Towards Better Evaluation of Topic Model Quality

Maria Khodorchenko, Nikolay Butakov, Denis Nasonov
ITMO University
Saint-Petersburg, Russia
mariyaxod@yandex.ru, alipoov.nb@gmail.com, denis.nasonov@gmail.com

Abstract—**Topic modelling is a popular unsupervised method for text corpora processing to obtain interpreted knowledge of the data. However, there is an automatic quality measurement gap between existing metrics, human evaluation and performance on the target tasks. That is a big challenge for automatic hyperparameter tuning methods as they heavily rely on the output signal to define the optimization direction. Currently, this process of evaluating the effectiveness of the topic model faces a number of difficulties and keeps being a labour-intensive routine performed manually due to the absence of a universal metric that may show strong correspondence with human assessment. The development of a quality metric that may satisfy this condition is essential to provide valuable feedback for the optimization algorithm when working with flexible and complex models, such as models based on additive regularisation or neural networks. To address the quality measurement gap, we performed an experimental study of existing scores on a specially created dataset containing topic models for several different text corpora in two languages accompanied by evaluated existing metrics and scores obtained from human assessment. The study results show how the situation with automatic quality estimation may be improved and pave the way to metrics learning with ensembles of machine learning algorithms.**

I. Introduction

Topic modelling is undeniably one of the most used methods for working with text due to its ability to convert unlabelled text data into explainable representation. The structures of topic models differ depending on the purposes of text analysis and the statistical characteristics of corpora. Classical methods such as Probabilistic Latent Semantic Analysis (PLSA) [1] or Latent Dirichlet Allocation (LDA) [2] are good baseline models for the modelling [3] as they can produce good topics representation which can be interpreted by humans while at the same time used as a feature engineering step in data preparation pipeline for classical machine learning methods. However, they have limitations which lead to poor performance on datasets with specific characteristics [4].

More flexible types of models have been proposed to improve the resulting quality. Few particularly successful examples of these classes are additively regularized topic models (ARTM) [5] and neural models [6]–[8]. These methods can produce better results by tuning a large number of hyperparameters with the appropriate quality and loss functions aligned with the task.

ARTM, being able to yield better quality, requires substantial effort to properly set its hyperparameters and often demands complementary skill and experience. It happens due to an extensive set of regularizers that can be combined and modified during the training, which leads to different modelling results. The whole procedure is often controlled by a specialist, which not only consumes a lot of time for a data scientist but eventually sets a high entrance threshold for newbies and hinders the method from gaining more popularity in the field.

To fix this situation and automate the labor-intensive step, one would require a metric that can be used to characterize the quality of resulting models and a framework that can tune the parameters according to the metric. Due to the fact that the data that is submitted to the input of topic models is an unlabelled set of documents, the quality of the resulting models solely depends on the chosen metric. In case of weak design of the metric, the resulting solution may be practically useless or lead to unbalanced efficiency between topics.

Currently, the process of evaluating the effectiveness of the topic model faces a number of difficulties. First of all, there is no unified quality estimation system, as modelling different corpora require different metrics, which complicates the comparison of performance. Common assessment metrics such as perplexity [9] allow first of all to determine whether the model has converged and to determine the criterion for stopping optimization, for example, when a reached difference e is no more than 5%. However, it does not show strong correspondence with human perception of quality and eventually with human assessments. The same is also true for other widespread metrics: coherence, npmi, kernel size, contrast, switchP and etc.

To overcome the problem with the assessment gap, we prepare the dataset consisting of a multitude of trained AR-based topic models with corresponding parameters and different existing metrics that have been evaluated for these models. All models are accompanied by a human assessment performed on Toloka [10] crowdsourcing platform. On top of this dataset, we conducted a study on metrics correspondence to human assessment and how it is influenced by the dataset itself to draw a connection between them. We also introduced several machine learning methods for automatic quality assessment and discussed possible ways of how to build a new composed metric that may yield better results.

Our research contributes the following:

1) Investigation of how different existing metrics correspond to human assessment based on an extensive experimental study with multiple datasets in different languages. The comparison is needed to estimate the connection between metrics.

2) Introduces a possible approach to improve automatic quality assessment of topic models based on machine learning that can be used in automatic or semi-automatic procedures for finding optimal parameters of topic models.
3) Introduces a new dataset (available via https://shorturl.at/giovx) for research of automatic quality estimation for topic modelling. The dataset contains resulting metrics for models trained with various input hyperparameters.

