entropy

MDPI

Article

Estimating Topic Modeling Performance with Sharma–Mittal Entropy

Sergei Koltcov *, **Vera Ignatenko** and **Olessia Koltsova**

St. Petersburg School of Physics, Mathematics, and Computer Science, National Research University Higher School of Economics, Kantemirovskaya Ulitsa, 3A, St. Petersburg 194100, Russia

* Correspondence: skoltsov@hse.ru; Tel.: +7-911-981-9165

Received: 26 April 2019; Accepted: 3 July 2019; Published: 5 July 2019

Abstract: Topic modeling is a popular approach for clustering text documents. However, current tools have a number of unsolved problems such as instability and a lack of criteria for selecting the values of model parameters. In this work, we propose a method to solve partially the problems of optimizing model parameters, simultaneously accounting for semantic stability. Our method is inspired by the concepts from statistical physics and is based on Sharma–Mittal entropy. We test our approach on two models: probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA) with Gibbs sampling, and on two datasets in different languages. We compare our approach against a number of standard metrics, each of which is able to account for just one of the parameters of our interest. We demonstrate that Sharma–Mittal entropy is a convenient tool for selecting both the number of topics and the values of hyper-parameters, simultaneously controlling for semantic stability, which none of the existing metrics can do. Furthermore, we show that concepts from statistical physics can be used to contribute to theory construction for machine learning, a rapidly-developing sphere that currently lacks a consistent theoretical ground.

Keywords: Sharma–Mittal entropy; topic modeling; optimal number of topics; stability

1. Introduction

The Internet and, particularly, social networks generate a huge amount of data of different types (such as images, texts, or table data). A large amount of collected data becomes comparable to physical mesoscopic systems. Correspondingly, it becomes possible to use machine learning methods based on methods of statistical physics to analyze such data. Topic Modeling (TM) is a popular machine learning approach to soft clustering of textual or visual data, the purpose of which is to define the set of hidden distributions in texts or images and to sort the data according to these distributions. To date, a relatively large number of probabilistic topic models with different methods of determining hidden distributions have been developed, and several metrics for measuring the quality of topic modeling results have been formulated and investigated. The lion's share of the research on TM has focused on the use of probabilistic models [1] such as variants of Latent Dirichlet Allocation (LDA) and probabilistic Latent Semantic Analysis (pLSA); therefore, we study and provide numerical experiments for these models. Non-probabilistic algorithms, such as Non-negative Matrix Factorization (NMF), can also be applied to the task of TM [2,3]; however, NMF approaches are less popular due to their inability to produce generative models. Other problems of NMF models were described in [4,5]. At the same time, despite broad usage of probabilistic topic models in different fields of machine learning [6–9], they, too, possess a set of problems limiting their usage for big data analysis.

A fundamental problem of probabilistic TM is finding the number of components in the mixture of distributions since the parameter determining this number has to be set explicitly [10–12]. A similar problem arises for the NMF approach since factorization rank has to be chosen [4]. A well-known exception is the Hierarchical Dirichlet Process model (HDP) [13] positioned by the authors as able to select the number of topics automatically. However, this class of models possesses a set of hidden parameters, which, according to the authors themselves, can influence the results of determining hidden distributions and the optimal number of topics correspondingly. The second unsolved problem in probabilistic TM is a certain level of semantic instability resulting from the ambiguity in retrieving the multidimensional density of the mixture of distributions. This ambiguity means that different runs of the algorithm on the same source data lead to different solutions. Solutions may differ both in terms of word and text composition of the resulting topics, which is usually incompatible with reliability requirements set by TM end users. The problems that have non-unique or non-stable solutions are termed ill-posed [14]. Let us mention that NMF is also an ill-posed problem [4] since factorization is not unique. A general approach to avoiding multiple solutions is given by Tikhonov regularization [14]. The essence of regularization is to redefine prior information that allows for narrowing the set of solutions. Regularization is implemented by introducing restrictions on hidden distributions [15], by modifying the sampling procedure [16], by using a combination of conjugate functions [10], or by incorporating different types of regularization procedures into the algorithm [15,17]. However, introduction of the regularization procedure, although it may contribute to higher stability, may also lead to the problem of determining regularization coefficients of probabilistic topic models since these parameters are, again, to be set by a user explicitly. All this leads users of machine learning methods to an understandable mistrust towards the obtained results [9,12].

The above problems naturally affect the quality of TM. Currently, the main methods for determining the quality of topic models are Gibbs–Shannon entropy [18,19], Kullback–Leibler divergence [20], log-likelihood [21], the Jaccard index [22,23], semantic coherence [17], and relevance [24]. However, first, each of these metrics measures only one of the aspects of TM performance. It is known that the distribution of words, at least in European languages, satisfies the so-called Zipf law (a power-law distribution), which is characteristic of complex systems, i.e., of systems with non-Markov processes [25,26]. It is known that the most effective way to investigate the behavior of complex systems is application of mathematical formalism borrowed from the theory of non-additive systems [26]. The goal of our research is thus to propose a metric that would be able to both measure different aspects of TM performance at the same time and would be more adequate for textual complex systems. For this, we adapt the mathematical formalism of non-extensive statistical physics, namely Sharma–Mittal entropy, and apply it for the analysis of the results of machine learning methods. We show that our metric combines the functionality of several existing metrics and is embedded in a more theoretically-grounded approach.

Before passing to our research, we briefly discuss the basics of TM and introduce notations. The key idea of TM is based on an assumption that any large document collection contains a set of topics or semantic clusters, while each word and each text of such a collection belongs to each topic with a certain probability. This gives TM an important ability to co-cluster both words by topics and topics by documents simultaneously. Topics are defined as hidden distributions of both words and texts that are to be restored from the observed co-occurrences of words in texts. Mathematically, topic models are based on the following propositions [27]:

1. Let $\tilde{D}$ be a collection of textual documents with D documents and $\tilde{W}$ be a set (dictionary) of all unique words with W elements. Each document $d \in \tilde{D}$ is a sequence of terms $w_1, ..., w_n$ from dictionary $\tilde{W}$.
2. It is assumed that there is a finite number of topics, T, and each entry of a word w in document d is associated with some topic $t \in \tilde{T}$. A topic is understood as a set of words that often (in the statistical sense) appear together in a large number of documents.
3. A collection of documents is considered a random and independent sample of triples $(w_i; d_i; t_i)$, $i = 1, ..., n$, from the discrete distribution $p(w; d; t)$ on a finite probability space $\tilde{W} \times \tilde{D} \times \tilde{T}$. Words w and documents d are observable variables, and topic t is a latent (hidden) variable.

4. It is assumed that the order of words in documents is unimportant for topic identification (the ”bag of words” model). The order of documents in the collection is also not important.

In TM, it is also assumed that the probability $p(w|d)$ of the occurrence of term w in document d can be expressed as a product of probabilities $p(w|t)$ and $p(t|d)$, where $p(w|t)$ is the probability of word w under topic t and $p(t|d)$ is the probability of topic t in document d. According to the formula of total probability and the hypothesis of conditional independence, one obtains the following expression [27]: $p(w|d) = \sum_{t \in \tilde{T}} p(w|t)p(t|d) \equiv \sum_{t \in \tilde{T}} \phi_{wt}\theta_{td}$. Thus, constructing a topic model means finding the set of latent topics $\tilde{T}$, i.e., the set of one-dimensional conditional distributions $p(w|t) \equiv \phi_{wt}$ for each topic t, which constitute matrix Φ (distribution of words by topics), and the set of one-dimensional distributions $p(t|d) \equiv \theta_{td}$ for each document d, which form matrix Θ (distribution of documents by topics), based on the observable variables d and w.

One can distinguish three types of models in the literature that allow solving this problem: (1) models based on likelihood maximization; (2) models based on Monte Carlo methods; and (3) models of the hierarchical Dirichlet process. A description of these models and their limitations can be found in Appendix A. In the process of TM, for algorithms based on the Expectation-Maximization (E-M) algorithm (first type) and the Gibbs sampling algorithm (second type), transition to a strongly non-equilibrium state occurs. The initial distributions of words and documents in matrices Φ and Θ for Gibbs sampling methods are flat; however, in E-M models, the initial distribution is determined by a random number generator. For both types of algorithm, the initial distribution corresponds to the maximum entropy of the topic model. Regardless of the algorithm type and the procedure of initialization, redistribution of words and documents by topics proceeds so that a significant portion of words (about 95% of all unique words) acquires probabilities close to zero and only about 3–5% receive probabilities above a threshold $1/W$ [28]. Numerical experiments demonstrate that the number of words with high probabilities depends on the number of topics and values of model parameters, which allows constructing a theoretical approach for analyzing such dependency using the perspective of statistical physics [29].

The rest of the paper proceeds as follows. Section 2.1 reviews the standard metrics, which are used in the field of machine learning, relationships between these metrics and their differences. Section 2.2 describes our concept and basic assertions of our new method. Section 2.3 is devoted to the adaptation of Renyi entropy for the analysis of TM results. Sections 2.4 and 2.5 represent adaptation of Sharma–Mittal entropy for the analysis of TM results leading to a new quality metric in the field of TM. The relations of this new metric to the standard ones are also presented throughout Section 2. Section 3 shows numerical results of the application of our new metric to the analysis of TM outputs. We demonstrate the results of simulations run on two datasets by using two TM algorithms, namely probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA) with Gibbs sampling. In Section 3, we also demonstrate the application of several standard metrics to TM results and compare them with our new metric. Section 4 summarizes the functionality of our two-parametric entropy approach and proposes directions for future research. Appendix A contains a short discussion of topic models and a detailed description of the models that were used in numerical experiments. Appendix B contains numerical results on another metric, which is called “semantic coherence”, to the outputs of TM and demonstrates difficulties when using this metric for tuning model parameters.

2. Materials and Methods

2.1. Methods for Analyzing the Results of Topic Modeling

The results of TM depend on the parameters of models, such as “number of topics”, hyper-parameters of Dirichlet distributions, or regularization coefficients, since these parameters are included explicitly in the mathematical formulation of the model. In the literature on TM, the most frequently-used metrics for analyzing topic models are the following.

1. Shannon entropy and relative entropy. Shannon entropy is defined according to the following equation [19,30,31]: $H = -\sum_{i=1}^{n} p(x_i)\log(p(x_i))$, where $p(x_i)$, $i = 1, ..., n$ are distribution probabilities of a discrete random value with possible values $\{x_1, ..., x_n\}$. Relative entropy is defined as follows [32]: $D_{KL}(p|q) = \sum_i p(x_i)\log(\frac{p(x_i)}{q(x_i)}) = -\sum_i p(x_i)\log(q(x_i)) + \sum_i p(x_i)\log(p(x_i))$, i.e., $D_{KL}(p|q)$ is the difference of cross-entropy $H(p,q) = -\sum_i p(x_i)\log(q(x_i))$ and Shannon entropy. Relative entropy is also known as Kullback–Leibler (KL) divergence. In the field of statistical physics, it was demonstrated that KL divergence is closely related to free energy. In the work [33], it was shown that in the framework of Boltzmann–Gibbs statistics, KL divergence can be expressed as follows: $D_{KL}(p|\tilde{p}) = q(F(p) - F(\tilde{p}))$, where p is the probability distribution of the system residing in the non-equilibrium state, $\tilde{p}$ is the probability distribution of the system residing in the equilibrium state, $q = 1/T$, T is the temperature of the system, and F is the free energy. Hence, KL divergence is nothing but the difference between the free energies of off-equilibrium and equilibrium. The difference between free energies is a key characteristic of the entropy approach [29], which is to be discussed further below in Sections 2.2 and 2.3. The variant of KL divergence used in TM is also discussed in Paragraph 3 of this section.
2. Log-likelihood and perplexity: One of the most-used metrics in TM is the log-likelihood, which can be expressed through matrices Φ and Θ in the following way [21,34]: $\ln(P(\tilde{D}|\Phi,\Theta)) = \sum_{d=1}^{D}\sum_{w=1}^{W} n_{dw}\ln(\sum_{t=1}^{T}\phi_{wt}\theta_{td})$, where n_{dw} is the frequency of word w in document d. A better model will yield higher probabilities of documents, on average [21]. In addition, we would like to mention that the procedure of log-likelihood maximization is a special case of minimizing Kullback–Leibler divergence [35]. Another widely-used metric in machine learning, and in TM, particularly, is called perplexity. This metric is related to likelihood and is expressed as: perplexity $= \exp(-\ln(P(\tilde{D}|\Phi,\Theta))/\sum_{d=1}^{D} n_d)$, where n_d is the number of words in document d. Perplexity behaves as a monotone decreasing function [36]. The score of perplexity is the lower the better. In general, perplexity can be expressed in terms of cross-entropy as follows: perplexity $= 2^{\text{entropy}}$ or perplexity $= e^{\text{entropy}}$ [37], where "entropy" is cross-entropy. The application of perplexity for selecting values of model parameters was discussed in many papers [10,17,21,34,38,39]. In a number of works, it was demonstrated that perplexity behaves as a monotonously-decreasing function of the number of iterations, which is why perplexity has been proposed as a convenient metric for determining the optimal number of iterations in TM [11]. In addition, the authors of [12] used perplexity for searching the optimal number of topics. However, the use of perplexity and log-likelihood has some limitations, which were demonstrated in [40]. The authors showed that perplexity depends on the size of vocabulary of the collection for which TM is implemented. The dependence of the perplexity value on the type of topic model and the size of the vocabulary was also demonstrated in [41]. Hence, comparison of topic models for different datasets and in different languages by means of perplexity is complicated. Many numerical experiments described in the literature demonstrate monotone behavior of perplexity as a function of the number of topics. Unlike the task of determining the number of iterations, the task of finding the number of topics is sensitive to this feature, and fulfillment of the latter task appears to be complicated by it. In addition, calculation of perplexity and log-likelihood is extremely time consuming, especially for large text collections.
3. Kullback–Leibler divergence: Another measure, that is frequently used in machine learning, is the Kullback–Leibler divergence (KL) or relative entropy [32,42,43]. However, in the field of TM, symmetric KL divergence is most commonly used. This measure was proposed by Steyvers and Griffiths [20] for determining the number of stable topics: $KL(i,j) = \frac{1}{2}\sum_{w=1}^{W}\phi'_{wi}\log_2(\frac{\phi'_{wi}}{\phi''_{wj}}) + \frac{1}{2}\sum_{w=1}^{W}\phi''_{wj}\log_2(\frac{\phi''_{wj}}{\phi'_{wi}})$, where ϕ' and ϕ'' correspond to topic-word distributions from two different runs; i and j are topics. Therefore, this metric measures dissimilarity between topics i and j. Let us note that KL divergence is calculated for the same words in different topics; thus, the semantic

component of topic models is taken into account. This metric can be represented as a matrix of size $T \cdot T$, where T is the number of topics in compared topic models. The minimum of $KL(i,j)$ characterizes the measure of similarity between topics i and j. If $KL(i,j) \approx 0$, then topics i and j are semantically identical. An algorithm for searching for the number of stable topics for different topic models was implemented [17] based on this measure. In this approach, pair-wise comparison for all topics of one topic's solution with all topics of another topic solution was done. Hence, if the topic is stable from the semantic point of view, then it reproduces regularly for each run of TM. In [16], it was shown that different types of regularization lead to different numbers of stable topics for the same dataset. The disadvantage of this method is that this metric does not allow comparing one topic solution with another as a whole, but one can only obtain a set of pair-wise compared word distributions for separate topics. No generalization of this metric for solution-level comparisons has been offered yet.

