Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2023"

June 14–16, 2023

Attention-based estimation of topic model quality

Veronika Kataeva
ITMO University
St Petersburg, Russia
kataevaveronika@niuitmo.ru

Maria Khodorchenko
ITMO University
St Petersburg, Russia
mkhodorchenko@niuitmo.ru

Abstract

Topic modeling is an essential instrument for exploring and uncovering latent patterns in unstructured textual data, that allows researchers and analysts to extract valuable understanding of a particular domain. Nonetheless, topic modeling lacks consensus on the matter of its evaluation. The estimation of obtained insightful topics is complicated by several obstacles, the majority of which are summarized by the absence of a unified system of metrics, the one-sidedness of evaluation, and the lack of generalization. Despite various approaches proposed in the literature, there is still no consensus on the aspects of effective examination of topic quality. In this research paper, we address this problem and propose a novel framework for evaluating topic modeling results based on the notion of attention mechanism and Layer-wise Relevance Propagation as tools for discovering the dependencies between text tokens. One of our proposed metrics achieved a 0.71 Pearson correlation and 0.74 ϕ_K correlation with human assessment. Additionally, our score variant outperforms other metrics on the challenging Amazon Fine Food Reviews dataset, suggesting its ability to capture contextual information in shorter texts.

Keywords: Topic modeling, evaluation metrics, language models, attention mechanism, Layer-wise Relevance Propagation

DOI: 10.28995/2075-7182-2023-22-215-224

Оценка качества тематических моделей на основе механизма внимания

Аннотация

Тематическое моделирование является важным инструментом для исследования и выявления скрытых закономерностей в неструктурированных текстовых данных, что позволяет исследователям и аналитикам извлекать ценную информацию о какой-либо конкретной области. Тем не менее, тематическое моделирование не имеет единого мнения по вопросу его оценки. Оценивание полученных тем осложняется несколькими препятствиями, большинство из которых сводится к отсутствию единой системы метрик, односторонности оценки и недостаточной обобщаемости. Несмотря на различные подходы, предложенные в литературе, до сих пор нет единого мнения об аспектах эффективной и качественной экспертизы полученных тем. В данной исследовательской работе мы рассматриваем эту проблему и предлагаем новую систему оценки результатов тематического моделирования, основанную на понятии механизма внимания и послойного распространения релевантности как инструментов для обнаружения зависимостей между текстовыми токенами. Одна из предложенных нами метрик достигла корреляции Пирсона 0,71 и корреляции ϕ_K 0,74 с сравнении с оценками человека. Кроме того, наш вариант метрики превосходит другие методы оценивания на сложном наборе данных Amazon Fine Food Reviews, что свидетельствует о его способности фиксировать контекстную информацию в более коротких текстах.

Ключевые слова: Тематическое моделирование, метрики оценки, языковые модели, механизм внимания, послойное распространение релевантности

1 Introduction

The tremendous growth of digital information in recent years has evoked an increasing need to effectively process and analyze an enormous amount of text in short time. As far as most of the information is not labeled and markup with assessors takes resources and time, there is a clear tendency to utilize such data with unsupervised methods.

Topic modeling has emerged as an essential instrument for identifying semantically related sets of words that holistically encapsulate the underlying information in the document collection. The topic model receives a corpus of text and outputs the topic distribution for each document and the word distribution for each topic. While such approaches as Latent Dirichlet Allocation (LDA) highly rely on the prior which significantly constrains the possible solutions, others, like Additive Regularization (ARTM) are much more flexible and thus demand to be tuned for each of the input text corpora (Bulatov et al., 2020; Khodorchenko et al., 2022b).