The remainder of this paper is organized as follows. Section II provides information on the background and related works on the question of topic models and quality measuring. Section III describes the assessment methodology, models sampling for human assessment and the resulting datasets. Experimental results are described in section IV. Section V provides the overall discussion, conclusions and vision of future works.

II. Background and related work

In this section, we provide a brief overview of topic models with their applications and quality-measuring approaches. Considering a rich history of topic modelling and an abundance of approaches and modifications, we highlight only broadly used models and quality metrics.

A. Topic Modeling

In general, the topic modelling task is to produce two matrices - latent distribution over topics for documents $p(t|d)$ and distribution over words for topics $p(w|t)$, in other words, we want to do soft clustering:

$$\begin{gathered} p(w|d) = \sum_{t \in T} p(w|t) p(t|d) = \\ \sum_{t \in T} \phi_{wt} \theta_{td}, d \in D, w \in W, \end{gathered} \tag{1}$$

where ϕ is a word-topic probabilities matrix, θ is a document-topic probabilities matrix, D is a collection of documents, W is a finite set of vocabulary words, and T is a set of topics. Here by latent variable topic component is assumed.

Topic models based on matrix decomposition - non-negative matrix decomposition (NMF) [11] and its modification SeaNMF [12] - are good at working with short documents such as posts in social media or hashtags.

Probabilistic models, like PLSA [13] or more popular LDA [2] are able to produce topic models with interpretable topics [14]. However, in cases with domain-specific corpora, such models in their general form experience difficulties and, thus, require to be modified along with the training procedure, which can be a labour-intensive procedure.

Models based on additive regularization (ARTM [5], [15], [16]) solve the ill-posed problem of matrix factorization by applying a set of regularizers on the task, which is represented as a simple sum. Due to the flexibility of the model, it is possible to create separate sets of topics and apply to them different sets of regularizers (the most common setting is to create specific and background topics). This can help to produce better and more informative topics without changing the logic of train and inference.

Due to the rapid development of neural models, there are a variety of models capable of producing highly optimized representations [6]–[8]. Such models can be roughly divided into two categories – ones that work on non-contextualized embeddings (like w2v, fasttext or Glove) and those that work with contextualized representations (most commonly BERT-like models). The latter can produce models with higher quality but require more calculation resources. Another problem is the length of texts because BERT-like models have limitations on the input length.

Though even with the development of neural models for specific task-solving topic models are still used not only for an understanding of contents of large corpora but also for other problems, such as information retrieval [17], downstream document classification [18], or sentiment extraction [19]. Resulting distributions over documents can be used to find the subsets of homogeneous data, which can be utilized to improve the performance of neural networks by fine-tuning the general model on domain-specific data [20].

B. Topic model quality

Due to the fact that topic modelling deals with unlabelled data, it is vital to have a good metric to measure the quality. However, there are difficulties in appropriate score selection, which is first of all related to the absence of a unified system of indicators. Different research works use distinct quality measuring approaches, which complicates performance comparison.

Initial works on topic model evaluation deal with global quality when only the topics (distribution of topics over documents) are considered for assessment and automatic scoring [21]–[23]. These approaches aim to approximate how people perceive whether the topics are clear and interpretable by utilizing co-occurrences and dependencies between words in corpora.

Authors of [21], [23] propose to estimate the quality of topic models by measuring the coherence score, which is positively correlated with human assessment of topics. In recent times effectiveness of this score is being argued in [24]. NPMI score [25] uses information on the context of the word and proved to work well for collocations. Scoring based on distributed representation [22] utilizes such properties of w2v like models as similarity notion. However, there is difficulty in training the w2v model for each dataset in the case of specific datasets.

In [5] more metrics are proposed for consideration, like topic kernel, which is a measure of the number of words which are the most probable for the topic. Topic purity and contrast measure the interpretability and difference between topics correspondingly. Sparsity phi and theta evaluate the structure of matrices. These metrics reveal the internal structure of the trained model and give an insight into its dynamics and characteristics.

TABLE I. CHARACTERISTICS OF DATASETS SELECTED FOR THE EXPERIMENTAL STUDIES

Dataset	# entries	Avg tokens	Lang	Dict size	Cls task
Lentaru	10000	119.5	ru	48874	topic tag
Amazon food reviews	10000	32.5	en	14678	-
20newsgroups	10000	117.8	en	59974	topic

Not only global scores are taken into consideration while measuring the quality of topic models [26]. It is also essential to pay attention to the obtained text representations. In [27], a new metric switchP is proposed to measure the local quality, which is a way to estimate how good the model is at describing topics of the document.

