Diagnostics of the Topical Model for a Collection of Text Messages Based on Latent Dirichlet Allocation

Alexander Sychev [1]

[1] Voronezh State University, Voronezh, Russia

sav@sc.vsu.ru

Abstract. The problem of constructing a correct topic model is relevant for the automatic processing of large collections of text messages tasks. This paper considers an approach to assessing the topic categorization of a collection of short text messages (labeled up by experts) based on latent Dirichlet allocation (LDA). Possible approaches aimed to solving this problem are discussed. As well the problem of the topics “orthogonality” metrics definition is discussed. The proposed metrics is based on the numerical results of the document-topic matrix analysis. Results of a computer experiment with two large collections of text messages demonstrating LDA based analysis of expert topic models are presented and discussed.

Keywords: text messages, terms, topic model, TF-IDF, lemmatization, latent Dirichlet allocation, clustering, correlation**.**

1 Introduction

The evaluation of an expert topic model for a collection of text messages is a real problem when structuring text content and organizing the search for relevant information in large collections of documents (messages). Both assessing the adequacy of the topic model chosen by the expert and searching for options for possible correction of the original model are included in the evaluation procedure. The description of a topic model is given through the representation of topic categories as a set of terms extracted from texts in a corpus.

It seems to us that the most important indicator of the "ideality" of a topic model is the "orthogonality" of topics that form it. This characteristic of a set of topics implies, on the one hand, the minimum possible degree of intersection of topics with each other and, on the other hand, the internal homogenity of each of the topics of the model. In practical applications, this intuitive definition of "orthogonality" will require the introduction of either a quantitative metric of this characteristic, or a relevant visualization that allows the expert to evaluate it himself. This problem is the focus of this paper.

This paper describes the proposed approach to evaluating an existing expert topic model presented in a collection of short text messages, as well as the results of its

experimental validation using two datasets obtained from online platforms of regional mass media.

This paper is organized as follows: In Section 2, we provide a review of some of the works related to the problem of topic modeling for short texts and relevant methodologies. Section 3 describes the proposal for approach that could be used to analyze a collection of short text messages. Practical validation of the approach is discussed in sections 4 (experimental setup) and 5 (results). Section 6 details the conclusions of the work carried out, with the aim of generating discussion points for future work.

2 Background and Related Work

Topic modeling is often applied to organize efficient access to large collections of texts. However, this application is associated with certain difficulties when working with short texts, where overlapping words in documents are rare, and topic modeling procedures have difficulties in the capturing the semantics of topics in such collections of texts from short documents. These difficulties are mainly due to the fact that topic modeling does not deal well with sparse term matrices. Therefore, before training the topic model, such sparse matrices are pre-compressed, for example, using singular value decomposition (SVD). In terms of statistics, SVD improves the quality of modeling, but the quality of the model suffers in terms of semantics [2].

Among the well-known classification algorithms, one can single out classical algorithms based on calculating the distance function between documents, for example, the iterative K-means algorithm or hierarchical algorithms, and based on the principle of maximizing the likelihood, for example, the probabilistic latent semantic analysis (PLSA) algorithm [3] and an algorithm based on the improved method of latent Dirichlet allocation (LDA) [4]. The last two, unlike the classical clustering algorithms (with hard clustering), assume that each document belongs to several topics at the same time with some probabilities (fuzzy clustering).

The Latent Dirichlet Allocation (LDA) is a generative probabilistic model of a corpus of natural language texts. The basic idea the model is that the content of the document is determined by mixing randomly selected hidden topics, each of which is characterized by a word distribution [4].

The principal advantage of the LDA model is the presentation of documents, which reflects the presence of several topics in it to varying degrees.

An actual aspect of the task of classifying documents is the problem of choosing classification features. A common text model is a representation in which individual words of the text are considered as separate features. In this case, the set of features is too large in size. The use of the LDA model is considered as one of the ways to reduce the dimension of the feature space. In particular, LDA reduces any document to a fixed set of real features, namely, posterior Dirichlet parameters associated with the document.