4. The Jaccard index and entropy distance: Another widely-used metric in the field of machine learning is the Jaccard index, also known as the Jaccard similarity coefficient, which is used for comparing the similarity and diversity of sample sets. The Jaccard coefficient is defined as the cardinality of the intersection of the sample sets divided by the cardinality of the union of the sample sets [23]. Mathematically, it is expressed as follows. Assume that we have two sets X and Y. Then, one can calculate the following values: a is the number of elements of X, which are absent in Y; b is the number of elements of Y, which are absent in X; c is the number of common elements of X and Y. The Jaccard coefficient is $J = \frac{c}{a+b+c}$, where $c = |X \cap Y|$, $|X \cup Y| = a + b + c$, $|\cdot|$ is the cardinality of a set. The Jaccard coefficient $J = 1$ if sets are totally similar and $J = 0$ if sets are totally different. This coefficient is used in machine learning due to the following reasons. Kullback–Leibler divergence characterizes similarity based on the probability distribution. This means that two topics are similar if words' distributions for them have similar values. At the same time, the Jaccard coefficient demonstrates the number of identical words in topics, i.e., it reflects another point of view of the similarity of topics. The combination of two similarity measures allows for deeper analysis of TM results. In addition, the Jaccard distance is often used, which is defined as [22]: $J(X,Y) = 1 - \frac{c}{a+b+c}$. This distance equals zero if sets are identical. The Jaccard distance also plays an important role in computer science, especially, in research on "regular language" [44,45] and is related to entropy distance as follows [22]: $D_H(X,Y) = 1 - I(X,Y)/H(X,Y) = J(X,Y) = 1 - J$, where $D_H(X,Y)$ is entropy distance, $I(X,Y)$ is the mutual information of X and Y, and $H(X,Y)$ is the joint entropy of X and Y. In the standard set-theoretic interpretation of information theory, the mutual information corresponds to the intersection of sets X and Y and the joint entropy to the union of X and Y, and hence, the entropy distance corresponds to the Jaccard distance [22]. Correspondingly, if $J(X;Y) = 0$, then $D_H(X,Y) = 0$ as well. The paper proposes to use the Jaccard coefficient as a parameter of entropy, but not for TM tasks, while we incorporate it into our two-parametric entropy approach to TM specifically.
5. Semantic coherence: This metric was proposed to measure the interpretability of topics and was demonstrated to correspond to human coherence judgments [17]. Topic coherence can be calculated as follows [17]: $C(t, W(t)) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log(\frac{D(v_m^t, v_l^t)+1}{D(v_l^t)})$, where $W(t) = (v_1^t, ..., v_M^t)$ is a list of M most probable words in topic t, $D(v)$ is the number of documents containing word v, and $D(v, v')$ is the number of documents where words v and v' co-occur. The authors of [17] proposed to consider the following values of $M = 5, ..., 20$. To obtain a single coherence score of a topic solution, one needs to aggregate obtained individual topic coherence values. In the literature, one can find that aggregation can be implemented by means of the arithmetic mean, median, geometric mean, harmonic mean, quadratic mean, minimum, and maximum [46]. Coherence can also be used for determining the optimal number of topics; however, in paper [47], it was demonstrated that the coherence score monotonously decreases if the number of topics increases.

6. Relevance: This is a measure that allows users of TM to rank terms in the order of their usefulness for topic interpretation [24]. This measure is similar to a measure proposed in [48], where a term's frequency is combined with the exclusivity of the word (exclusivity is the degree to which a word's occurrences are limited to only a few topics). The relevance of term w to topic t given a weight parameter λ $(0 \leq \lambda \leq 1)$ can be expressed as: $r(w, k|\lambda) = \lambda \cdot \log(\phi_{wt}) + (1-\lambda)\log(\frac{\phi_{wt}}{p_w})$, where λ determines the weight given to ϕ_{wt} relative to its lift and p_w is the empirical term probability, which can be calculated as: $p_w = \frac{\sum_{d=1}^{D} n_{dw}}{\sum_{d=1}^{D} n_d}$ with n_{dw} being a count of how many times the term w appears in document d and n_d being total term-count in document d, namely, $n_d = \sum_w n_{dw}$. The authors of [24] proposed to take the default value of $\lambda = 0.6$ according to their user study; however, in general, it is not clear how to chose the optimal value of λ for a particular dataset. Furthermore, relevance is a topic-level measure that cannot be generalized for an entire solution, which is why it is not used further in this research.

2.2. Minimum Cross-Entropy Principles in Topic Modeling

As was shown above, TM parameter estimation and assessment of semantic stability are separate processes based on several unrelated metrics. Therefore, it is necessary to develop a single approach that would include a number of metrics and would allow solving simultaneously two problems, namely optimization of both semantic stability and other parameters. Such an approach can be developed on the basis of the cross-entropy minimum principle (minimum of KL divergence). In doing so, this principle can be implemented in two ways: (1) by constructing an entropic metric and searching for the minimum of this metric under variation of different topic model parameters, where TM is conducted using standard algorithms; (2) by creating an algorithm of restoring hidden distributions based on cross-entropy minimization. A version of the TM algorithm, close to the second approach, was considered in [49], where symmetric KL divergence was added to the model based on log-likelihood maximization. However, this model included regularization using only matrix Θ, and one has to set explicitly the regularization coefficient (the parameter called η). In our work, we consider only the first approach, i.e., searching for optimal parameters of the topic model based on the entropy metric, which takes into account the distribution of words by topics and the semantic stability of topics under the condition of the variation of different model parameters. By the "optimal" number of topics for a dataset, we mean the number of topics that corresponds to human judgment. We propose a method for tuning topic models, which is based on the following assertions [29,50], which create a linkage between TM and statistical physics and reformulate the problem of model parameter optimization in terms of thermodynamics: (1) A collection of documents is considered a mesoscopic information system: a statistical system where the elements are words and the documents number in the millions. Correspondingly, the behavior of such a system can be studied by application of models from statistical physics. (2) The total number of words and documents in the information system under consideration is constant (i.e., the system volume is not changed). (3) A topic is a state (an analogue of spin direction) that each word and document in the collection can take. Here, a word and a document can belong to different topics (spin states) with different probabilities. (4) A solution of topic modeling is a non-equilibrium state of the system. (5) Such information system is open and exchanges energy with the environment via changing the temperature. Here, the temperature of the information system is the number of topics that is a parameter and should be selected by searching for a minimum KL divergence. (6) Since KL divergence is proportional to the difference of free energies, to measure the degree to which a given system is non-equilibrium, one can use the following expression: $\Lambda_F = F(T) - F_0$, where F_0 is the free energy of the initial state (chaos) of the topic model and $F(T)$ is the free energy after TM for a fixed number of topics T [50]. (7) The minimum of Λ_F depends on topic model parameters such as the number of topics and other hyper-parameters. (8) The optimal number of topics and the set of optimal hyper-parameters of the topic model correspond to the situation when the information maximum (in terms of non-classical entropy) is reached. If one does not take semantic stability into account, then the information maximum corresponds to the Renyi entropy minimum [29].

However, in our work, we aim to consider the semantic stability of topics; hence, the information maximum will depend on the semantic component.

It is known that in topic models, the sum of probabilities of all words equals the number of topics $T = \sum_{t=1}^{T} \sum_{w=1}^{W} p_{wt}$, where $p_{wt} \in [0,1]$ for all $w = 1,..,W$; $t = 1,...,T$. In the framework of statistical physics, it is common to investigate the distribution of statistical systems by energy levels, where energy is expressed in terms of probability. In accordance with such approach, we divide the range of probabilities $[0,1]$ by a fixed number of intervals, determine energy levels corresponding to these intervals, and then seek the number of words belonging to each energy level. Let us note that these values depend on the number of topics and the values of the hyper-parameters of a topic model. Division into intervals is convenient from a computational point of view. If the lengths of such intervals tend to zero, the distribution of words by intervals will tend to the probability density function. However, for simplification, we will consider a two-level system, where the first level corresponds to words with high probabilities and the second level corresponds to words with small probabilities close to zero. Therefore, we introduce the density-of-states function for words with high probabilities under a fixed number of topics and a fixed set of parameters: $\rho = N/(WT)$, where N is the number of words with high probabilities. By high probability, we mean the probability satisfying: $p > 1/W$. The choice of such a level is informed by the fact that the values $1/W$ are the initial values of matrix Φ for a topic model. The value $W \cdot T$ determines the total number of micro-states of the topic model (the size of matrix Φ), and normalizes the density-of-states function. During the process of TM, the probabilities of words redistribute with respect to the above threshold $1/W$. A small part of the words has probabilities higher than the threshold level, while the larger part of words has probabilities lower than that. The energy of the upper level containing states with high probabilities is expressed as follows:

$$E = -\ln(\tilde{P}) = -\ln\left(\frac{1}{T}\sum_{wt}(p_{wt} \cdot \Omega(p_{wt} - 1/W))\right), \tag{1}$$

where the step function $\Omega(\cdot)$ is defined by $\Omega(p_{wt} - 1/W) = 1$ if $p_{wt} \geq 1/W$ and $\Omega(p_{wt} - 1/W) = 0$ if $p_{wt} < 1/W$. Therefore, in Equation (1), we sum only the probabilities that are greater than $1/W$. The energy of the lower level is expressed analogously, except that summing occurs for probabilities that are smaller than $1/W$. A level is characterized by two parameters: (1) the normalized sum of probabilities of micro-states, that lie in the corresponding interval, $\tilde{P}$; (2) the normalized number of micro-states (density-of-states function), ρ, whose probabilities lie in this interval. Let us note that the density-of-states function is sometimes called the statistical weight of a complex system's level. For a two-level system, the main contribution to the entropy and energy of the whole system is made by the states with high probabilities, that is mainly by the upper level. Respectively, the free energy of the whole system is almost entirely determined by the entropy and the energy of the upper level. The free energy of a statistical system can be expressed through Gibbs–Shannon entropy and the internal energy in the following way [51]: $F = E - TS = E - S/q$, where $q = 1/T$. The entropy of such a system can be expressed through the number of micro-states belonging to the same level [52]: $S = \ln(N)$. It follows that the difference of free energies of the topic model is expressed through $\tilde{P}$ and ρ in the following way:

$$\begin{aligned}\Lambda_F =& F(T) - F_0 = (E(T) - E_0) - (S(T) - S_0)T \\ =& -\ln(\tilde{P}) - T\ln(\rho),\end{aligned} \tag{2}$$

where E_0 and S_0 are the energy and the entropy of the initial state of the system, with $E_0 = -\ln(T)$ and $S_0 = \ln(WT)$. Hence, the degree to which a given system is non-equilibrium can be defined as the difference between the two free energies and expressed in terms of experimentally-determined values ρ and $\tilde{P}$. Values ρ and $\tilde{P}$ were calculated for each topic model under variation of parameter T and hyper-parameters, i.e., Λ_F is a function of the number of topics T, hyper-parameters, and size of vocabulary W.