Still, the automatic evaluation is not well aligned with human evaluation [24], [28] and the problem of the lack of a good metric for measuring the quality of topic models is intensified, including with the emergence and active development of neural topic models [6]–[8], which have the ability to optimize significantly the provided quality metric, but at the same time get worse results based on the results of human perception [24].

Due to the variety of metrics, each reflects only a certain aspect of the quality of the produced model - at one of the levels. Thus, the quality evaluation of the model becomes biased towards one or another metric. It is a common practice to select local quality metrics by default, but still, they are not able to produce highly efficient models. It results in the need for more research into the topic models scoring methods.

III. EVALUATION OF TOPIC MODELS

A. Data preparation

1) Datasets: For our experiments, we selected datasets with different properties to estimate the quality of the resulting models. Characteristics of the datasets are provided in Table I. *Lentaru* and *20newsgroups* datasets contain metadata that can be used for data classification purposes (”Cls task” in the table), such as tags for the former and topics for the latter.

- 20newsgroups dataset [29] - a well-known dataset which contains newsgroups posts on 20 topics varying from religion to sport. The whole dataset has 18 000 entries.
- Lentaru dataset [30] - a collection of news from Russian electronic resource for 20 years. This resource covers a huge range of different local and global events.
- Amazon food reviews [31] - a dataset with relatively short plain reviews on various food categories from Amazon. Entries were collected for the ten years period, and the total amount of entries is 500 000.

For each of the datasets, we sampled 10000 documents to reduce computational cost and, at the same time, have a sufficient amount of data to capture the range of presented topics and subtopics.

To train topic models we performed a set of preprocessing steps which included removal of punctuation; cleaning out HTML-artifacts, links, digits; lemmatization (*Mystem* lemmatizer from *pymystem3* for Russian and *WordNetLemmatizer* from *nltk* for English). After that, stop-words were removed, and the texts with less than five tokens left were filtered out.

2) Models and topics sampling: We used ARTM models to get all the results presented in the paper. Firstly, we defined the topic modelling task for additive regularization approach [5].

The main idea of additive regularization is based on the maximization of the log-likelihood with the addition of regularizers weighted sum to produce a unique solution for matrix factorization task, in other words, to find the Φ and Θ matrices that satisfy the objective.

$$argmax_{\Phi,\Theta} \sum_{d\in D} \sum_{w\in W} n_{dw} ln \sum_{t\in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta), \quad (2)$$

where n_{dw} is a counter of word w appear in a single document d, R is the weighted sum of regularizers.

By combining existing regularizers or creating a new one, it is possible to train topic models with different characteristics that give resulting topics suitable for a particular corpus. For example, in ARTM-based models, it is possible to reduce the influence of frequent words in the documents by making a separate set of specific and background topics with different regularizers. At the same time, there are no universal hyperparameter values for various datasets that will result in a good model.

We prepared a set of topic models with additive regularization from various optimization generations with the help of the evolutionary approach described in [32]. The main idea of the method is to effectively optimize the hyperparameters of topic models with additive regularization paying attention to the iterative improvement of the models. The genetic algorithm that is used for optimization has good exploration abilities and produces mixed results.

There were five runs of the optimization algorithm made for each of the topic counts (10, 25, 50, 75, 100). We saved all the models that were trained during the runs with corresponding evaluated quality metrics.

In order to obtain models with different characteristics of ϕ and θ matrices we added a coefficient α to reward the optimizer for getting good regularization values. When the sparsity value of θ is small topic model gives a probability for each of the topics to be present in the document. In the case of values close to 1 model tends to select one topic for each document (which can be a background topic as well) which is not a desirable outcome. Thus we set preferable sparsity values in the range of [0.2,0.8].

$$Q(x) = \alpha \cdot (mean(coh_{50}(i)) + min(coh_{50}))) \quad (3)$$

where

$$\alpha = \begin{cases} 1, & \text{if } 0.2 \leq Sp_\theta \leq 0.8 \\ 0.7, & \text{otherwise} \end{cases} \quad (4)$$

where Sp_θ - sparsity of θ matrix and Coh_{50} is coherence for the 50 most probable tokens in topic.

In the following text we will refer to $Q(x)$ as a fitness function as it is a natural naming for the quality function in evolutionary framework that we use for topic models optimization.

To leave only valid topic models we filtered out ones with number of topics less than expected and checked that all of them have at least 15 tokens per topic.