The LDA models often lack scalability with respect to the vocabulary size. To learn interpretable word embeddings and topics even in corpora with large vocabularies the Embedded Topic Model (ETM) extending the LDA was developed [5]. The

ETM uses embedding representations of both words and topics and contains two notions of latent dimension. Firstly, the vocabulary is embeded in an L-dimensional space. Then each document is represented in terms of K latent topics. In fact the ETM incorporates word similarity into the topic model. Each word is modeled with a categorical distribution whose natural parameter is calculated as the inner product between the word's embedding and an embedding of its assigned topic. One of the popular word embeddings technique is the word2vec algorithm using a neural network model to learn word associations from a large corpus of text [6].

Another problem of the LDA approach is in producing flat topics. In [7] the method for topic detection called Hierarchical Latent Tree Analysis (HLTA) is proposed, where patterns of word co-occurrence and co-occurrences of those patterns are modeled using a hierarchy of discrete latent variables.

The purpose of our research is an exploratory study, and fairly simple and effective approach based on LDA was chosen for this study.

Further the following notation from the LDA model will be used.

• A word (term) is a basic unit of discrete data, defined as an element of the dictionary T, in which the elements are indexed from 1 to V.

• Document $\mathbf{w}$ is a sequence of N words: $\mathbf{w} = (w_1, w_2, ..., w_N)$.

• Corpus D is a set of M documents: $D = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$.

The plate notation of the LDA model is shown in figure 1 (adapted from [4]).

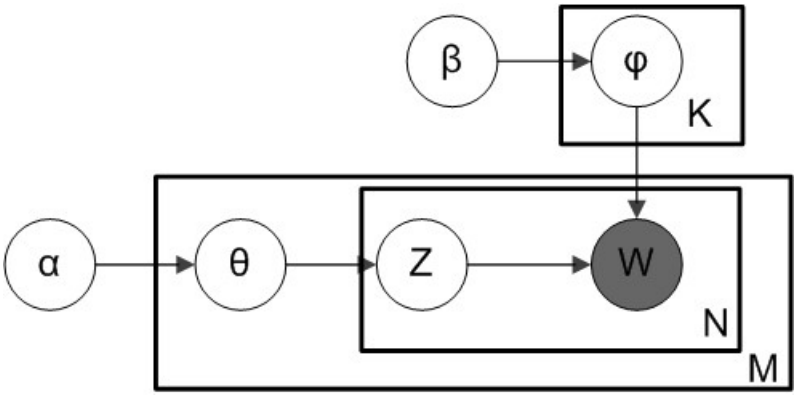

Figure 1. The LDA model.

According to this model, the words w_{ij} are the only observable variables (the W variable is highlighted in gray), and the rest of the variables are latent variables. This model is based on the intuitive assumption that the probability distribution of the words in a topic is highly skewed, so that the set of words that have a high probability value is small in size [4].

The figure shows the following model variables:

— M is the total number of documents in the corpus.
— K is the total number of topics presented in the corpus.
— N is the number of words in the current document.

— α is the a priori Dirichlet parameter of the distribution of topics among documents.
— β is the a priori Dirichlet parameter of the distribution of words by topics.
— θ_i is the distribution of topics for document i.
— φ_k is the distribution of words on topic k.
— z_{ij} is the topic for the j-th word in document i.
— w_{ij} is the j-th word in document i.

Most of publications on the topic modeling are dedicated to the problem of topics discovery from a corpus of texts. The issue of applying topic modeling methods to the task of evaluating the set of topics already existing in the corpus of texts (namely expert topics) has not received due attention in publications.

Results of our previous research on the problem of diagnostics of the topic model for a collection of text messages based on hierarchical clustering of terms are presented and discussed in [1]. As part of the experiment, a graph of terms was constructed that reflected the relations and some numerical features of the terms that form the topic model for a collection of messages. It was demonstrated that the analysis of the structure of the constructed graph helps in formulation some practical recommendations for reorganization the topic model presented in the expert labeled text collection.