2.3. Renyi Entropy of the Topic Model

Using partition function:

$$Z_q = e^{-q\Lambda_F} = e^{-qE+S} = \rho(\tilde{P})^q, \tag{3}$$

$q = 1/T$ [53], one can express Renyi entropy in Beck notation through free energy [54] and through experimentally-determined values ρ and $\tilde{P}$:

$$S_q^R = \frac{\ln(Z_q)}{q-1} = \frac{\ln(e^{-q\Lambda_F})}{q-1} = \frac{-q\Lambda_F}{q-1} = \frac{q\ln(\tilde{P}) + \ln(\rho)}{q-1}, \tag{4}$$

where, again, $q = 1/T$. The choice of entropy in Beck notation is determined by the following considerations. Firstly, constructing topic models with just one or two topics is meaningless in terms of their informativeness for end users. Correspondingly, the entropy of such a model should be large. Secondly, excessive increase of the number of topics leads to a flat distribution of words by topics that, again, should lead to a large value of entropy. Thirdly, both q and Z_q calculated for words with high probabilities are less than one. Correspondingly, if we normalize this value by $1-q$, we will obtain a negative value of Renyi entropy. Taking into account the necessity to have maximum entropies at the boundaries of the range of the number of topics, the normalization coefficient $q-1$ should be used. Summing up the advantages of Renyi entropy application to TM, the following can be said. First, since calculation of Renyi entropy is based on the difference of free energies (i.e., on KL divergence or relative entropy), it is convenient to use Renyi entropy as a measure of the degree to which a given system is in non-equilibrium, and this is what we do in our approach. Second, Renyi entropy, in contrast to Gibbs–Shannon entropy, allows taking into account two different processes: a decrease in Gibbs–Shannon entropy and an increase in internal energy, both of which occur with the growth of the number of topics. The difference between these two processes can have an area of balance when two processes counterbalance each other. In this area, Renyi entropy reaches its minimum. Third, the search for the Renyi entropy minimum (i.e., minimum of KL divergence) can be convenient for optimizing regularization coefficients in topic modeling. As mentioned above, a relative drawback of Renyi entropy here is the impossibility of taking into account the semantic component of topic models since it is expressed only through the density-of-states function and energy of the level. However, this drawback can be overcome by using two-parametric Sharma–Mittal entropy, where one of deformation parameters is taken as $q = 1/T$ and the second deformation parameter corresponds to the semantic component of a topic model.

2.4. Sharma–Mittal Entropy in Topic Modeling

Sharma–Mittal two-parametric entropy proposed in [55] has been discussed in many works [56–58]. The main emphasis in these papers was made on the investigation of its mathematical properties [56,59,60] or application of this entropy when constructing generalized non-extensive thermodynamics [61]. In the field of machine learning, Sharma–Mittal entropy is used in a few works, for instance, in [62]. Two-parametric Sharma–Mittal entropy can be written as:

$$S_{SM} = \frac{(\sum_i p_i^q)^{(1-r)/(q-1)} - 1}{1-r}, \tag{5}$$

where r and q are deformation parameters. The essence of deformation parameters r and q for TM can be determined based on consideration of limit cases. One can show that $\lim_{r\to 1} S_{SM} = S_q^R$ and $\lim_{r\to 0} S_{SM} = \exp(S_q^R) - 1$. Since in TM, deformation parameter q can be defined through the number of topics ($q = 1/T$), in order to use Sharma–Mittal entropy for the purposes of TM, one has to define the meaning of parameter r. Let us note that $r \in [0;1]$ according to [55]. In addition, if $r \to 1$, then Sharma–Mittal entropy transforms into Renyi entropy; hence, in this case, the quality of topic model is defined only by Renyi entropy and deformation parameter q, i.e., by the number of topics.

If $r \to 0$, then the value of entropy becomes large since $\lim_{r\to 0} S_{SM} = \exp(S_q^R) - 1$. Based on the principle that maximum entropy corresponds to the information minimum, we conclude that the minimum value of parameter r corresponds to the minimum information and maximum entropy. Taking into account that entropy can be parameterized by the Jaccard coefficient and that semantic distance between two topic solutions can be estimated by entropy distance, we define r as a parameter being responsible for the semantic stability of the topic model under variation of the number of topics or hyper-parameters. Therefore, we define the value of r as equal to the value of the Jaccard coefficient (i.e., $r := J$, where J is the Jaccard coefficient calculated for the sets of the most probable words for each pair of topic solutions). Consequently, $1 - r = J(W', W'')$ is the entropy distance or Jaccard distance, where W' and W'' are the sets of the most probable words of the first topic solution and the second topic solution, correspondingly.

2.5. Sharma–Mittal Entropy for a Two-Level System

Based on Equations (4) and (5) and the statistical sum (3), the Sharma–Mittal entropy of the topic model in terms of experimentally-determined values ρ and $\tilde{P}$ can be defined as:

$$S_{SM} = \frac{Z_q^{(1-r)/(q-1)} - 1}{1 - r} = \frac{(\tilde{P}^q \rho)^{(1-r)/(q-1)} - 1}{1 - r}. \qquad (6)$$

On the one hand, application of Sharma–Mittal entropy allows estimating the optimal values of topic model parameters, such as hyper-parameters, and the number of topics, by means of searching for the minimum entropy, which, in turn, is characterized by the difference of entropies between the initial distribution and the distribution obtained after TM. On the other hand, it allows estimating the contribution of the semantic difference between any two topic solutions that, in turn, is influenced by values of hyper-parameters and the number of topics. Hence, the optimal values of topic model parameters correspond to the minimum Sharma–Mittal entropy, and the worst values of parameters correspond to the maximum entropy.

3. Results

3.1. Data and Computational Experiments

For our numerical experiments, the following datasets were used:

- Russian dataset (from the Lenta.ru news agency): a publicly-available set of 699,746 news articles in the Russian language dated between 1999 and 2018 from the Lenta.ru online news agency (available at [63]). Each news item was manually assigned to one of ten topic classes by the dataset provider. We considered a class-balanced subset of this dataset, which consisted of 8624 news texts (containing 23,297 unique words). It is available here at [64]. Below, we provide statistics on the number of documents with respect to categories (Table 1).

Table 1. Statistics on the Russian dataset.

Category	Number of Documents
business	466
culture	499
economy and finance	667
incidents	712
media	628
policy	1231
security services	863
science and tech	580
society and travel	1957
sports	1022

Some of these topics are strongly correlated with each other. Therefore, the documents in this dataset can be represented by 7–10 topics.

- English dataset (the well-known "20 Newsgroups" dataset http://qwone.com/~jason/20Newsgroups/): 15,404 English news articles containing 50,948 unique words. Each of the news items belonged to one or more of 20 topic groups. Since some of these topics can be unified, 14–20 topics can represent the documents of this dataset [65]. This dataset is widely used to test machine learning models.

We conducted our numerical experiments using pLSA and LDA with Gibbs sampling. These models represent two different types of algorithms. The LDA model used here was based on the Gibbs sampling procedure, and the pLSA model was based on the E-M algorithm. A detailed description of these models can be found in Appendix A. Experiments on these models allowed us to estimate the usability of Sharma–Mittal entropy for two main types of algorithms. Topic modeling was conducted using the following software implementation: the package "BigARTM" (http://bigartm.org) was used for pLSA; GibbsLDA++ (http://gibbslda.sourceforge.net) for LDA (Gibbs sampling). All source codes were integrated into a single package "TopicMiner" (https://linis.hse.ru/en/soft-linis) as a set of dynamic link libraries. Each model was calculated under variation of the number of topics in the range of $[2; 50]$ in increments of one topic, and for LDA model, also values of hyper-parameters α and β were varied in the range of $[0; 1]$ in increments of 0.1 for each dataset. For each model and for each dataset, the following metrics were calculated: (1) log-likelihood; (2) Jaccard index; (3) Sharma–Mittal entropy; (4) semantic coherence.

3.1.1. Results for the pLSA Model

The choice of pLSA model was determined by the fact that this model has only one parameter: the number of topics. Correspondingly, we can isolate the effect of this parameter on the values of the above metrics. Figure 1 plots the log-likelihood as a function of the number of topics for both datasets. One can see that increasing the number of topics led to a smooth increase of the log-likelihood. Thus, these curves did not allow determining the optimal number of topics due to the absence of any clear extrema. The difference between these two curves resulted from different sizes of vocabularies and the different amounts of documents in the corresponding datasets.

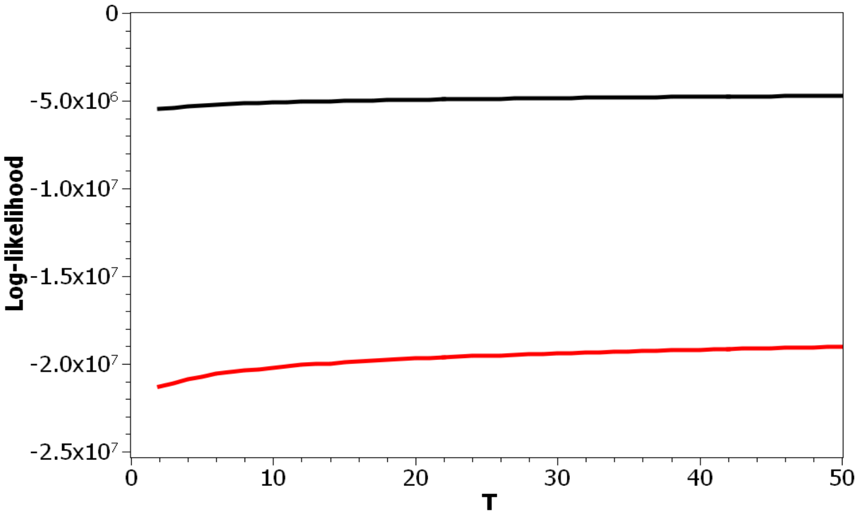


Figure 1. Log-likelihood distribution over T (probabilistic Latent Semantic Analysis (pLSA)). Russian dataset, black; English dataset, red.

Figure 2 demonstrates Renyi entropy curves for the pLSA model on both datasets. The entropy was calculated according to Equation (4). The exact minimum of Renyi entropy for the Russian dataset was seven and for the English dataset 16. However, as was noted, being an ill-posed problem, topic modeling produced different results on different runs of the same algorithm, which was especially true for pLSA. From the previous research [29], it is known that the range of such variation between the runs is approximately ±3 topics. Therefore, it makes more sense to look at the range of the neighboring minima rather than at the exact minimum. It can be seen that the numbers of topics defined by humans, when corrected for inter-topic correlation, lied within the discovered ranges in both datasets, which suggests the language-independent character of this metric (at least for European languages). As Renyi entropy does not include an instrument to evaluate the semantic stability of topic models, we calculated Jaccard coefficients under variation of the number of topics. Figure 3 presents a "heat map" of Jaccard coefficients for the dataset in the Russian language. The matrix containing Jaccard coefficients was symmetric with respect to the main diagonal, and this is the reason why only half of this matrix is depicted. The structure of the "heat map" of the Jaccard index for the English dataset was similar to that for the Russian dataset and can be found in [29].

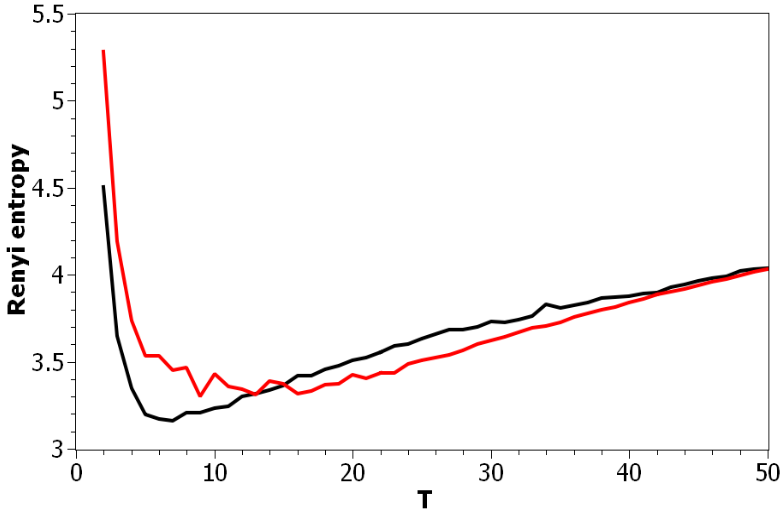


Figure 2. Renyi entropy distribution over the number of topics T (pLSA). Russian dataset, black; English dataset, red.

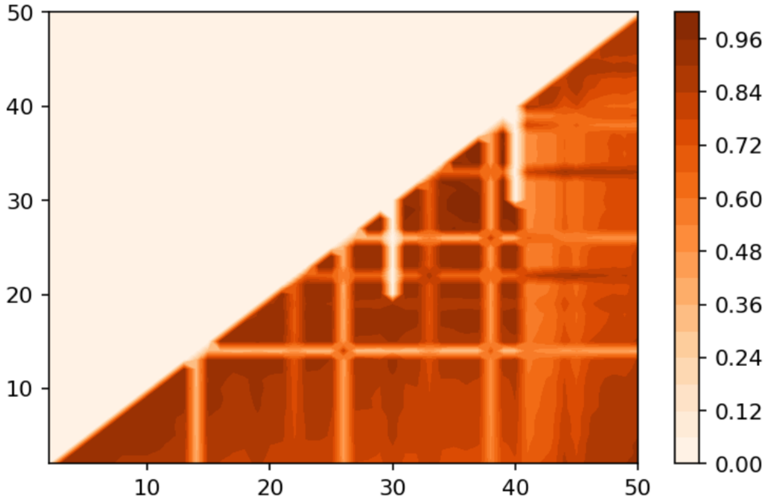


Figure 3. Heat map of the Jaccard index for the Russian dataset (pLSA).

Figure 4 presents a pairwise comparison of topic solutions with the number of topics equal to T and $T+1$ correspondingly, under variation of T for the Russian and English datasets. As demonstrated in Figures 3 and 4, there are areas of sharp decreases in semantic similarity between topic solutions with different numbers of topics. In order to incorporate the "density-of-states" function, the probabilities of words, and semantic similarity under variation of the parameter "number of topics", we calculated Sharma–Mittal entropy according to Equation (6) for the pLSA model on both datasets.