From selected datasets which were described in previous subsection III-A1 two of them have targets for classification tasks which we used for sampling strong models with high and average classification scores. For amazon food reviews datasets we created pseudo-labels based on clustering results with k-means on 20 clusters.

Further sampling of filtered by classification quality models was done with the help of clustering by metrics mechanism to have different topic models for assessment. As a result we obtained 20 models for each of the topic counts. After that topics sampling was done according to the corresponding counts - 10 for model with 10 topics, 15 for models with 25 and 50 topics, 25 for 75 and 30 for 100 (100 models per dataset in total).

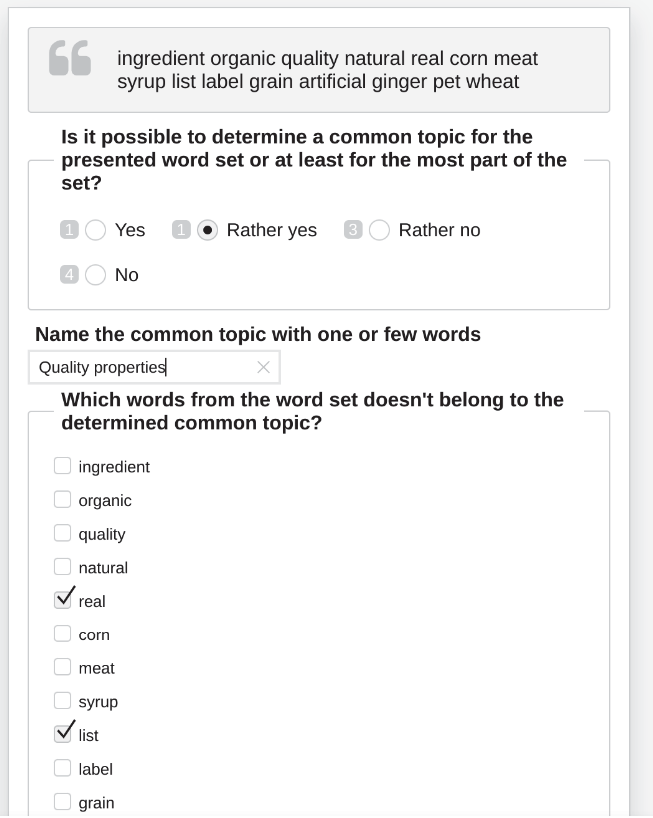
B. Methodology

Each chosen model went through a human assessment performed with the following methodology. Each of the assessors was given N tasks, which consisted of two parts:

1) Select one of four categories-characteristics of the provided set of tokens by answering the question *”Is it possible to determine a common topic for the presented word set or at least for the most part of the set?”*. They could answer one of the following: *yes* - if they agree with the statement and words have a strong connection between them, *rather yes* - if some words are too common or out of topic, *rather no* - if the amount of irrelevant words is high to determine a topic or there is a mixture of topics, *no* - when words seem to be unconnected.
2) In case of answering *yes* or *rather yes* on the first part of the task, assessors were asked to mark the words that they think are out of the topic. Also, they are asked to enter the topic's name in free form (*Name the common topic with one or few words*).

Toloka task interface is provided in Fig. 1, and it consists of three subtasks. The last two open only for answers *yes* and *rather yes*, which indicate that the words are connected by some topic, and the assessor is able to give details on his decision (naming the topic and selecting the excess words).

To get reliable results, each individual task (where the task is one topic) was given to five assessors. Each of the assessors was required to pass the training set of questions with a quality of no less than 70%. Also, we prepared control examples which were mixed into the task examples. For them, we also set a threshold of 70% correctness. Results from assessors who failed training or control tasks or completed the set of six tasks in less than a minute were excluded from further processing.



ingredient organic quality natural real corn meat syrup list label grain artificial ginger pet wheat

Is it possible to determine a common topic for the presented word set or at least for the most part of the set?

1 ○ Yes 1 ● Rather yes 3 ○ Rather no

4 ○ No

Name the common topic with one or few words

Quality properties

Which words from the word set doesn't belong to the determined common topic?

□ ingredient
□ organic
□ quality
□ natural
☑ real
□ corn
□ meat
□ syrup
☑ list
□ label
□ grain

Fig. 1. Task interface which is available for assessors

The resulting topic score is the most probable choice (majority) between the assessors. All the categories was provided with weights: 2 - *yes*, 1 - *rather yes*, -1 - *rather no*, -2 - *no*. If there is no agreement between ”good” categories (*yes*, *rather yes*) and ”bad” ones (*no*, *rather no*) the topic was given 0 score.

