Terragni, Debora Nozza, Elisabetta Fersini, and Enza Messina. 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, Insights 2020*, pages 32–40. Association for Computational Linguistics.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4):20:1–20:38.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray L. Buntine. 2021. Topic modelling meets deep neural networks: A survey. *arXiv*, abs/2103.00498.