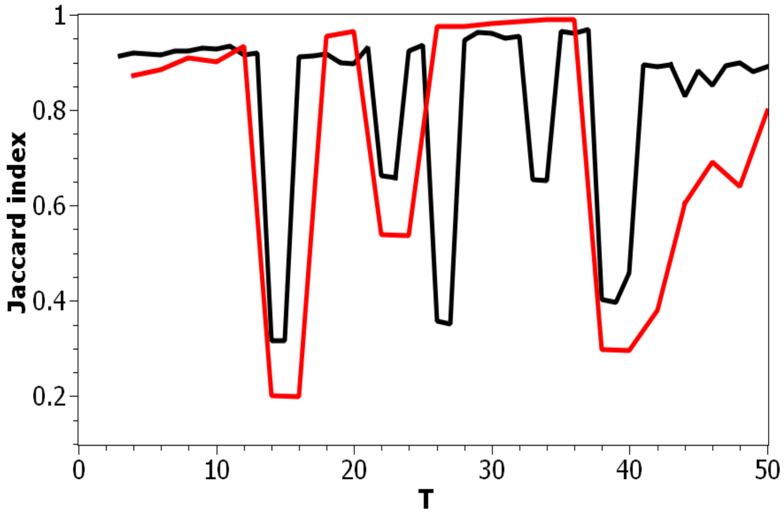


Figure 4. Distribution of Jaccard coefficients of the pairwise comparison for neighboring topic solutions with the number of topics T and $T+1$ (pLSA). Russian dataset, black; English dataset, red.

Figure 5 plots Sharma–Mittal entropy as a function of the number of topics calculated only on the data from pairwise comparisons of topic solutions with the neighboring values of T (i.e., T and $T+1$). The values of the Jaccard index used for this calculation constitute over-diagonal elements taken from the full matrix of pairwise comparisons of all topic solutions in the range $T = [2; 50]$. Figures 6 and 7 demonstrate Sharma–Mittal entropy for the Russian and English datasets where large values (≥ 5) were replaced by five to make the global minimum more visible.

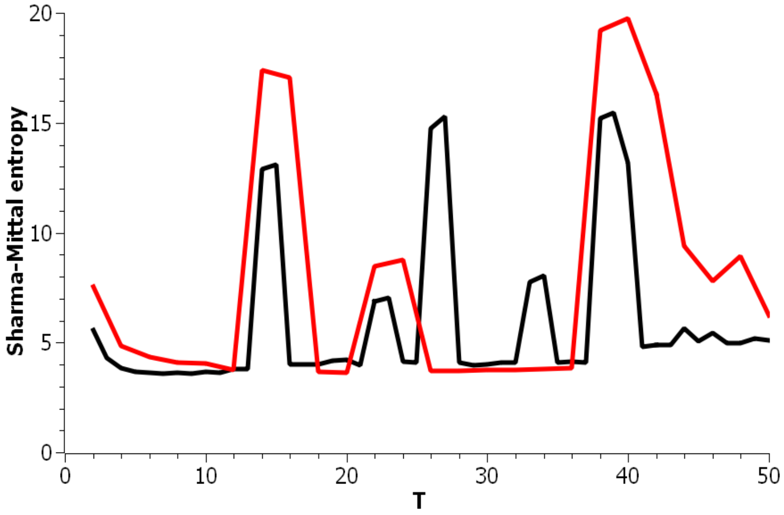


Figure 5. Sharma–Mittal entropy distribution over the number of topics T (pLSA). Russian dataset, black; English dataset, red.

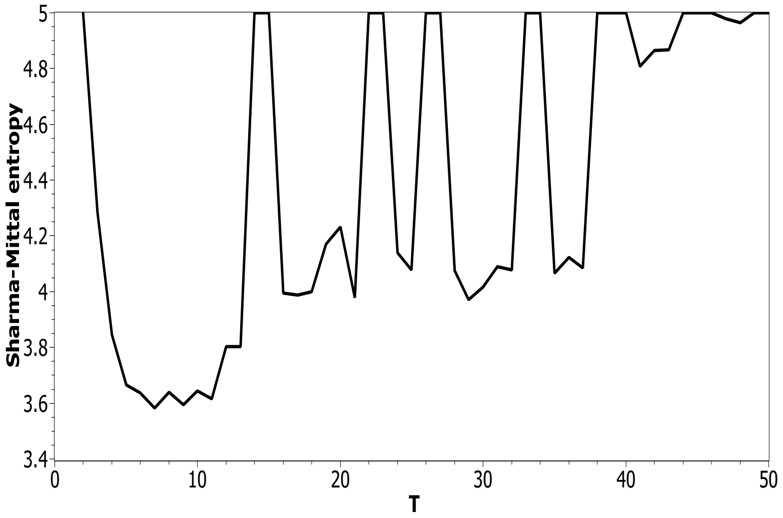


Figure 6. Sharma–Mittal entropy distribution over T with $S_{SM} > 5$ reduced to five (Russian dataset). pLSA.

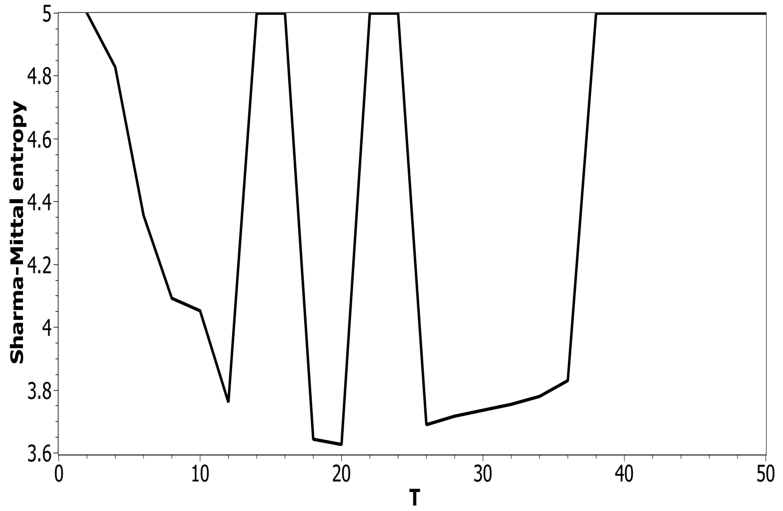
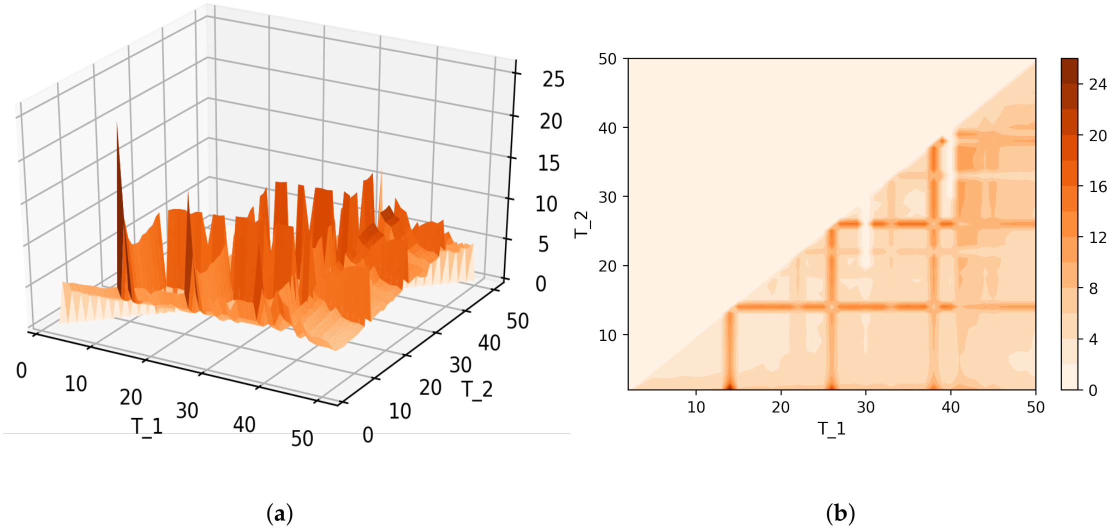


Figure 7. Sharma–Mittal entropy distribution over T with $S_{SM} > 5$ reduced to five (English dataset). pLSA.

Figure 8 shows Sharma–Mittal entropy for the pLSA model in two versions: a 3D picture and its view from above. Together, they show that Sharma–Mittal entropy has areas of minima and maxima, the overall shape of the curve being determined by the number of topics and the local fluctuations resulting from the fluctuations of the Jaccard distance. In practice, however, we propose to consider only two-dimensional versions of this figure (e.g., Figure 6), where the Jaccard index is calculated only for the neighboring solutions. Such plots are easier to interpret, and at the same time, they demonstrate the influence of semantic stability. The exact values of the Sharma–Mittal entropy minimum are the following: $T = 20$ for the English dataset and $T = 7$ for the Russian dataset. Horizontal shift of the Sharma–Mittal entropy minimum as compared to the Renyi entropy minimum on the English dataset is an effect of the sharp fall of the Jaccard coefficient observed in the range of 14–16 topics. It follows that application of Sharma–Mittal entropy for models based on the E-M algorithm allows determining the optimal number of topics involving the semantic stability of topics. Figures that demonstrate the behavior of semantic coherence for these datasets can be found in Appendix B. We do not provide them here since they monotonously decrease, with some fluctuations, but without any clear extrema, thus providing no criteria for choosing topic number.



(**a**) (**b**)

Figure 8. Sharma–Mittal entropy for the pLSA model (Russian dataset). (**a**) 3D plot of Sharma–Mittal entropy; (**b**) projection of Sharma–Mittal entropy to OT_1T_2.

3.1.2. Results for the LDA with Gibbs Sampling Model

The difference between the pLSA model and LDA Gibbs sampling model is not only in the application of the Monte Carlo algorithm for determining hidden distributions, but also in the presence of a regularization procedure. The level of regularization in LDA with Gibbs sampling is determined by hyper-parameters α and β. In our numerical experiments, we used the algorithm [11] where hyper-parameters of the LDA model were fixed and did not change from iteration to iteration since our goal was to analyze the results of the LDA model with respect to different values of hyper-parameters. Figure 9 plots the log-likelihood for the Russian dataset as a function of T for pLSA and for LDA with different fixed values of α or β. The behavior of the log-likelihood for the English dataset was similar to that for the Russian dataset, and therefore, we do not provide the figure.

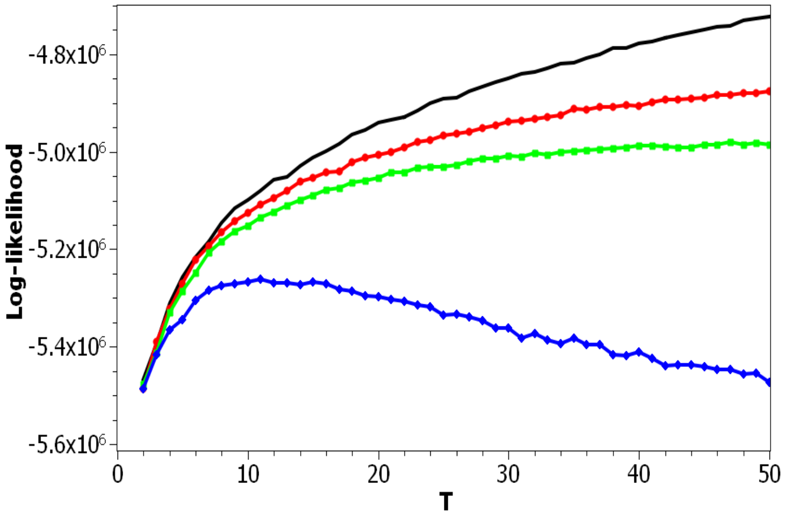


Figure 9. Log-likelihood distribution over T for different α and β (Russian dataset). pLSA, black; LDA (α = 0.1, β = 0.1), red; LDA (α = 0.5, β = 0.1), green; LDA (α = 1, β = 1), blue.

Using the results of calculations (Figure 9), one can conclude that the log-likelihood metric allows estimating the effect of regularization in the LDA Gibbs sampling model. Namely, it can be seen that the largest values of regularization coefficients (blue curve) led to the lowest values of the log-likelihood, while according to [21,34], the optimal topic model should correspond to the maximum log-likelihood. According to our numerical results, the maximum log-likelihood corresponds to the pLSA model, that is to the zero regularization of LDA. Let us note that a similar result was obtained in [66], where, according to human mark-up, pLSA was shown to perform better than LDA, as regularized pLSA, and than pLSA regularized with decorrelation and sparsing-smoothing approaches, for the task of revealing ethnicity-related topics.

Figures 10 and 11 plot Renyi entropy as functions of T for different values of α and β for the Russian and English datasets. Calculations demonstrated that application of Renyi entropy and the log-likelihood allowed estimating the influence of regularization in TM. Namely, larger regularization coefficients led to higher entropy, i.e., to the model's deterioration. The exact minima of Renyi entropy were the following: (1) Russian dataset: $T = 7$ for $\alpha = 0.1, \beta = 0.1$; $T = 9$ for $\alpha = 0.5, \beta = 0.1$; $T = 14$ for $\alpha = 1, \beta = 1$; (2) English dataset: $T = 17$ for $\alpha = 0.1, \beta = 0.1$; $T = 15$ for $\alpha = 0.5, \beta = 0.1$; $T = 13$ for $\alpha = 1, \beta = 1$. It follows that Renyi entropy is useful for estimating topic model hyper-parameters for different datasets, at least in European languages. In addition, Renyi entropy is less sensitive to the size of vocabulary since this metric is normalized with respect to initial states (chaos). However, as Renyi entropy for the LDA Gibbs sampling model and pLSA model does not allow taking into account semantic stability, we further do not present our results on Sharma–Mittal entropy.