3 Methods

Variables $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ in the LDA model can be represented as matrices obtained by decomposition of the original document-word matrix describing the entire corpus of simulated documents. In this representation, $\boldsymbol{\theta}$ consists of rows defined by documents and columns defined by topics, and $\boldsymbol{\varphi}$ consists of rows defined by topics and columns defined by words.

Reflected in $\boldsymbol{\theta}$ the decomposition of messages $\boldsymbol{w}_i$ in corpus D by basis of hidden topics z makes it possible to compare subjective expert decisions on the topics of messages with predictions based on calculations on a large set of text data.

Based on the matrix $\boldsymbol{\theta}$ formed for corpus D, the topic attribution vector for messages $\boldsymbol{w_i}$ was calculated:

$$t_i^* = \arg\max_k \theta_{ik}, i = 1..M \quad (1)$$

The comparing of known expert topics of messages with topics calculated by (1) in an aggregated form was used to diagnose the existing topic model in the corpus D. To diagnose the thematic model of the corpus, two forms of aggregation were proposed: tabular and graphical.

As an additional diagnostic tool, it could be possible to carry out procedures for classifying and clustering messages in the corpus.

The document-topic matrix $\boldsymbol{\theta}$ can be considered as the basis for subsequent clustering (in the case of an unlabeled collection) or classification (for a collection of labeled documents). The probabilities presented in the matrix can be used as message attributes during clustering (classification). A simpler version of document clustering

is possible by choosing, for example, z-topic with the maximum probability (by the maximum value in the matrix row) as the topic of the message.

The presence of a matrix of distribution of topics by terms $\boldsymbol{\varphi}$ allows us to get the initial data for clustering the terms of the dictionary. The simplest variant of dictionary clustering by topics can be implemented by choosing the maximum value in the matrix column.

The presence of a pair of matrices $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ calculated using LDA makes it possible to propose a simple algorithm for topic attribution of new documents (not represented in the training set):

1. Vectorization of a document **w** in the framework of the "bag of words" model with the calculation of the weights of terms according to the TF-IDF metric: **w** = (w_1, w_2,..., w_N).
2. Selection (from the "topics-terms" matrix) of column-vectors corresponding to the terms w_j from the document vector: φ_{kj}, $k = 1,...,K$.
3. Aggregating a set of selected term vectors, for example, by sum the normalized term vectors:
$$\theta_i^* = \sum_{j=1}^{N} (\varphi_{kj} / \sum_{k=1}^{K} \varphi_{kj}), \qquad k = 1, \dots, K$$
4. The resulting vector θ^*_i after appropriate normalization is interpreted as a document vector and shows the probability distribution of topics $k=1,...,K$ for document $\mathbf{w}_i$.

4 Experimental Setup

For the experiment, we used two collections of text messages. The first collection D_1 contained 65091 text messages marked by experts as related to 10 topic categories (table 1a). The second collection D_2 contained 57509 text messages marked by experts as belonging to 8 topic categories (table 1b).

The datasets for the machine experiment were formed by downloading news messages from the online platforms of two popular regional media.

As a result of preliminary processing of the text messages collection, the matrix *MT* ("messages-terms") was formed. The lemmatized forms of words (obtained as a result of the messages analysis according to the bag-of-words model) were chosen as terms. Stop list words were excluded from the terms list in advance. Among 300-400 thousands of extracted lemmas, 1000 most frequent terms were selected into the dictionary *T* for research. The value of the element mt_{ij} of the constructed matrix *MT* was calculated according to the TF-IDF scheme and indicated the frequency of term w_{ij} in the document $\mathbf{w}_i$.

When calculating the LDA for a corpus, the values of the parameters α and β were set equal to $1/K$.

The number of z-topics for the LDA was chosen to be equal to the number of expert topics in the corpora, i.e. $K = 10$ for D_1 and $K = 8$ for D_2.

Table 1a. Characteristics of the messages corpus D_1.

Topic #	Topic title	Number of messages	Average message size, words	Average message lemmas list size
1	city	3727	146	109
2	control	2883	153	113
3	culture	5658	225	162
4	ecology	1347	151	114
5	incidents	24285	100	82
6	money	5294	149	107
7	ofitsialno	869	282	189
8	people	3752	222	159
9	society	13732	143	108
10	sport	3544	150	110
total:		65091	142	107

Table 1b. Characteristics of the messages corpus D_2.