We calculated the overall model score as an average value of all calculated topics in the model.

IV. Experimental study

A. Metrics correlation

To measure the effectiveness of quality metrics, we calculated pairwise Pearson correlation between all the metrics and human evaluation results (Fig. 2 - 4). Fitness function (3) is denoted as ”avg_coherence_score”, and human model scoring is ”total_score”.

Considering the Fig. 2 - 4, it is clear that there is no universal metric which is equally good on all the presented datasets. For 20newsgroups high positive correlation is seen for npmi and fitness function, while the coefficient is close to zero for the classification task. A bit different situation with lentaru dataset where correlation is seen for fitness as well, coherence and background tokens ratio. On Amazon food dataset, the best metric is kernel size. Classification score is moderately correlated with human evaluation. The lowest correlation values are for amazon food dataset where there are no scores that have high coefficients.

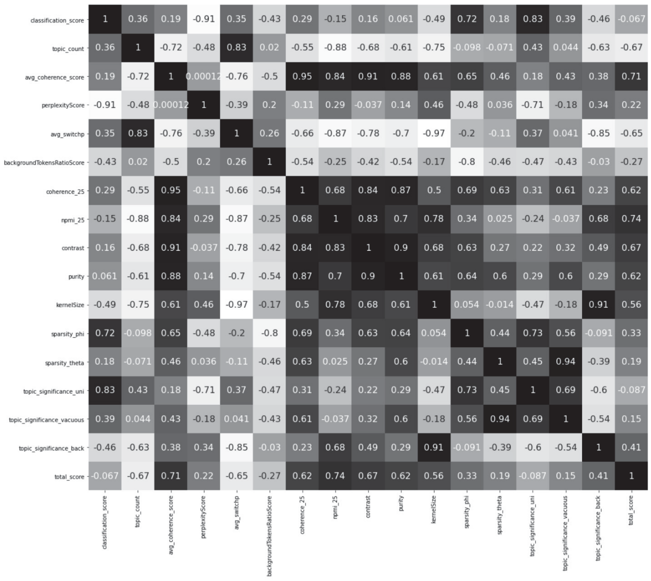


Fig. 2. Scores correlation on 20 newsgroups dataset. More intensive color means higher positive correlation.

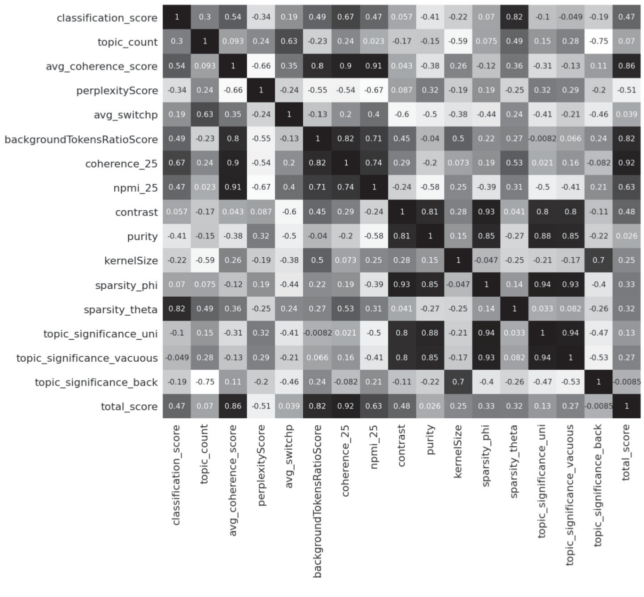


Fig. 3. Scores correlation on lentaru. More intensive color means higher positive correlation.

To ensure that the models have differences in quality, we calculated the overlap of chosen words between assessors. Tokens which were selected by more than half of the assessors are considered to be out-of-topic. For topics that have quality ”bad” or ”rather bad”, all the tokens are assumed to be selected. For each of the models, the final amount of bad words is an average of all the labelled topics. Figure 5 illustrates a stable decline in the number of selected words with evaluation score growth for 20 newsgroups dataset. The trend stays the same for the rest two datasets, namely amazon food and lentaru.