Nonetheless, the estimation of resulting topic models is complicated due to several obstructions, primarily caused by a lack of a unified system of metrics. Across various papers, researchers conduct their experiments differently and employ a variety of metrics, hence intricating the comparison of performances (Abdelrazek et al., 2023). Furthermore, Doogan and Buntine substantiate the need for new evaluation measures, as the new models may be incompatible with older metrics. Another negative aspect of evaluation is the absence of generalization in experimental settings. This problem is exacerbated by the non-availability of benchmark datasets, compared to, for example, classification tasks (Doogan and Buntine, 2021). Furthermore, the metrics may reflect only a particular side of the produced model quality (Hoyle et al., 2021). Additionaly, best evaluation metrics can differ from dataset to dataset (Khodorchenko et al., 2022a). The suboptimal decision of the best topic model may cause an inaccurate representation of data and, therefore, its biased understanding. Thorough and comprehensive manual control of topic modeling outputs is still required, and it is critical for obtaining unbiased and high-quality results (Rüdiger et al., 2022). Various metrics emerged in attempts of evaluating topic quality, such as Normalized Pair-wise Mutual Information (NPMI), Perplexity, Topic Switch Percent (SwitchP), Coherence, Topic Significance, etc. However, they are not capable of closing the gap in evaluation.

As well as most of the existing topic model quality estimation scores do not fully encounter the context and rely mostly on statistics of the corpora at hand, in this paper we're addressing the power of language models. Transformer (Vaswani et al., 2017) models specifically have proven their effectiveness for numerous natural language understanding tasks, making them state-of-the-art architecture. One of the essential features of the models is attention mechanisms, which facilitate selective focus on certain parts of the input when making predictions, as well as allowing to identify the relationship of input with itself.

In this study, we propose to calculate the frequency and strength of the relationships between pairs of words in the topics with regard to attention scores. Each topic consists of words that are present in the text corpus, and a fine-tuned language model, having a deep understanding of the structure and semantics of data it has been trained on, can detect latent associations and their strength between them.

Our main contributions can be summarized as follows: (1) a novel approach for topic model quality estimation based on attention extraction with Layer-wise Relevance Propagation (LRP) mechanism, (2) an analysis of various ways to utilize attention information for the presented task, (3) a comparative study of correlations between different metrics with human evaluation to justify the proposed approach.

2 Related Work

2.1 Topic Modeling

Topic modeling has a long development story and includes wide range of models with the goal to extract latent component of the corpora which defines the topic starting from matrix factorization approaches (NMF (Févotte and Idier, 2011), SeaNMF (Shi et al., 2018)) to neural-based models (Card et al., 2018; Bai et al., 2018). The task, in general, can be formulated as follows:

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

where ϕ is a matrix of token probabilities in the topic, θ is a matrix of topic probabilities in the documents, D is a collection of documents, W is a finite set of vocabulary tokens, and T is a set of topics.

Among probabilistic models LDA approach (Blei et al., 2003) is being used as a solid baseline for modeling purposes despite a range of criticism which include weekly explainable Dirichlet prior and difficulties in inference adaptation to domain-specific corpora, though they can be enhanced in terms of parameter learning (Deeva et al., 2023).

In recent years, active development of neural topic models results in a wide range of new models (Card et al., 2018; Tian et al., 2020). At the same time, such models are vulnerable to overfitting and thus demand a carefully designed quality metric and loss function.

A semi-probabilistic additive regularization approach (Vorontsov and Potapenko, 2014) is one of the most flexible in terms of domain-specific models creation, as it allows combining regularizers to produce models with specific characteristics. Still, while providing such wide tools for model designing, it significantly increases the number of hyperparameters to be tuned.

2.2 Evaluation Metrics

Throughout the development of topic modeling, a range of automated metrics have been developed to quantify the performance of topic models.

The earliest and later heavily criticized (Hoyle et al., 2021; Doogan and Buntine, 2021) for low correlation with human assessments and unreliability are perplexity (Blei et al., 2003) which measures predictive likelihood of document given topic matrix and hyperparameters and topic coherence (Newman et al., 2010) that is based on pairwise words concurrences in the corpora. One of the prominent variations of coherence is NPMI has shown a substantial correlation with human judgment on word relatedness in previous studies. It compares joint probability of words to the probability of them occurring independently.