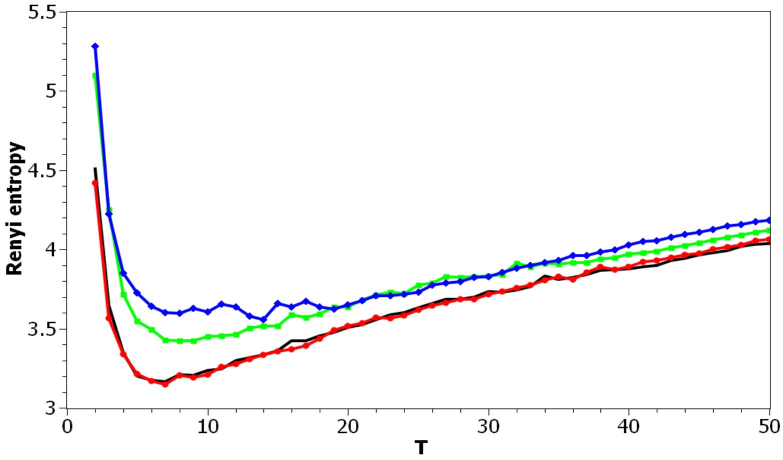


Figure 10. Renyi entropy distribution over T for different α and β (Russian dataset). pLSA—black, LDA (α = 0.1, β = 0.1)—red, LDA (α = 0.5, β = 0.1)—green, LDA (α = 1, β = 1)—blue.

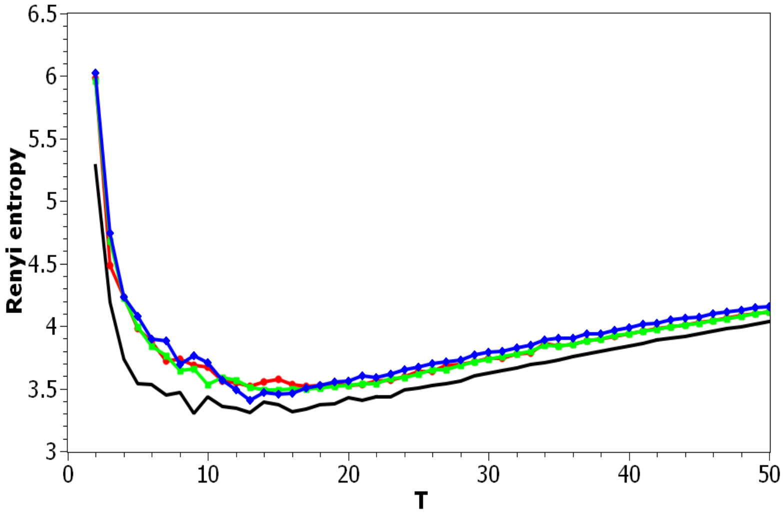


Figure 11. Renyi entropy distribution over T for different α and β (English dataset). pLSA, black; LDA (α = 0.1, β = 0.1), red; LDA (α = 0.5, β = 0.1), green; LDA (α = 1, β = 1), blue.

Figures 12 and 13 show curves of Sharma–Mittal entropy for the LDA Gibbs sampling model under variation of hyper-parameters α and β for the Russian and English datasets. Figures 14 and 15 demonstrate Sharma–Mittal entropy curves where large values ($\geq$6) are replaced by six in order to demonstrate clearly the location of the global minimum. These figures show that for small values of hyper-parameters, the behavior of Sharma–Mittal entropy for LDA is similar to that for the pLSA model. The exact minima of Sharma–Mittal entropy were: (1) Russian dataset: $T = 7$ for $\alpha = 0.1, \beta = 0.1$; $T = 7$ for $\alpha = 0.5, \beta = 0.1$; $T = 19$ for $\alpha = 1, \beta = 1$; (2) English dataset: $T = 21$ for $\alpha = 0.1, \beta = 0.1$; $T = 21$ for $\alpha = 0.5, \beta = 0.1$; $T = 13$ for $\alpha = 1, \beta = 1$ Furthermore, these figures demonstrate that the location of jumps of Sharma–Mittal entropy, which are related to semantic stability, are almost independent of the regularization coefficients. However, in general, entropy curves were lifted along the Y axis if regularization coefficients increased. It follows that for LDA Gibbs sampling, the optimal values of both α and β coefficients were small. It can be concluded that the results of regularization coefficients' selection by means of Sharma–Mittal entropy were similar to those obtained with the log-likelihood and Renyi entropy; however, two-parametric entropy, unlike other considered metrics, allowed incorporating semantic stability using the Jaccard distance. Sharma–Mittal entropy under variation of the number of topics and incorporation of the Jaccard coefficient represents a three-dimensional structure with a set of local minima, which are determined by the number of topics and by semantic stability. These areas of local minima represent islands of stability. Figures 16 and 17 demonstrate the three-dimensional surfaces of Sharma–Mittal entropy for the Russian and English datasets and its projections to the horizontal plane OT_1T_2.

Numerical results on semantic coherence for LDA with Gibbs sampling can be found in Appendix B (Figures A3 and A4). However, as with pLSA, this metric fell monotonously and did not provide any criteria for the choice of the topic number.

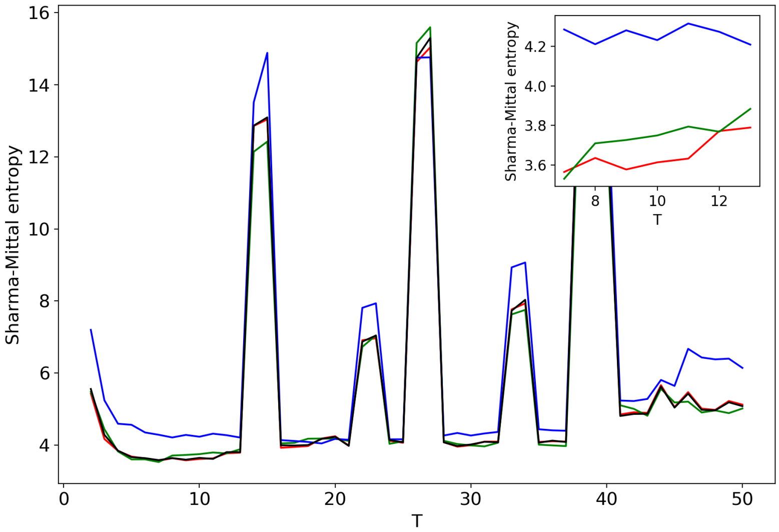


Figure 12. Sharma–Mittal entropy distribution over topics (Russian dataset). pLSA, black; LDA (α = 0.1, $\beta = 0.1$), red; LDA ($\alpha = 0.5$, $\beta = 0.1$), green; LDA ($\alpha = 1$, $\beta = 1$), blue.

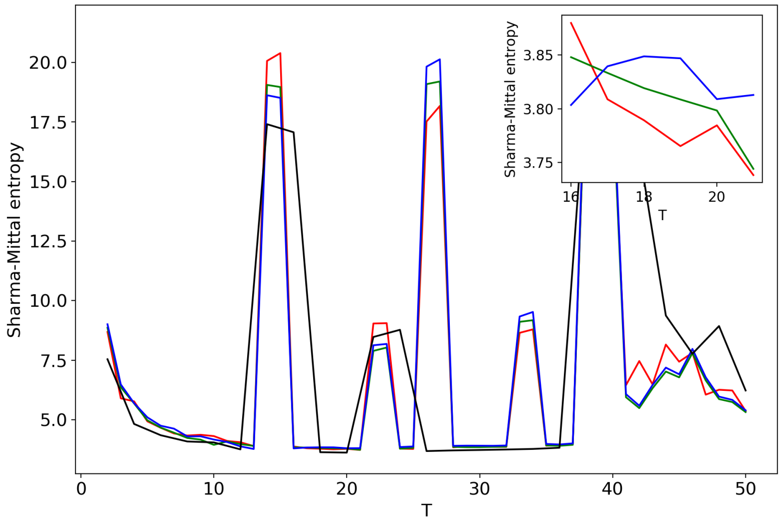


Figure 13. Sharma–Mittal entropy distribution over topics (English dataset). pLSA, black; LDA ($\alpha = 0.1$, $\beta = 0.1$), red; LDA ($\alpha = 0.5$, $\beta = 0.1$); green, LDA ($\alpha = 1$, $\beta = 1$), blue.

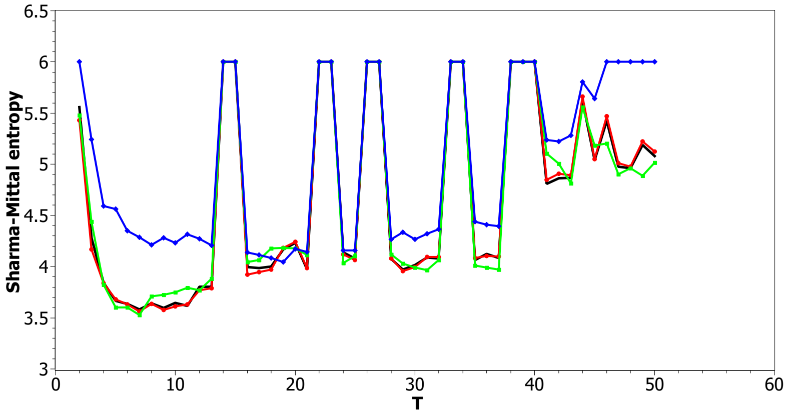


Figure 14. Sharma–Mittal entropy distribution over topics (Russian dataset). pLSA, black; LDA (α = 0.1, $\beta = 0.1$), red; LDA ($\alpha = 0.5$, $\beta = 0.1$), green; LDA ($\alpha = 1$, $\beta = 1$), blue.

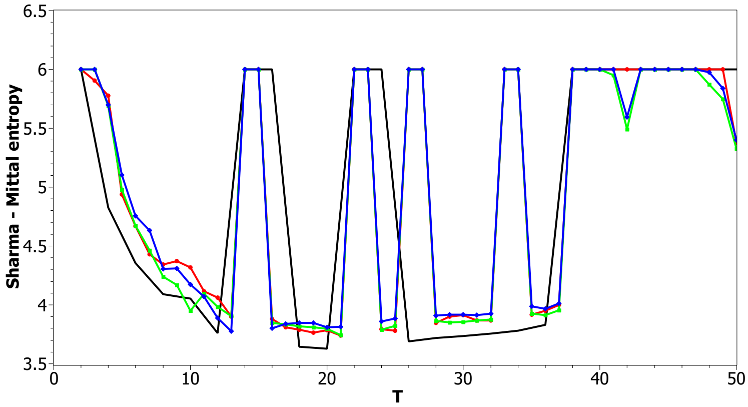
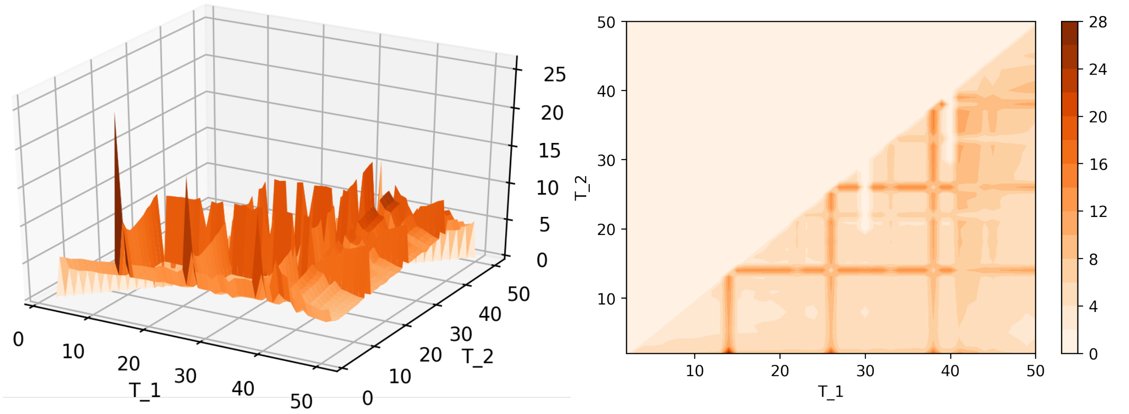
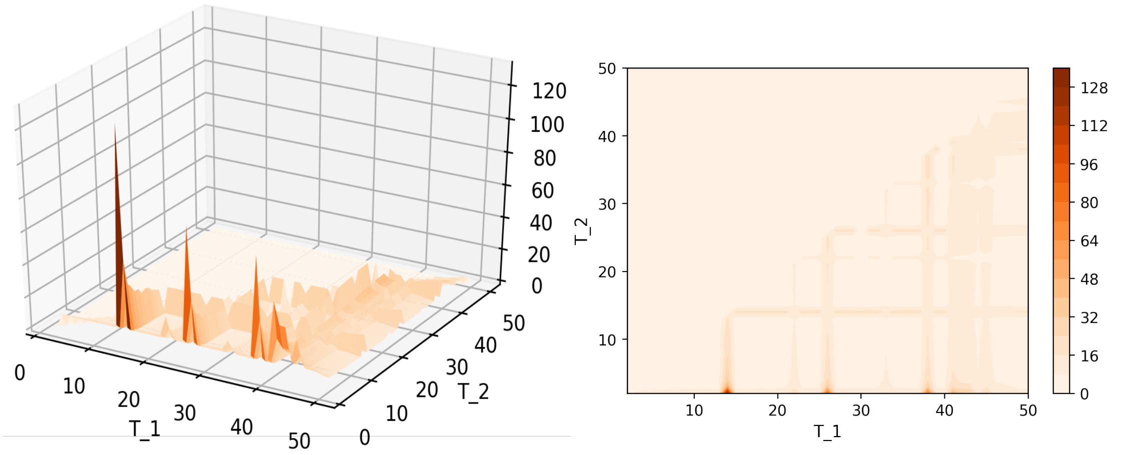


Figure 15. Sharma–Mittal entropy distribution over topics (English dataset). pLSA, black; LDA ($\alpha = 0.1$, $\beta = 0.1$), red; LDA ($\alpha = 0.5$, $\beta = 0.1$), green; LDA ($\alpha = 1$, $\beta = 1$), blue.