Topic #	Topic title	Number of messages	Average message size, words	Average lemmas list size
1	culture	8363	262	108
2	development	2112	215	103
3	economy	6062	175	96
4	health	4614	126	75
5	sport	13237	156	87
6	criminal	5864	129	90
7	society	11186	279	89
8	accidents	6071	100	73
total:		57509	142	107

5 Results

The aggregated visualization of the matrix $\boldsymbol{\theta}$ obtained for D_1 and D_2 corpora is shown in figures 2a, 2b. This figures show the distribution of messages in the space “expert topics – LDA topics”. Heat scale indicates the concentration (ranged from 0 to 1) of messages among expert topics for fixed z topic values. Due to visualization

limitations, only 3% of all messages from corpus D_1 (D_2) were selected in random for each expert topic.

According to the assumption made us in the introduction, “ideality” of topic model implies “orthogonality” of topics in some sense. Visual interpretation of the “orthogonality” implies that in the heatmap the greatest concentration of messages related to expert topic #i is achieved only in a small area corresponding to any LDA calculated topic #j. At the same time, in the column corresponding to LDA topic #j, there should be no other highly concentrated expert topics besides expert topic #i. In figure 2a the expert topic #10 meets this criterion to the best degree among all expert topics, and expert topic #3 corresponds to a lesser degree. In figure 2b expert topics #4 and #5 are the best, expert topics #1 and #6 are noticeably inferior to them.

An alternative representation of the same corpus can be given in the numerical form as aggregated result of comparing the vector of known expert topics in the messages corpus and the calculated vector t^*. For both corpora D_1 and D_2 the result of analysis is presented in tables 2a, 2b. The numerical values in the tables reflect the proportion of messages related to the k-th expert topic, for which the calculated topic turned out to be equal to z ($z = 1, \dots, K$). Columns in the tables are rearranged in a way that maximum element in j-*th* column is placed to diagonal as closer as possible.

A tabular representation in tables 2a, 2b turns out to be more convenient for introducing a formal metric of topic "orthogonality" via taking into account the distribution of elements c_{ij} in the matrix.

This metric for expert topics #i should take into account the degree of dominance of the value of the element c_{ij} and the degree of heterogeneity in the distribution of elements both in row i and in column j in the matrix.

Based on the data from tables 2a, 2b, the correlation of expert topics #i and #j in the corpus was also calculated from the ith and jth rows in the tables 2a, 2b. The results are presented in tables 3a, 3b.

To understand the topics formed as a result of the LDA experiment, for each expert topic in corpus the list of the first 10 terms that were topically identified according to formula (1) and had the highest estimate of the skew coefficient are provided in tables 4a, 4b. Corresponding skew values are provided in tables 4c, 4d. The ranking order of terms in the list was set by sorting by the value of the skew coefficient Sk, which reflects the uneven distribution of ratings by topic for the dictionary term in the matrix $\boldsymbol{\varphi}$.