One of the recent approaches to assessing topic quality is the Topic significance (Lund et al., 2017). The metric considers the entire topic-word distribution, unlike the coherence measure. SwitchP (Lund et al., 2019) estimates local topic quality, that regards to the quality of a topic within a specific document. SwitchP demonstrates a higher positive correlation with human judgments in comparison to coherence (Lund et al., 2019; Rezaee and Ferraro, 2020).

While coherence is viewed as a measure of topic interpretability, there are introduced several attempts to evaluate other topic qualities, such as topic stability (Xing and Paul, 2018) and topic diversity (Dieng et al., 2020).

It is essential to note that some researchers apply combination of metrics (Dieng et al., 2020) or view topic modeling as classification or clustering (Harrando et al., 2021).

The scope of this paper is to study the usefulness of neural networks for topic quality assessment.

3 Attention-based topic model evaluation

The proposed attention-based topic evaluation consists of several steps, which include 1) language model fine-tuning to acquire the connections in input text; 2) performing layer-wise propagation to understand which heads and words connections are important; 3) identifying co-dependency value of individual pairs of tokens in topic and averaging the values; 4) calculating the final model quality by averaging scores from step 3 for all topics.

The first stage of our research involves fine-tuning the language model BERT (Devlin et al., 2018) to solve text classification tasks. One of the key features of the model is the multi-head attention mechanism. Learned in parallel, multiple attention heads produce versatile representations that provide various aspects of the input (Vaswani et al., 2017). For our experiments, we choose BERT and RuBERT (Kuratov and Arkhipov, 2019), BERT adaptation for the Russian language (further both referred as BERT). Both models consist of 12 Transformer blocks, where each layer has 12 heads.

As far as the original architecture of the BERT model appears as a black box, several approaches attempt to explain the decision-making behind the model predictions by addressing attention mechanisms. This work employs these methods to trace which interconnections between tokens BERT may detect: Layer-wise propagation (LRP) (Bach et al., 2015); Improved LRP (Chefer et al., 2020); raw output attentions.

Output layer attentions in Transformer-based architectures are calculated with multi-head attention mechanism which concatenates the results from all layer heads (eq. 2-4).

$$A(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

$$head_i = A(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

$$Multihead(Q, K, V) = Concat_i(head_i)W^o \tag{4}$$

where Q - query matrix, K - key, V - value, d_k - key dimensionality, W_i and W^o are parameter matrices.

LRP algorithm intends to examine the individual contribution of input to model output by propagating the relevance of the output back through the network layers consecutively to the input, using the same set of weights that have been used to compute the output (Voita et al., 2019).

Propagating relevance scores at a given layer are calculated as:

$$R_{i\leftarrow k}^{(l,l+1)} = R_k^{(l+1)} \frac{a_i w_{ik}}{\sum_h a_h w_{hk}}, \tag{5}$$

where k, i are neurons from layer $l+1$, and proceeding layer l, provided that i has a forward connection to k, w is weights, and a is an output from an activation function.

In Improved LRP, the local relevances are assigned based on the Deep Taylor Decomposition (Montavon et al., 2015), a method based on the Taylor series. This propagation involves an advanced approach to operating with matrix multiplication and skip connections.

The output of the method is defined through the weighted attention relevance:

$$\bar{A}^{(b)} = I + E_h(\nabla A_{(b)} \odot R^{n_b}) \tag{6}$$

$$C = \bar{A}^{(1)} \cdot \bar{A}^{(2)} \cdot ... \cdot \bar{A}^{(B)}, \tag{7}$$

where $\odot$ is Hadamard product, E_h is the mean across attention heads, $A^{(b)}$ is attention map of block b and $\nabla A_{(b)}$ is its gradients, R^{n_b} is layer's relevance.

Since our focus is solely on the relevance of the heads, we do not continue propagating down to the input variables and instead stop at desired self-attention layer, producing relevance matrices.

The next step is dedicated to determining BERT attention heads that carry non-trivial information. Since multi-head attention block learns different representations, each attention function may dissimilarly contribute to forming a prediction.