(**a**) (**b**)

Figure 16. Sharma–Mittal entropy for the LDA model (Russian dataset). (**a**) 3D plot of Sharma–Mittal entropy; (**b**) projection of Sharma–Mittal entropy to OT_1T_2.



(**a**) (**b**)

Figure 17. Sharma–Mittal entropy for the LDA model (English dataset). (**a**) 3D plot of Sharma–Mittal entropy; (**b**) projection of Sharma–Mittal entropy to OT_1T_2.

4. Discussion

In this work, we proposed a new entropy-based approach for the multi-aspect evaluation of the performance of topic models. Our approach was based on two-parametric Sharma–Mittal entropy, that is twice deformed entropy. We considered the deformation parameter, q, being the inverse value of the number of topics, and the second parameter, r, being the Jaccard coefficient, while $1 - r$ the entropy distance. Our numerical experiments demonstrated that, firstly, Sharma–Mittal entropy, as well as Renyi entropy allowed determining the optimal number of topics. Secondly, as the minimum of Sharma–Mittal entropy corresponded to the maximum of the log-likelihood, the former also allowed choosing the optimal values of hyper-parameters. Thirdly, unlike Renyi entropy or the log-likelihood, it allowed optimizing both hyper-parameters and the number of topics, simultaneously accounting for semantic stability. This became possible due to the existence of areas of semantic stability that have been shown to be characterized by low values of Sharma–Mittal entropy. According to our numerical results, the location of such areas did not depend on the hyper-parameters. However, on the whole, larger values of hyper-parameters in the LDA Gibbs sampling led to higher entropy, while small values made the LDA model almost identical to pLSA. This means that new methods of regularization are needed that would not impair TM performance in terms of entropy. We concluded that Sharma–Mittal entropy is an effective metric for the assessment of topic models performance since it includes the functionality of several metrics.

However, our approach had certain limitations. First of all, topic models have an obvious drawback, which is expressed by the fact that the probabilities of words in topics depend on the number of documents containing these words. This means that if a topic is represented in a small number of documents, then the topic model will assign small probabilities to the words of this topic, and correspondingly, a user will not be able to see this topic. Thus, topic models can detect topics that are represented in many documents and poorly identify topics with a small number of documents. Therefore, Renyi entropy and Sharma–Mittal entropy allow determining the number of those large topics only. Secondly, in our work, Sharma–Mittal entropy was tested only for two European languages, while there are papers on the application of topic models for the Chinese, Japanese, and Arabic languages. Correspondingly, our research should be extended and tested on non-European languages. Thirdly, our metric allowed finding the global minimum when topic modeling was performed in a wide range of the number of topics; however, this process was resource-intensive and in practice can be applied to datasets containing up to 100–200 thousand documents. For huge datasets, this metric is not applicable. This problem might be partially solved by means of renormalization, which can be adapted for topic models from statistical physics. Research on application of renormalization for fast search of Renyi entropy and Sharma–Mittal entropy minima deserves a separate paper. Fourthly, we would like to note that our method was not embedded in algorithms of topic modeling. Therefore, in future research, the metric of quality based on Sharma–Mittal entropy can be used for the development of new topic models. Sharma–Mittal entropy can be embedded in the algorithms based on the Gibbs sampling procedure, where walks in the multi-dimensional space of words, hyper-parameters, and the number of topics will be determined by the level of this entropy. Correspondingly, transition along different axes of multi-dimensional space can be guided by the entropy minimization principle. An algorithm similar to the algorithm of annealing based on searching for the minimum Tsallis entropy [67] can be used in this case. However, unlike the algorithm proposed by Tsallis, one can use deformation parameter q as a parameter that controls the number of components in the mixture of distributions and search for the minimum when changing the number of components. Therefore, the walk in the multi-parameter space can be determined by the direction of the minimum of deformed entropy when changing the dimension of the space.

For topic models based on the maximum log-likelihood principle, the sizes of matrices are included in the model as external parameters, which are selected by the user. Correspondingly, new topic models can be developed in the future by using the principle of deformed logarithm maximization, where one of deformation parameters corresponds to the sizes of matrices (namely, the number of topics) and the other parameter corresponds to semantic stability (e.g., the Jaccard index). Note that both parameters here are maximization parameters. A more detailed discussion of these possible directions for research is out of the scope for this paper and can be used as a starting point for new research.

Author Contributions: Conceptualization, S.K. and V.I.; methodology, S.K.; software, S.K.; validation, O.K. and V.I.; formal analysis, S.K., O.K. and V.I.; investigation, V.I.; resources, O.K.; data curation, S.K.; writing, original draft preparation, S.K. and V.I.; writing, review and editing, O.K. and V.I.; visualization, S.K.; supervision, S.K. and O.K.; project administration, S.K.; funding acquisition, O.K.

Funding: This research was funded by the Center for Basic Research of the National Research University Higher School of Economics in 2018.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

E-M	expectation-maximization
HDP	Hierarchical Dirichlet Process
KL	Kullback–Leibler
LDA	Latent Dirichlet Allocation
pLSA	probabilistic Latent Semantic Analysis
TM	Topic Modeling

Appendix A. Types of Topic Models

Appendix A.1. Models Based on Likelihood Maximization

Mathematical formulation of this family of models is based on the fact that matrix containing distributions of words by documents, F, is represented as a product of Φ and Θ [15,27], so the problem $F = \Phi\Theta$ is considered as a problem of stochastic matrix decomposition, which is carried out using the Expectation-Maximization (E-M) algorithm. Logarithm of likelihood is maximized for searching the approximation of solution [10]. Let us note that stochastic matrix decomposition is not unique and is defined with accuracy up to a non-degenerate transformation: $F = \Phi\Theta = (\Phi R)(R^{-1}\Phi) = \Phi'\Theta'$ [15]. It means that different topic solutions with the same number of topics can be assigned to the initial set of words and documents (matrix F). The elements of matrices Φ and Θ can differ under variation of matrix of transformation R. It follows that the problem of TM is ill-posed. Nowadays, many models with different types of regularization exist. One of the most used regularizer found in literature is a product of conjugate distributions, namely, multinomial and Dirichlet distributions. In this case, the final distribution of words by topics and distribution of topics by documents are Dirichlet distributions [10] in accordance with properties of conjugate distributions. Another variant of widely used regularization is additive regularization developed by Vorontsov [15]. In the framework of this approach, a set of functions characterizing the variant of regularization is added to the logarithm of likelihood in the framework of maximization problem, the level of regularization is defined by the value of regularization coefficient. For example, Gibbs–Shannon entropy or Kullback–Leibler divergence can play a role of regularizer. Despite the large number of possibilities of this approach, the method of additive regularization does not assist in choosing regularization parameters and selecting the combination of regularizers [68]. Basically, the selection of regularization coefficients is carried out manually taking into account the perplexity stabilization [68]. Generally, the problem of selecting regularizers and their coefficient values is still in the research core for this type of models. Alternative variant of regularization is a model taking into account the relations between words. Such information is taken externally (for instance, from the dataset) and is expressed in form of covariance matrix C of size $W \times W$, where W is the number of unique words. A significant disadvantage of this type of regularization is the problem of calculating the covariance matrix, since the size of the dictionary of unique words can exceed one million of words. However, in our point of view, this type of regularization is potentially perspective since "word embedding" method is actively developed in the framework of neural networks. This method allows calculating word co-occurrence matrices which, in turn, can be incorporated in topic models. Nowadays, there are two hybrids of topic models and "word embedding" model [69,70]. However, these models also possess instability since "word embedding" algorithms possess instability [71,72].

Let us consider the Probabilistic Latent Semantic Analysis (PLSA) model, which is used in our numerical experiments, in detail. In the framework of this model, the determination of the matrices Φ and Θ is performed as described in [27]. The entire dataset is generated as:

$$p(D) = \prod_{d \in D} \prod_{w \in W} p(d,w)^{n(d,w)} = \prod_{d \in D} \prod_{w \in W} p(d)^{n(d,w)} p(w|d)^{n(d,w)}$$
$$= \prod_{d \in D} \prod_{w \in W} p(d)^{n(d,w)} \sum_{t \in T} p(w|t)^{n(d,w)} p(t|d)^{n(d,w)}$$

where $p(d,w)$ is the joint probability distribution, $n(d,w)$ counts the appearance frequency of the term w in the document d. Note that this model involves a conditional independence assumption, namely, d and w are independently conditioned on the state of the associated latent variable [27].

The estimation of the one-dimensional distributions is based on log-likelihood maximization with linear constraints:

$$L(\phi, \theta) = \sum_{d \in D} \sum_{w \in W} n(d,w) \ln \left(p(d) \sum_{t \in T} \phi_{wt} \theta_{td} \right) \to \max_{\phi, \theta} L(\phi, \theta),$$

where $\phi_{wt} \geq 0, \sum_{w \in W} \phi_{wt} = 1, \theta_{td} \geq 0, \sum_{t \in T} \theta_{td} = 1$.

The determination of the local maximum of $L(\phi, \theta)$ is carried out using Expectation-Maximization (E-M) algorithm. The initial approximation of ϕ_{wt} and θ_{td} is chosen randomly or uniformly before the first iteration.

E-step: using Bayes' rule, conditional probabilities $p(t|d,w)$ are calculated for all $t \in T$ and each $w \in W, d \in D$ [73], namely:

$$p(t|d,w) = \frac{p(d,w|t)p(t)}{p(d,w)} = \frac{p(d|t)p(w|t)p(t)}{p(d) \sum_{s \in T} p(w|s)p(s|d)}$$
$$= \frac{p(w|t)p(t|d)}{\sum_{s \in T} p(w|s)p(s|d)} = \frac{\phi_{wt} \theta_{td}}{\sum_{s \in T} \phi_{ws} \theta_{sd}}.$$

M-step: using conditional probabilities, new approximations of ϕ_{wt}, θ_{td} are estimated, namely:

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

Thus, alternating E and M steps in a cycle, $p(t|d)$ and $p(w|t)$ can be estimated. Note that this model has no additional parameters except of "the number of topics", which defines the size of matrices Φ, Θ.

Appendix A.2. Models Based on Monte-Carlo Methods

This class of models represents a variant of Potts model adapted for text analysis. Each document of the collection is considered a one-dimensional grid, where a node is a word. The number of the states of the spin is considered as the number of topics. The difference between topic model and Potts model is that in TM a large amount of documents is used. Probabilities of distribution of words and documents can be estimated by means of expectation under condition of known integrand. According to approach of Blei [10], probability density functions of multinomial and Dirichlet distributions are used as integrand. Determining the hidden distributions is implemented by means of Gibbs sampling. Due to the presence of a set of local minima and maxima of integrand, this type of models also possess a certain degree of instability. One can also introduce regularizers for these types of models, for example, by fixing belongings of some words in a range of topics [74]. This type of regularization behaves as process of crystallization, where a layer of words (which are often found together with words from the core) is formed around the core containing fixed words. Another variant of regularization represents a modified Gibbs-sampling procedure, where sampling is implemented not for one word, but for several words, which are placed inside a window of fixed size [16]. As demonstrated in experiments, this variant of regularization gives a high level of stability, however, there are a lot of "garbage topics", which can not be interpreted. One can claim that the problem of optimization of regularization procedure for models based on Gibbs sampling and determining the optimal number of topics is not completely solved.

Let us consider Latent Dirichlet Allocation (LDA) model with Gibbs sampling procedure in detail. LDA is a topic model, in which each topic is smoothed by the same regularizer in the form of Dirichlet function [11]. According to Blei et al. [10], it is assumed to use Dirichlet distributions with one-dimensional parameters β and α, correspondingly, in order to simplify the derivation of analytical expressions for the matrices Φ and Θ. In LDA, documents are generated by picking a distribution over topics θ from a Dirichlet distribution with parameter α, then the words in the document are generated by picking a topic t from this distribution and then picking a word from that topic according to probabilities which are determined by $\phi_{\cdot t}$ [11], where ϕ_{cdott} is drawn from a Dirichlet distribution with parameter β. On this basis, the probability of the i-th word in a given document d is defined as follows [11]:

$$p(w_{i,d}) = \sum_{j=1}^{T} p(w_{i,d}|z_{i,d} = j)p(z_{i,d} = j) = \sum_{j=1}^{T} \phi_{wj}\theta_{dj} = \sum_{j=1}^{T} \frac{c_{d,j} + \alpha}{\sum_{j=1}^{T} c_{d,j} + \alpha T} \cdot \frac{c_{w,j} + \beta}{\sum_{w=1}^{W} c_{w,j} + \beta W},$$

where $z_{i,d}$ is a latent variable (topic), $p(w_{i,d}|z_i = j)$ is the probability of the word w_i in document d under the j-th topic, $p(z_{i,d} = j)$ is the probability of choosing a word from topic j in the current document d, $w_{i,d}$ is the i-th word in document d, counter $c_{d,j}$ is the number of words in document d assigned to topic j, counter $c_{w,j}$ is the number of word w is assigned to topic j; $\sum_{j=1}^{T} c_{d,j}$ is the total number of words in document d (i.e., length of document d), $\sum_{w=1}^{W} c_{w,j}$ is the total number of words assigned to topic j. Correspondingly, θ and ϕ can be obtained as follows:

$$\theta_{dj} = \frac{c_{d,j} + \alpha}{\sum_{j=1}^{T} c_{d,j} + \alpha T}, \tag{A1}$$

$$\phi_{wj} = \frac{c_{w,j} + \beta}{\sum_{w=1}^{W} c_{w,j} + \beta W}. \tag{A2}$$

The algorithm of calculation consists of three phases. The first one is the initialization of matrices, counters and parameters α and β, in addition to specifying the number of iterations. Counters, which define the initial values of matrices Φ and Θ, are set as constants. So, matrices are filled with constants, for example, Φ can be filled with uniform distribution, where all elements of the matrix are equal to $1/W$, where W is the number of unique words in a collection of documents.