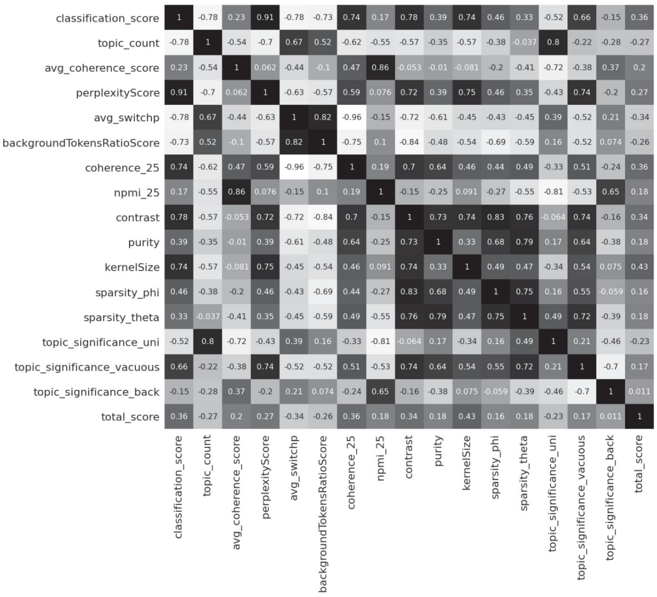


Fig. 4. Scores correlation on amazon fine food dataset. More intensive color means higher positive correlation.

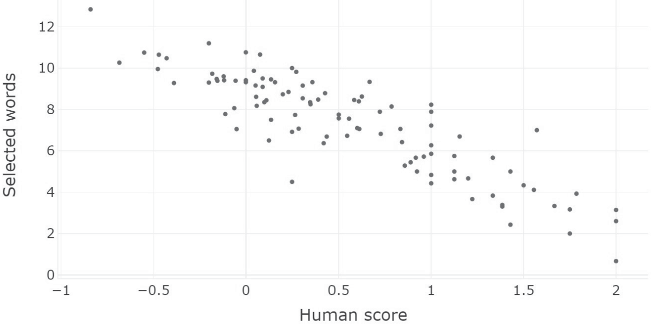


Fig. 5. Connection between average selected out-of-topic words and human score.

B. Score development

To develop a new approach to quality estimation, we decided to apply metric learning based on an ensemble of classical machine learning algorithms, namely linear regression and gradient boosting.

All the training was performed with the help of LightAutoml framework [33] as it is a high-speed solution for building ensembles of ML models and their hyperparameter tuning and feature selection. Moreover, the framework can prune one or several models if it sees that a subset of initial ML algorithms performs better than the whole ensemble.

We trained ensembles following the two next settings:

- ”Native” - the dataset is split into several folds. We choose one fold to be for testing purposes, and the reset folds are used for training. This routine is repeated a number of times equal to a number of folds. Predictions on this testing fold are combined with predictions of models on their corresponding testing folds and form predictions for the whole initial dataset. These predictions were later used to measure correlation with the human assessment.
- ”General” - an ensemble is trained on two datasets and tested on the third one. In this setting, we wanted to check if it is possible to create a generalizable model.

Results of 5-fold validation are provided in Fig. 6. The quality of the ”native” model is significantly higher on all the datasets in the experiment. The largest gaps are for lentaru and 20newsgroups datasets which have a higher topic diversity. Also, it should be noted that the highest variance is for amazon food dataset, which may be due to short texts modelling that requires more time to find good hyperparameters.

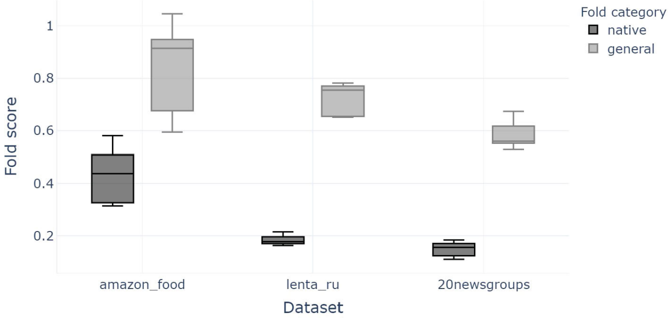


Fig. 6. Results of MAE calculation on 5 folds for native and general settings. Lower values indicate better prediction capabilities.

To compare the models trained in different settings with each other and the outputs of existing metrics, we calculated the Pearson correlation. Results in Table II indicate that the best correlation is for the native score as it shows a solid linear connection with the human assessment. It should be noted that for the table, we selected the metrics with the highest correlation with the calculated new scores.