We use head confidence to handle raw output attentions, which is proportional to the average highest attention weight assigned across all instances of the evaluation dataset, excluding the end-of-sentence token. LRP-based techniques estimate head importance by calculating head relevance as the sum of the neurons' relevances within the head, normalized across all heads within a layer. The final relevance of a head is the mean relevance in the evaluation dataset (Voita et al., 2019).

The task of topic model quality estimation in case of attention-based approach can be generalized as an average quality of the resulting topics

$$Q^b = \frac{1}{T} \sum_t^T \sum_s^S \sum_h^H \sum_{i,j,i\neq j}^N a_{tshij}^b, \tag{8}$$

where T is the overall number of topics, S is overall number of texts, H is overall number of confident/important heads, N - amount of tokens in text and size of attention matrix, i is the i-th token of n-th text that attends to j-th token, $a^b \in \{a^{Attn\_sum}, a^{Imp\_LRP\_sum}, a^{LRP\_sum}, a^{Count}\}$ is an element of one of attention/relevance matrix.

Depending on the type of a^b we defined 4 alternative quality functions:

1. Q^{Attn_sum} denoting *Attention sum*, which derives information from output attentions matrix (eq. 4) from head with maximum relevance according calculated as $\frac{1}{n}max_i^L(head_i^l)$, where n is amount of layers,
2. $Q^{Imp\_LRP_sum}$ denoting *Improved LRP sum*, which identifies important heads by relevance matrices obtained through propagation by employing Improved LRP approach (eq. 6-7),
3. Q^{LRP_sum} denoting *LRP sum*, which uses LRP matrices as a source of information on tokens interconnections according to eq. 5,
4. Q^{Count} denoting *Count*, which uses binary matrix ($a_{ij} = 1$ if $a_{ij} > 0$ obtained from relevance matrix) that shows any present co-dependent token pairs.

4 Experimental study

4.1 Datasets

In this work, we use three datasets in Russian and English languages:

1. 20 Newsgroups dataset (Lang, 1995), a collection of news posts that covers 20 various topics including sports, religion, science, and politics.
2. Lenta.ru dataset, which comprises Russian news from an electronic resource spanning 20 years.
3. Amazon Fine Food Reviews (Amazon Reviews) dataset (McAuley and Leskovec, 2013), which consists of short reviews on food categories gathered over 10 years.
4. The dataset with evaluation of topics (Khodorchenko et al., 2022a) contains automatic and human scores for a variety of sampled topics outputted by 100 variously configured ARTM models with different amounts of topics built on the datasets 1-3 from this list. To measure the quality of topics, they were presented as tasks in Toloka (Tol, 2023) crowdsourcing platform interface. The assessors were asked whether a common topic for the presented word set is distinguishable. If the answer is positive, they are asked to name the topic and identify irrelevant to the topic words. Each of the topics was evaluated by several assessors to fit into weighted categories: a score of 2 for *yes*, 1 for *rather yes*, -1 for *rather no*, and -2 for *no*, with a score of 0 awarded in cases of inhomogeneous evaluations of human assessors. To compute correlation coefficients between human and automated model quality, we average the quality scores for each topic, enabling a comparison with human decisions.

Each dataset (1-3 from datasets list) is reduced to contain approximately 10 000 samples in order to diminish computational costs. Finally, the texts are pre-processed by removing any HTML tags, punctuation, links, tags, digits, and stop words, as well as by being lemmatized and filtered out to contain more than five tokens.

Both 20 Newsgroups and Lenta.ru contain pre-defined topic-related labels, whereas, for Amazon Reviews, we establish the labels by K-means clustering on reduced via Truncated SVD TF-IDF vectors with the number of clusters equal to 20.

We fine-tuned BERT instances on the classification task for each of the datasets. Details on hyperparameters of BERT are presented in Table 1. Models achieved F1-scores of 0.8545, 0.8056, and 0.8933 correspondingly, denoting sufficient understanding of texts BERT models have learned.