The second phase (sampling procedure) is an exhaustive search through all the documents and all words in each document in a cycle. Each word w_i in a given document d is matched with the topic number, which is generated as follows:

$$p(z_i = j|z_{-i}) \approx \frac{c_{d,j}^{-i} + \alpha}{\sum_{j=1}^{T} c_{d,j}^{-i} + \alpha T} \cdot \frac{c_{w_i,j}^{-i} + \beta}{\sum_{w=1}^{W} c_{w,j}^{-i} + \beta W},$$

where $c_{d,j}^{-i}$ is the number of words from document d assigned to topic j not including the current word w_i, $c_{w,j}^{-i}$ is the number of instances of word w assigned to topic j not including the current instance i,

$c_{d,j}^{-i}$ and $c_{w,j}^{-i}$ are called counters. Here, the probabilities of belonging of the current word to different topics are calculated, then z_i is sampled according to this distribution. The initial probability of word-topic matching is defined only by $1/W$ when considering a uniform distribution as the initial approximation of matrix Φ. However, after each word matching to a topic, the values of counters change and, hence, after an important number of iterations, counters contain the full statistics of document collection under study.

At the third phase, Φ and Θ are calculated according to the Equations (A1) and (A2). Finally, the matrices are ready for manual analyses, where for sociological analysis, only the most probable words and documents for each topics are considered. Note that the coefficients α and β defining Dirichlet distribution are parameters of this model, which one has to select. The hyper-parameter β determines whether topics will have more sparse or more uniform distributions over words [21]. The hyper-parameter α determines the level of sparsity of vectors $\theta_{\cdot d}$. If $\alpha = 1$ then Dirichlet distribution transforms into uniforms distribution while small values of α cause more sparse vectors $\theta_{\cdot d}$. Therefore, in general, the hyper-parameters α and β influence the sparsity of matrices Φ and Θ [34]. The sparsity of matrices influences, in turn, the number of topics, which can appear in a document collection. Consequently, the number of topics may implicitly depend on the values of hyper-parameters. Work [11] suggests a rule to select hyper-parameters: $\alpha = 50/T$ and $\beta = 0.01$, where T is the number of topics. Such values of parameters were widely used in different studies [75–77].

Appendix A.3. Models Based on Hierarchical Dirichlet Process

Alternative approach in TM is hierarchical model based on Dirichlet processes (HDP) [13]. In paper [78], a two-level version of hierarchical Dirichlet process (with Split-Merge Operations) based on Chinese Restaurant Franchise (CRF) is used. According to this paper, Chinese Restaurant Franchise is associated to topic model in the following way: “restaurant” corresponds to “document”; customer corresponds to “word”; “dish” corresponds to “topic”. In this approach, “customers” are partitioned at the group-level and “dishes” are partitioned at the top level. The customer partition represents the per-document partition of words; the top level partition represents the sharing of topics between documents [78]. Let us note that the list of dishes (topics) is the same for all restaurants. Despite the fact, that this type of models is referred to the class of non-parametric methods in literature, this model has a set of pre-defined parameters, which influence on the results and on the number of topics.

Appendix B. Numerical Results on Semantic Coherence

Appendix B.1. PLSA

In order to calculate semantic coherence, we considered 30 top-words for each topic of topic solutions for the two datasets from Section 3. Figure A1 demonstrates behavior of individual topic coherence for topic solution on 30 topics for the Russian and English datasets, where topics are ordered in descending order with respect to the coherence. It is not obvious how to separate “good” and “bad” topics for the Russian dataset since topic coherence change is nearly smooth. For the English dataset one can see a dramatic decrease in the region of 25 topics, however it does not correspond to the mark-up of this dataset.

Figure A2 demonstrates aggregated semantic coherence for topic solutions with the different topic numbers T for the Russian and English datasets. This figure does not allow us to choose the “optimal” number of topics since the maximum coherence for the Russian dataset corresponds to $T = 4$, however, it is not close to the human mark-up. For the English dataset we observe a number of peaks that does not assist in selecting the number of topics. It follows that semantic coherence does not allow us to determine the single “optimal” number of topics for pLSA model.

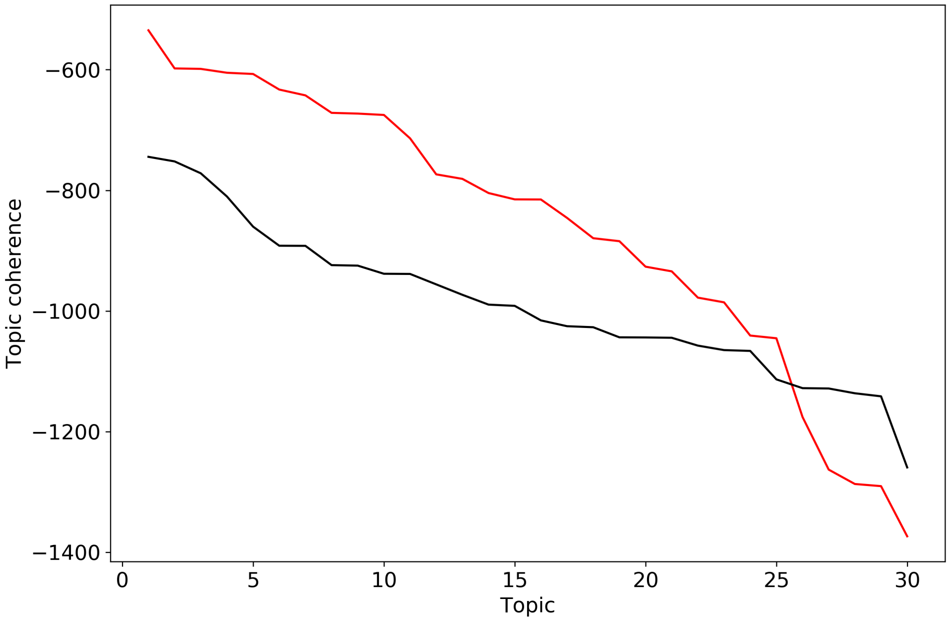


Figure A1. Topic coherence for solution on 30 topics (pLSA). Russian dataset, black; English dataset, red.

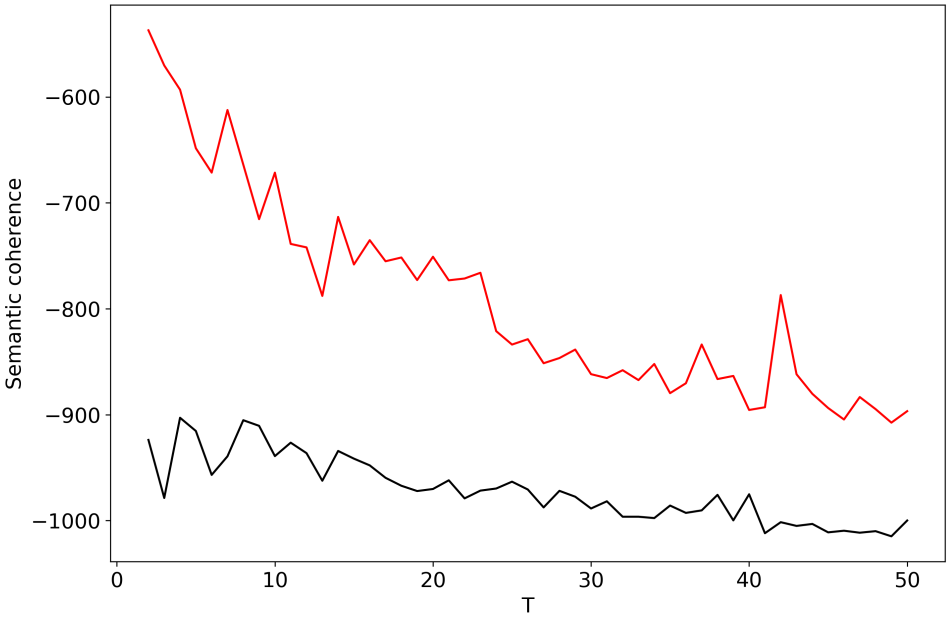


Figure A2. Distribution of semantic coherence over T (pLSA). Russian dataset, black; English dataset, red.

Appendix B.2. LDA with Gibbs Sampling

For calculation of semantic coherence for LDA models we, again, considered 30 top-words for each topic. Figure A3 demonstrates behavior of individual topic coherences for topic solutions on 30 topics for the Russian and English datasets, where topics are ordered in descending order with respect to the coherence. However, it is not obvious how to choose the optimal number of topics, i.e., where to cut the line in order to separate "good" and "bad" topics for the Russian dataset. For the English dataset we observe a sharp fall for $T = 25$, however this number of topics does not correspond to the description of the dataset. Figure A4 demonstrates aggregated semantic coherence for topic solutions with different topic numbers T for the Russian and English datasets. One can see that the maximum is reached for $T = 4$ for the Russian dataset which does not correspond to the human annotation. For the English dataset one can see a peak for $T = 19$, however this pick is not unique and it is not obvious which one we should choose. Thus, we have demonstrated limitations of semantic coherence as a method for selecting the number of topics.

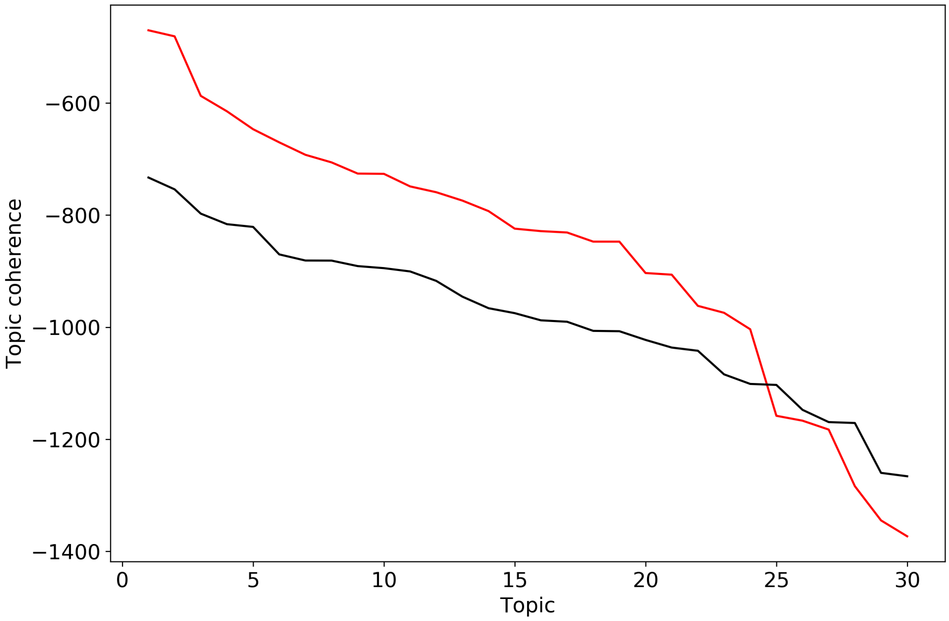


Figure A3. Topic coherence for solution on 30 topics. LDA ($\alpha = 0.1, \beta = 0.1$). Russian dataset, black; English dataset, red.

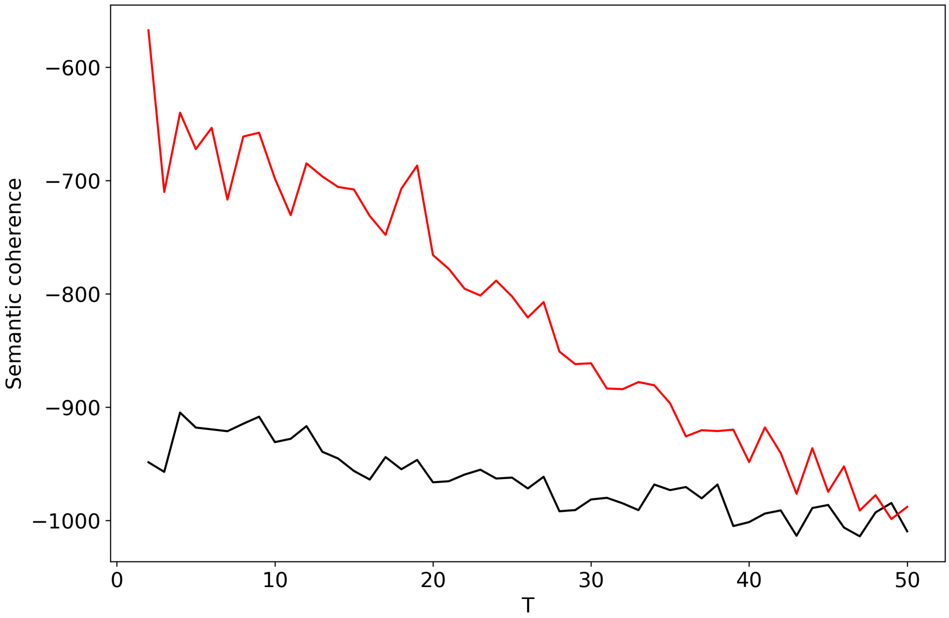


Figure A4. Distribution of semantic coherence over T. LDA ($\alpha = 0.1, \beta = 0.1$). Russian dataset, black; English dataset, red.