TABLE II
GENERAL AND NATIVE SCORES CORRELATION RESULTS WITH EXISTING EVALUATION METRICS

	Gen. score	Native score	Fitness	SwitchP	Coh 25	Npmi 25	Kernel size
20ng	0.72	**0.96**	0.71	-0.65	0.62	0.74	0.56
lenta ru	0.87	**0.93**	0.86	0.04	0.92	0.63	0.25
amazon food	0.17	**0.84**	0.2	-0.34	0.36	0.18	0.43

V. DISCUSSION AND CONCLUSION

From the presented results, several conclusions may be drawn.

For text corpora with various characteristics (average number of tokens in documents, language, size of the resulting dictionary, etc.), there is a different correlation level for the same selected quality metric, which in turn leads to the non-optimality of using the same metric when training models on different text corpora. Though, it should be noted that there are a number of metrics that show relatively high results on different enclosures, which suggests their limited versatility.

A quality assessment model trained on the texts of the target corpora is likely to be better than a quality assessment model trained on several combined sets of models trained on available text corpora, excluding the target one. However, an increasing amount of such corpora may improve the situation as the trained model will be able to see more connections. It also should be noted that the correlation on its own may be enough to guide the optimization algorithm that is used for hyperparameter tuning.

It follows from the previous conclusion related to the superior quality of the native quality assessment model that using the human-in-the-loop approach for partial markup can help to get a better quality assessment. For instance, a scheme with sending several sampled topics for labelling to crowdsourcing platform, such as Toloka or Amazon Mechanical Turk, while training may lead to high quality model for the particular dataset.

At the same time, the general model, though it does not show the highest correlation for specific datasets, demonstrates one of the best correlations on all datasets compared to the entire set of metrics from which one will have to make a choice when a new text corpus arrives.

Thus the new approach may be used for the optimization procedure. It may split the optimization process into three stages. In the first stage, an optimization algorithm using the general model (or even using several independent ”islands” with different quality metrics that show the highest average correlation on labelled datasets) as a quality estimator grows initial sets of various solutions for which individual metrics are evaluated. In the second step, the grown topic models are sampled, and the tasks for human assessment are formed and submitted automatically through Toloka platform API. In the final third stage, the new quality estimation model is trained on the labelled data and is used to guide the optimization algorithm further to grow the final solution. Such an approach requires only a small part of data to be marked up and, thus, can be applied in practical tasks.

In future work, we plan to increase the number of labelled models and datasets, which will be used to train more specific metrics. Also, we are going to extend the assessment methodology and introduce a task which will aim at measuring the quality of resulting document-topic distributions. At the same time, we are going to develop ideas of transfer learning and fine-tuning in topic model quality measuring.

Acknowledgment

This research is financially supported by The Russian Science Foundation, Agreement #20-11-20270.

References

[1] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '99. New York, NY, USA: Association for Computing Machinery, 1999, p. 50–57. [Online]. Available: https://doi.org/10.1145/312624.312649

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. null, p. 993–1022, mar 2003.

[3] I. Vayansky and S. A. Kumar, "A review of topic modeling methods," *Information Systems*, vol. 94, p. 101582, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306437920300703

[4] M. Hajjem and C. Latiri, "Combining ir and lda topic modeling for filtering microblogs," *Procedia Computer Science*, vol. 112, pp. 761–770, 2017, knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050917315235

[5] K. Vorontsov, O. Frei, M. Apishev, P. Romov, and M. Dudarenko, "Bigartm: Open source library for regularized multimodal topic modeling of large collections," 2015, pp. 370–381.

[6] D. Card, C. Tan, and N. A. Smith, "Neural models for documents with metadata," in *ACL*, 2018.

[7] H. Bai, Z. Chen, M. R. Lyu, I. King, and Z. Xu, "Neural relational topic models for scientific article analysis," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, ser. CIKM '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 27–36. [Online]. Available: https://doi.org/10.1145/3269206.3271696

[8] M. Rezaee and F. Ferraro, "A discrete variational recurrent topic model without the reparametrization trick," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[9] C. Meister and R. Cotterell, "Language model evaluation beyond perplexity," 2021. [Online]. Available: https://arxiv.org/abs/2106.00085

[10] "Toloka ai: Powering data-centric ai," https://toloka.ai/.

[11] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the -divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[12] T. Shi, K. Kang, J. Choo, and C. K. Reddy, "Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations," in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, p. 1105–1114. [Online]. Available: https://doi.org/10.1145/3178876.3186009

[13] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, p. 289–296.