Input data	Model	Max len	Batch size	Learning rate	# of epochs
20 Newsgroups	$BERT_{BASE}$	128	8	5e-5	5
Lenta.ru	$RuBERT_{BASE}$	256	8	1e-5	5
Amazon Reviews	$BERT_{BASE}$	256	8	1e-5	5

Table 1: Hyperparameters of fine-tuned models.

4.2 Attention-based metrics performance

To understand the effectiveness of the developed metrics, we conduct a correlation analysis, using Pearson's r to detect linear relationships and ϕ_K to trace non-linear ones.

Firstly, correlation is measured between the human assessments and proposed metrics. Specifically, we measure dependencies for the quality values of distinct topics. The results of our experiments indicate that the Count metric of interconnections between tokens exhibits a greater degree of correlation than other attention-based methods. However, it is characterized by larger fluctuations of coefficient values across different datasets. In contrast, two LRP-based approaches demonstrate significantly greater stability.

Secondly, we measure the correlation between human assessments and model scores (see Table 2) calculated as the mean quality of all topics within each model.

In assessing the performance of attention-based metrics at the model level, our examination has revealed that the Improved LRP approach demonstrates higher correlation values than other techniques. Nonetheless, the LRP approach remains a viable and competitive option, displaying significant performance on the 20 Newsgroups dataset.

Dataset	Corr.	Attn sum	Imp. LRP sum	LRP sum	Count
20 Newsgroups	r	0.74	0.51	**0.78**	0.73
	ϕ_K	0.68	0.64	**0.86**	0.61
Lenta.ru	r	0.65	**0.79**	0.65	0.20
	ϕ_K	**0.80**	0.63	0.69	0.75
Amazon Reviews	r	0.65	**0.75**	0.69	0.61
	ϕ_K	0.61	**0.73**	0.69	0.66

Table 2: The correlation between human assessments of model quality and scores from developed automated metrics, as measured by Pearson's r and ϕ_K coefficients.

Figure 1 shows the propagated relevance matrix derived via applying Improved LRP to Amazon Reviews text. One of the topics obtained during topic modeling is *dog treat chew toy bone teeth ball puppy training liver piece bread pet play vet*, which was unanimously voted by assessors as a good topic. As we can see, the explainability method can identify significant relationships between words in the text *dog*, *treat*, and *pet*, which are part of the aforementioned topic as well and therefore contribute to a higher automated quality score during the proposed attention-based metrics calculations.

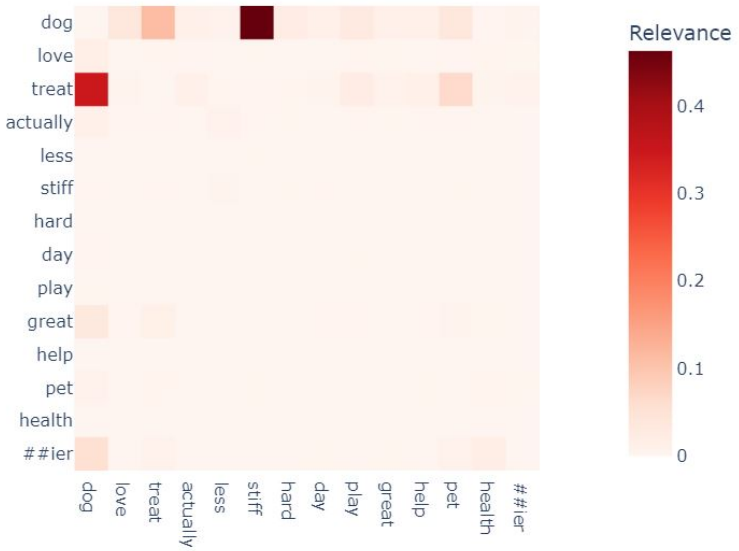

Figure 1: Relevance matrix of Amazon Reviews text derived using Imporved LRP. By employing this matrix with topic word set *dog treat chew toy bone teeth ball puppy training liver piece bread pet play vet*, we notice the substantial dependencies between topic and text words *dog*, *treat*, and *pet*.