References

1. Greene, D.; O'Callaghan, D.; Cunningham, P. How Many Topics? Stability Analysis for Topic Models. In *Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 498–513.
2. Arora, S.; Ge, R.; Moitra, A. Learning Topic Models–Going Beyond SVD. In Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science, New Brunswick, NJ, USA, 20–23 October 2012.
3. Wang, Q.; Cao, Z.; Xu, J.; Li, H. Group Matrix Factorization for Scalable Topic Modeling. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, OR, USA, 12–16 August 2012.
4. Gillis, N. The Why and How of Nonnegative Matrix Factorization. *arXiv* **2014**, arXiv:1401.5226.
5. Gaussier, E.; Goutte, C. Relation Between PLSA and NMF and Implications. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, 15–19 August 2005.
6. Roberts, M.; Stewart, B.; Tingley, D. Navigating the local modes of big data: The case of topic models. In *Computational Social Science: Discovery and Prediction*; Cambridge University Press: New York, NY, USA, 2016.
7. Chernyavsky, I.; Alexandrov, T.; Maass, P.; Nikolenko, S.I. A Two-Step Soft Segmentation Procedure for MALDI Imaging Mass Spectrometry Data. In Proceedings of the German Conference on Bioinformatics 2012, GCB 2012, Jena, Germany, 20–22 September 2012; pp. 39–48.
8. Tu, N.A.; Dinh, D.L.; Rasel, M.K.; Lee, Y.K. Topic Modeling and Improvement of Image Representation for Large-scale Image Retrieval. *Inf. Sci.* **2016**, *366*, 99–120. [CrossRef]
9. Chang, J.; Boyd-Graber, J.; Gerrish, S.; Wang, C.; Blei, D.M. Reading Tea Leaves: How Humans Interpret Topic Models. In Proceedings of the 22nd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009; pp. 288–296.
10. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
11. Griffiths, T.L.; Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5228–5235. [CrossRef]
12. Agrawal, A.; Fu, W.; Menzies, T. What is wrong with topic modeling? And how to fix it using search-based software engineering. *Inf. Softw. Technol.* **2018**, *98*, 74–88. [CrossRef]
13. Teh, Y.W.; Jordan, M.I.; Beal, M.J.; Blei, D.M. Hierarchical Dirichlet Processes. *J. A Stat. Assoc.* **2006**, *101*, 1566–1581. [CrossRef]
14. Tikhonov, A.N.; Arsenin, V.Y. *Solutions of Ill-Posed Problems*; V. H. Winston & Sons: Washington, DC, USA, 1977.
15. Vorontsov, K.V. Additive regularization for topic models of text collections. *Dokl. Math.* **2014**, *89*, 301–304. [CrossRef]
16. Koltsov, S.; Nikolenko, S.; Koltsova, O.; Filippov, V.; Bodrunova, S. Stable Topic Modeling with Local Density Regularization. In *Internet Science: Third International Conference*; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; Volume 9934, pp. 176–188.
17. Mimno, D.; Wallach, H.M.; Talley, E.; Leenders, M.; McCallum, A. Optimizing Semantic Coherence in Topic Models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011; pp. 262–272.
18. Zhao, W.; Chen, J.J.; Perkins, R.; Liu, Z.; Ge, W.; Ding, Y.; Zou, W. A heuristic approach to determine an appropriate number of topics in topic modeling. In Proceedings of the 12th Annual MCBIOS Conference, Little Rock, AR, USA, 13–14 March 2015.
19. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
20. Steyvers, M.; Griffiths, T. Probabilistic Topic Models. In *Handbook of Latent Semantic Analysis*; Landauer, T., Mcnamara, D., Dennis, S., Kintsch, W., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2007.
21. Wallach, H.M.; Mimno, D.; McCallum, A. Rethinking LDA: Why Priors Matter. In Proceedings of the 22nd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009; pp. 1973–1981.
22. Galbrun, E.; Miettinen, P. *Redescription Mining*; Springer Briefs in Computer Science; Springer: Berlin/Heidelberg, Germany, 2017.

23. Jaccard, P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles* **1901**, *37*, 241–272. (In French)
24. Sievert, C.; Shirley, K.E. LDAvis: A method for visualizing and interpreting topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, MD, USA, 27 June 2014.
25. Mehri, A.; Jamaati, M. Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations. *Phys. Lett. A* **2017**, *381*, 2470–2477. [CrossRef]
26. Piantadosi, S.T. Zipf's word frequency law in natural language: A critical review and future directions. *Psychon. Bull. Rev.* **2014**, *21*, 1112–1130. [CrossRef]
27. Hofmann, T. Probabilistic Latent Semantic Indexing. In Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 15–19 August 1999; pp. 50–57.
28. Koltcov, S.; Koltsova, O.; Nikolenko, S. Latent Dirichlet Allocation: Stability and Applications to Studies of User-generated Content. In Proceedings of the 2014 ACM Conference on Web Science, Bloomington, IN, USA, 23–26 June 2014; pp. 161–165.
29. Koltcov, S. Application of Rényi and Tsallis entropies to topic modeling optimization. *Phys. A Stat. Mech. Appl.* **2018**, *512*, 1192–1204. [CrossRef]
30. Hall, D.; Jurafsky, D.; Manning, C.D. Studying the History of Ideas Using Topic Models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; pp. 363–371.
31. Misra, H.; Cappé, O.; Yvon, F. Using LDA to Detect Semantically Incoherent Documents. In Proceedings of the Twelfth Conference on Computational Natural Language Learning, Manchester, UK, 16–17 August 2008; pp. 41–48.
32. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Statist.* **1951**, *22*, 79–86. [CrossRef]
33. Akturk, E.; Bagci, G.B.; Sever, R. Is Sharma–Mittal entropy really a step beyond Tsallis and Renyi entropies? *arXiv* **2007**, arXiv:cond-mat/0703277.
34. Heinrich, G. *Parameter Estimation for Text Analysis*; Technical Report; Fraunhofer IGD: Darmstadt, Germany, May 2005.
35. Abbas, A.E.; Cadenbach, A.; Salimi, E. A Kullback–Leibler View of Maximum Entropy and Maximum Log-Probability Methods. *Entropy* **2017**, *19*, 232. [CrossRef]
36. Asuncion, A.; Welling, M.; Smyth, P.; Teh, Y.W. On Smoothing and Inference for Topic Models. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, 18–21 June 2009; pp. 27–34.
37. Goodman, J.T. A Bit of Progress in Language Modeling. *Comput. Speech Lang.* **2001**, *15*, 403–434. [CrossRef]
38. Newman, D.; Asuncion, A.; Smyth, P.; Welling, M. Distributed Algorithms for Topic Models. *J. Mach. Learn. Res.* **2009**, *10*, 1801–1828.
39. Liu, L.; Tang, L.; Dong, W.; Yao, S.; Zhou, W. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* **2016**, *5*, 1608. [CrossRef]
40. De Waal, A.; Barnard, E. Evaluating topic models with stability. In Proceedings of the Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa, Cape Town, South Africa, 27–28 November 2008; pp. 79–84.
41. Rosen-Zvi, M.; Chemudugunta, C.; Griffiths, T.; Smyth, P.; Steyvers, M. Learning Author-topic Models from Text Corpora. *ACM Trans. Inf. Syst.* **2010**, *28*, 4:1–4:38. [CrossRef]
42. Bishop, C.M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*; Springer: Berlin/Heidelberg, Germany, 2006.
43. Bigi, B., Using Kullback–Leibler Distance for Text Categorization. In *Advances in Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 305–319.
44. Ramakrishnan, N.; Kumar, D.; Mishra, B.; Potts, M.; Helm, R.F. Turning CARTwheels: An alternating algorithm for mining redescriptions. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004; pp. 266–275.
45. Parker, A.J.; Yancey, K.B.; Yancey, M.P. Regular Language Distance and Entropy. *arXiv* **2016**, arXiv:1602.07715.

46. Röder, M.; Both, A.; Hinneburg, A. Exploring the Space of Topic Coherence Measures. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Shanghai, China, 2–6 February 2015; pp. 399–408.
47. Stevens, K.; Kegelmeyer, P.; Andrzejewski, D.; Buttler, D. 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, Jeju Island, Korea, 12–14 July 2012; pp. 952–961.
48. Bischof, J.M.; Airoldi, E.M. Summarizing Topical Content with Word Frequency and Exclusivity. In Proceedings of the 29th International Coference on International Conference on Machine Learning, Edinburgh, UK, 26 June–1 July 2012; pp. 9–16.
49. Du, J.; Jiang, J.; Song, D.; Liao, L. Topic Modeling with Document Relative Similarities. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, 25–31 July 2015; pp. 3469–3475.
50. Koltcov, S.N. A thermodynamic approach to selecting a number of clusters based on topic modeling. *Tech. Phys. Lett.* **2017**, *43*, 584–586. [CrossRef]
51. Tsallis, C. *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World*; Springer: New York, NY, USA, 2009.
52. Tkačik, G.; Mora, T.; Marre, O.; Amodei, D.; Palmer, S.E.; Berry, M.J.; Bialek, W. Thermodynamics and signatures of criticality in a network of neurons. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 11508–11513. [CrossRef]
53. Mora, T.; Walczak, A.M. Renyi entropy, abundance distribution and the equivalence of ensembles. *arXiv* **2016**, arXiv:abs/1603.05458.
54. Beck, C. Generalised information and entropy measures in physics. *Contemp. Phys.* **2009**, *50*, 495–510. [CrossRef]
55. Sharma, B.D.; Garg, A. Nonadditive measures of average charge for heterogeneous questionnaires. *Inf. Control* **1979**, *41*, 232–242. [CrossRef]
56. Nielsen, F.; Nock, R. A closed-form expression for the Sharma–Mittal entropy of exponential families. *J. Phys. A Math. Theor.* **2011**, *45*. [CrossRef]
57. Scarfone, A. Legendre structure of the thermostatistics theory based on the Sharma–Taneja–Mittal entropy. *Phys. A Stat. Mech. Appl.* **2006**, *365*, 63–70. [CrossRef]
58. Scarfone, A.M.; Wada, T. Thermodynamic equilibrium and its stability for microcanonical systems described by the Sharma-Taneja-Mittal entropy. *Phys. Rev. E* **2005**, *72*, 026123. [CrossRef]
59. Frank, T.; Daffertshofer, A. Exact time-dependent solutions of the Renyi Fokker–Planck equation and the Fokker–Planck equations related to the entropies proposed by Sharma and Mittal. *Phys. A Stat. Mech. Appl.* **2000**, *285*, 351–366. [CrossRef]
60. Kaniadakis, G.; Scarfone, A. A new one-parameter deformation of the exponential function. *Phys. A Stat. Mech. Appl.* **2002**, *305*, 69–75. [CrossRef]
61. Kolesnichenko, A.V. Two-parameter functional of entropy Sharma–Mittal as the basis of the family of generalized thermodynamices of non-extensive systems. *Keldysh Inst. Prepr.* **2018**, *104*, 35.
62. Elhoseiny, M.; Elgammal, A. Generalized Twin Gaussian Processes Using Sharma—Mittal Divergence. *Mach. Learn.* **2015**, *100*, 399–424. [CrossRef]
63. News Dataset from Lenta.Ru. Available online: https://www.kaggle.com/yutkin/corpus-of-russian-news-articles-from-lenta (accessed on 4 July 2019).
64. Yandex Disk. Available online: https://yadi.sk/i/RgBMt7lJLK9gfg (accessed on 4 July 2019).
65. Basu, S.; Davidson, I.; Wagstaff, K. (Eds.) *Constrained Clustering: Advances in Algorithms, Theory, and Applications*; Taylor & Francis Group: Boca Raton, FL, USA, 2008.
66. Apishev, M.; Koltcov, S.; Koltsova, O.; Nikolenko, S.; Vorontsov, K. Additive Regularization for Topic Modeling in Sociological Studies of User-Generated Texts. In Proceedings of the 15th Mexican International Conference on Artificial Intelligence, MICAI 2016, Cancún, Mexico, 23–28 October 2016.
67. Tsallis, C.; Stariolo, D.A. Generalized simulated annealing. *Phys. A Stat. Mech. Appl.* **1996**, *233*, 395–406. [CrossRef]
68. Vorontsov, K.; Potapenko, A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. In *Analysis of Images, Social Networks and Texts*; Communications in Computer and Information Science; Springer International Publishing: Berlin/Heidelberg, Germany, 2014.

69. Moody, C.E. Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. *arXiv* **2016**, arXiv:1605.02019.
70. Newman, D.; Bonilla, E.V.; Buntine, W. Improving topic coherence with regularized topic models. In *Neural Information Processing Systems (NIPS), Proceedings of Advances in Neural Information Processing Systems 24 (NIPS 2011), Granada, Spain, 12–14 December 2011*; Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Eds.; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2011; pp. 496–504.
71. Liu, Y.; Liu, Z.; Chua, T.S.; Sun, M. Topical Word Embeddings. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; pp. 2418–2424.
72. Wendlandt, L.; Kummerfeld, J.K.; Mihalcea, R. Factors Influencing the Surprising Instability of Word Embeddings. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; pp. 2092–2102.
73. Hofmann, T. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Mach. Learn.* **2001**, *42*, 177–196. [CrossRef]
74. Nikolenko, S.I.; Koltcov, S.; Koltsova, O. Topic Modelling for Qualitative Studies. *J. Inf. Sci.* **2017**, *43*, 88–102. [CrossRef]
75. Naili, M.; Chaibi, A.H.; Ghézala, H.B. Arabic topic identification based on empirical studies of topic models. *Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées* **2017**, *27*. Available online: https://arima.episciences.org/3830 (accessed on 4 July 2019).
76. Andrzejewski, D.; Zhu, X. Latent Dirichlet Allocation with Topic-in-set Knowledge. In Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, Boulder, CO, USA, 4 June 2009; pp. 43–48.
77. Maier, D.; Waldherr, A.; Miltner, P.; Wiedemann, G.; Niekler, A.; Keinert, A.; Pfetsch, B.; Heyer, G.; Reber, U.; Häussler, T.; et al. Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *J. Commun. Methods Meas.* **2018**, *12*, 1–26. [CrossRef]
78. Wang, C.; Blei, D.M. A Split-Merge MCMC Algorithm for the Hierarchical Dirichlet Process. *arXiv* **2012**, arXiv:1201.1657.