[14] J. Rieger, C. Jentsch, and J. Rahnenführer, "RollingLDA: An update algorithm of Latent Dirichlet Allocation to construct consistent time series from textual data," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2337–2347. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.201

[15] V. Bulatov, V. Alekseev, K. V. Vorontsov, D. Polyudova, E. Veselova, A. Goncharov, and E. S. Egorov, "Topicnet: Making additive regularisation for topic modelling accessible," in *LREC*, 2020.

[16] D. Kochedykov, M. Apishev, L. Golitsyn, and K. Vorontsov, "Fast and modular regularized topic modelling," in *2017 21st Conference of Open Innovations Association (FRUCT)*, 2017, pp. 182–193.

[17] J. Guo, Y. Cai, Y. Fan, F. Sun, R. Zhang, and X. Cheng, "Semantic models for the first-stage retrieval: A comprehensive review," *ACM Trans. Inf. Syst.*, vol. 40, no. 4, mar 2022. [Online]. Available: https://doi.org/10.1145/3486250

[18] A. Zamiralov, M. Khodorchenko, and D. Nasonov, "Detection of housing and utility problems in districts through social media texts," *Procedia Computer Science*, vol. 178, pp. 213–223, 2020, 9th International Young Scientists Conference in Computational Science, YSC2020, 05-12 September 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920323978

[19] T. Sokhin and N. Butakov, "Semi-automatic sentiment analysis based on topic modeling," *Procedia Computer Science*, vol. 136, pp. 284–292, 2018, 7th International Young Scientists Conference on Computational Science, YSC2018, 02-06 July2018, Heraklion, Greece. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050918315965

[20] E. Nevezhin, N. Butakov, M. Khodorchenko, M. Petrov, and D. A. Nasonov, "Topic-driven ensemble for online advertising generation," in *COLING*, 2020.

[21] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *NAACL*, 2010.

[22] S. I. Nikolenko, "Topic quality metrics based on distributed word representations," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1029–1032. [Online]. Available: https://doi.org/10.1145/2911451.2914720

[23] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*. Potsdam, Germany: Association for Computational Linguistics, Mar. 2013, pp. 13–22. [Online]. Available: https://aclanthology.org/W13-0102

[24] A. M. Hoyle, P. Goel, D. Peskov, A. Hian-Cheong, J. L. Boyd-Graber, and P. Resnik, "Is automated topic model evaluation broken?: The incoherence of coherence," in *NeurIPS*, 2021.

[25] D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin, "Evaluating topic models for digital libraries," in *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, ser. JCDL '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 215–224. [Online]. Available: https://doi.org/10.1145/1816123.1816156

[26] C. Doogan and W. Buntine, "Topic model or topic twaddle? re-evaluating semantic interpretability measures," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 3824–3848. [Online]. Available: https://aclanthology.org/2021.naacl-main.300

[27] J. Lund, P. Armstrong, W. Fearn, S. Cowley, C. Byun, J. Boyd-Graber, and K. Seppi, "Automatic evaluation of local topic quality," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 788–796. [Online]. Available: https://aclanthology.org/P19-1076

[28] M. Khodorchenko and N. Butakov, "Developing an approach for lifestyle identification based on explicit and implicit features from social media," *Procedia Computer Science*, vol. 136, pp. 236–245, 2018, 7th International Young Scientists Conference on Computational Science, YSC2018, 02-06 July2018, Heraklion, Greece. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050918315679

[29] K. Lang, "Newsweeder: Learning to filter netnews," in *Proceedings of the Twelfth International Conference on Machine Learning*, 1995, pp. 331–339.

[30] D. Yutkin, "Corpus of russian news articles collected from lenta.ru." [Online]. Available: https://github.com/yutkin/Lenta.Ru-News-Dataset

[31] J. J. McAuley and J. Leskovec, "From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews," in *Proceedings of the 22nd International Conference on World Wide Web*, ser. WWW '13. New York, NY, USA: Association for Comp. Machinery, 2013, p. 897–908.

[32] M. Khodorchenko, S. Teryoshkin, T. Sokhin, and N. Butakov, "Optimization of learning strategies for artm-based topic models," in *Hybrid Artificial Intelligent Systems*, E. A. de la Cal, J. R. Villar Flecha, H. Quintián, and E. Corchado, Eds. Cham: Springer International Publishing, 2020, pp. 284–296.

[33] A. Vakhrushev, A. Ryzhkov, M. Savchenko, D. Simakov, R. Damdinov, and A. Tuzhilin, "Lightautoml: Automl solution for a large financial services ecosystem," 2021. [Online]. Available: https://arxiv.org/abs/2109.01528