Examples of obtained top-10 most probable words per topics from Lenta.ru dataset and corresponding scoring are presented in Table 3.

Quality	Score	Topic	Attn sum	Imp. LRP sum $\times 1e6$	LRP sum $\times 1e4$	Count
High	2	уголовный статья убийство обвинение следствие срок расследование преступление прокуратура комитет	349.94	650	790	11944
	1	олимпийский япония японский спортсмен спорт олимпиада сочи алкоголь спортивный клуб	31.82	62	2.3	2500
	2	продажа строительство квартира стоимость метр продавать квадратны жилье площадь сделка	207.06	240	490	7628
Low	-2	святой медиа сенат австрия действительность уверять алиев добывать окончательно разрушение	2	0.024	0.12	36
	-2	относиться певица опрос потеря свидетельствовать опрашивать сегодняшний рождаться рождение треть	5.94	37	2.2	400
	-1	деньги знать пытаться жить узнавать удаваться говорить решать вернуть помогать	43.79	17	-4	3192

Table 3: Examples of obtained high- and low-quality top-10 most probable words per topics obtained from Lenta.ru dataset with corresponding human labeling and received with proposed metrics scores.

Figure 2 illustrates how individual topics are scored by different metrics. Improved LRP is showing the best ability in distinguishing between high and low quality topics. It should be also noted that different automatic metrics make mistakes on different examples, so it is potentially possible to make an ensemble approach to better approximate human labelling.

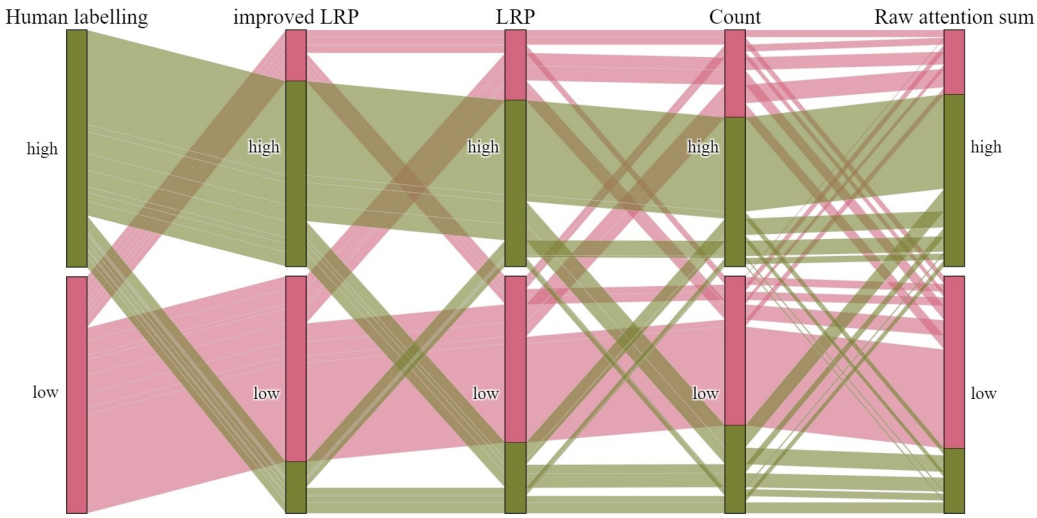

Figure 2: Quality comparison of individual topics. This illustration shows a scoring for each individual sentence. For “Human labelling” scores -1 and -2 were combined to “low” category and +1 and +2 - to “high” category. We also balanced amount of “high” and “low” scores by human labelling with random sampling. For other metrics intervals were divided by median value. “Improved LRP” shows better abilities in high and low quality topics.

4.3 Model-level correlation with human assessment comparison for automatic metrics

To estimate the general performance of the proposed metric, we compare the best attention-based variants (LRP and Improved LRP), and other commonly used metrics in both academia and applied settings, based on their ability to approximate human judgment. The results are presented in Table 4.

Dataset	20 Newsgroups		Lenta.ru		Amazon Reviews		Average	
Correlation	r	ϕ_K	r	ϕ_K	r	ϕ_K	r	ϕ_K
Improved LRP sum (our)	0.51	0.64	**0.79**	0.63	**0.75**	**0.73**	0.68	0.67
LRP sum (our)	0.78	0.86	0.65	0.69	0.69	0.69	**0.71**	**0.74**
NPMI	**0.86**	0.72	0.75	0.78	0.43	0.52	0.68	0.67
Perplexity	0.28	0.71	-0.43	0.75	0.49	0.58	0.4	0.68
Background Tokens Ratio	-0.22	0.58	0.73	0.8	-0.58	0.47	0.51	0.62
Avg SwitchP	-0.75	0.68	-0.2	0.7	-0.73	0.67	0.56	0.68
Coherence	0.71	0.71	0.75	0.8	0.53	0.69	0.66	0.73
Contrast	0.67	**0.94**	0.59	**0.87**	0.51	0.39	0.59	0.73
Purity	0.73	0.71	0.19	0.77	0.35	0.0	0.42	0.49
Kernel Size	0.65	0.64	0.24	0.59	0.44	0.62	0.44	0.62
Topic Significance Avg	0.29	0.72	0.19	0.67	0.25	0.54	0.25	0.64

Table 4: The Pearson's r and ϕ_K correlation coefficients between human assessments of model quality and scores produced by automated metrics, including our novel attention-based approaches and widely-used automated metrics. Best scores are indicated by bold text, while worst results are underlined. Average is calculated as an average correlation strength without sign.

One of the most stable results according to an average of the scores is demonstrated by LRP sum metric for both of the correlations. In this case, attention-based metric is showing good performance regardless of the dataset. At the same time, Improved LRP sum shows superior performance on Lenta.ru (linear correlation) and Amazon reviews (both correlations) while being worse on average. Proposed attention-based scores in general indicate good linear and non-linear correlations with human assessment.

Considering the results, our findings demonstrate the lack of consensus in the observed results, highlighting the existence of varying degrees of linear and non-linear correlation with human judgment across the different metrics evaluated. However, proposed LRP Sum metric can be used as a good metric for topics on average and Improved LRP version – for shorter text cases.

5 Conclusion and Future Work

In this paper, we presented an attention-based method to evaluate topic model quality. Results indicate that the proposed utilization of LRP approach to extract and summarize the interconnections between words in the topics based on fine-tuned BERT architecture is showing a better quality compared to other existing metrics, reaching 0.71 Person and 0.74 ϕ_K correlations with human assessment for LRP sum score. At the same time, Improved LRP sum score variant is revealing superior quality on the most difficult for other metrics dataset – Amazon Reviews, indicating its ability to catch more context-based information in case of shorter texts.

In future work, we are going to conduct experiments in the setting where BERT fine-tuning is done on the task of masked language model task to omit the necessity of label creation in case of their absence. We will also investigate ways to speed up LRP computations to insert the proposed scores into a topic models tuning framework.

Acknowledgements

This research is financially supported by the Foundation for National Technology Initiative's Projects Support as a part of the roadmap implementation for the development of the high-tech field of Artificial Intelligence for the period up to 2030.

References

Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102–131.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07.

Haoli Bai, Zhuangbin Chen, Michael R. Lyu, Irwin King, and Zenglin Xu. 2018. Neural relational topic models for scientific article analysis. // *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, P 27–36, New York, NY, USA. Association for Computing Machinery.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar.

Victor Bulatov, Vasiliy Alekseev, Konstantin Vorontsov, Darya Polyudova, Eugenia Veselova, Alexey Goncharov, and Evgeny Egorov. 2020. TopicNet: Making additive regularisation for topic modelling accessible. // *Proceedings of the Twelfth Language Resources and Evaluation Conference*, P 6745–6752, Marseille, France, May. European Language Resources Association.

Dallas Card, Chenhao Tan, and Noah A. Smith. 2018. Neural models for documents with metadata. // *ACL*.

Hila Chefer, Shir Gur, and Lior Wolf. 2020. Transformer interpretability beyond attention visualization. *Computing Research Repository*, arXiv:2012.09838.

Irina Deeva, Anna Bubnova, and Anna V. Kalyuzhnaya. 2023. Advanced approach for distributions parameters learning in bayesian networks with gaussian mixture models and discriminative models. *Mathematics*, 11(2).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *Computing Research Repository*, arXiv:1810.04805.

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.

Caitlin Doogan and Wray Buntine. 2021. Topic model or topic twaddle? re-evaluating semantic interpretability measures. // *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, P 3824–3848. Association for Computational Linguistics, June.

Cédric Févotte and Jérôme Idier. 2011. Algorithms for nonnegative matrix factorization with the -divergence. *Neural Computation*, 23(9):2421–2456.

Ismail Harrando, Pasquale Lisena, and Raphael Troncy. 2021. Apples to apples: A systematic evaluation of topic models. // *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, P 483–493. INCOMA Ltd., September.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Lee Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. // *Proceedings of 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*.

Maria Khodorchenko, Nikolay Butakov, and Denis Nasonov. 2022a. Towards better evaluation of topic model quality. // *2022 32nd Conference of Open Innovations Association (FRUCT)*, P 128–134.

Maria Khodorchenko, Nikolay Butakov, Timur Sokhin, and Sergey Teryoshkin. 2022b. Surrogate-based optimization of learning strategies for additively regularized topic models. *Logic Journal of the IGPL*, 02. jzac019.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *ArXiv*, abs/1905.07213.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. // Armand Prieditis and Stuart Russell, *Proceedings of the 12th International Conference on Machine Learning*, P 331–339, San Francisco (CA). Morgan Kaufmann.

Jeffrey Lund, Connor Cook, Kevin Seppi, and Jordan Boyd-Graber. 2017. Tandem anchoring: a multiword anchor approach for interactive topic modeling. // *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 896–905, Vancouver, Canada, July. Association for Computational Linguistics.

Jeffrey Lund, Piper Armstrong, Wilson Fearn, Stephen Cowley, Courtni Byun, Jordan Boyd-Graber, and Kevin Seppi. 2019. Automatic evaluation of local topic quality. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P 788–796, Florence, Italy, July. Association for Computational Linguistics.

Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. // *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, P 897–908, New York, NY, USA. Association for Computing Machinery.

Grégoire Montavon, Sebastian Bach, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2015. Explaining nonlinear classification decisions with deep taylor decomposition. *Computing Research Repository*.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. // *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, P 100–108, USA. Association for Computational Linguistics.

Mehdi Rezaee and Francis Ferraro. 2020. A discrete variational recurrent topic model without the reparametrization trick. // *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020)*. Curran Associates Inc.

Matthias Rüdiger, David Antons, Amol M. Joshi, and Torsten-Oliver Salge. 2022. Topic modeling revisited: New evidence on algorithm performance and quality metrics. *PLOS ONE*, 17:1–25, 04.

Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K. Reddy. 2018. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. // *Proceedings of the 2018 World Wide Web Conference*, WWW '18, P 1105–1114, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Runzhi Tian, Yongyi Mao, and Richong Zhang. 2020. Learning VAE-LDA models with rounded reparameterization trick. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 1315–1325, Online, November. Association for Computational Linguistics.

2023. Toloka ai: Powering data-centric ai. `https://toloka.ai/`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Computing Research Repository*, arXiv:1706.03762.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P 5797–5808, Florence, Italy, July. Association for Computational Linguistics.

Konstantin Vorontsov and Anna Potapenko. 2014. Additive regularization of topic models. *Machine Learning*, 101:1–21, 12.

Linzi Xing and Michael Paul. 2018. Diagnosing and improving topic models by analyzing posterior variability. // *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, volume 32, P 6005–6012, apr.