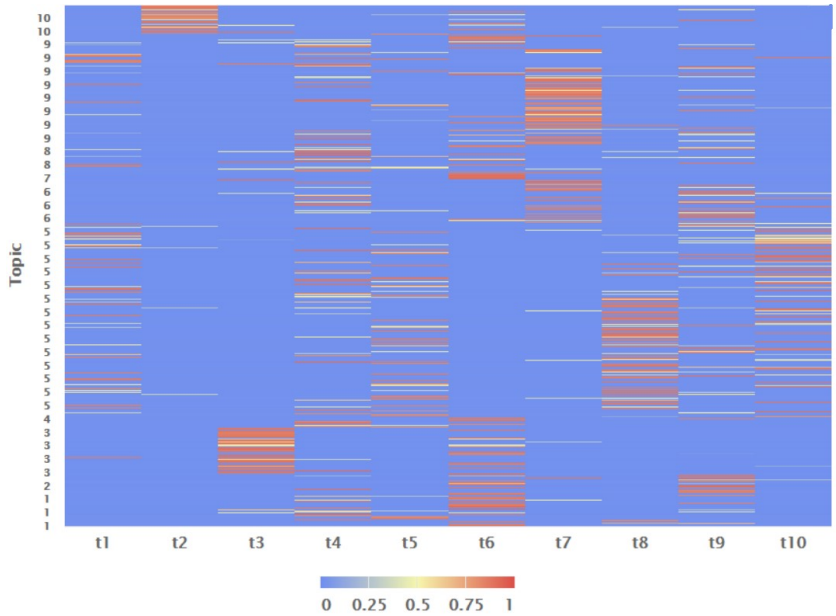

Figure 2a. Heatmap of the distribution of messages in space of expert topics (from *#1* to *#10*) and LDA calculated *z* topics (from t_1 to t_{10}). Messages were taken from D_1 corpus.

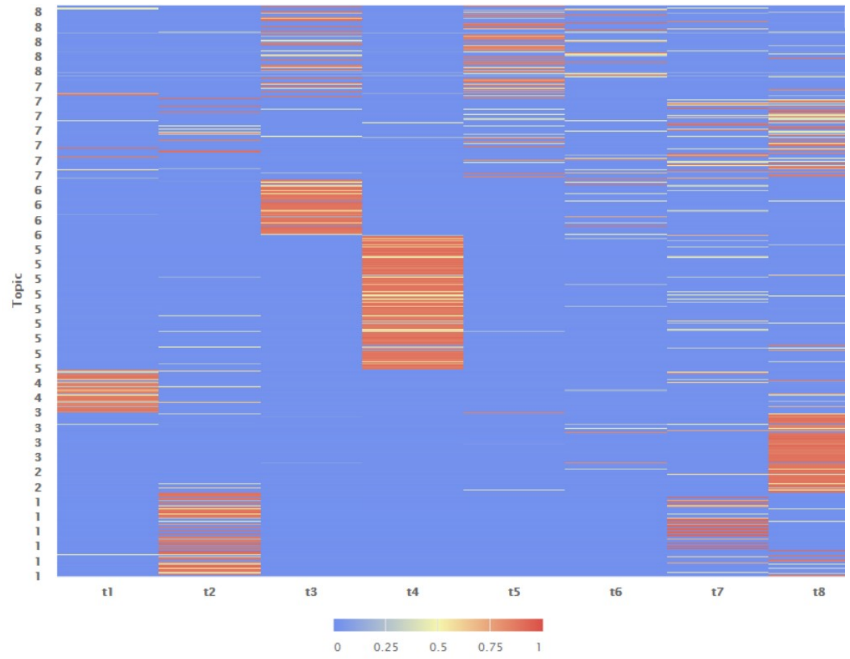

Figure 2b. Heatmap of the distribution of messages in space of expert topics (from *#1* to *#8*) and LDA calculated *z* topics (from t_1 to t_8). Messages were taken from D_2 corpus.

Table 2a. Correspondence of expert and LDA topics for messages corpus D_1

		LDA topics (z)									
		9	**3**	**4**	**10**	**5**	**8**	**7**	**6**	**2**	**1**
LDA topics (z)	**1**	0,11	0,03	0,20	0,01	0,16	0,01	0,10	0,38	0,01	0,00
	2	**0,43**	0,01	0,06	0,11	0,01	0,01	0,19	0,17	0,01	0,00
	3	0,00	**0,70**	0,04	0,00	0,00	0,00	0,00	0,25	0,00	0,00
	4	0,15	0,00	**0,42**	0,02	0,10	0,00	0,09	0,22	0,00	0,00
	5	0,07	0,00	0,08	**0,29**	**0,21**	**0,22**	0,01	0,01	0,00	**0,11**
	6	0,32	0,01	0,10	0,08	0,00	0,00	**0,42**	0,06	0,00	0,00
	7	0,15	0,02	0,06	0,00	0,01	0,00	0,12	**0,63**	0,01	0,00
	8	0,07	0,16	0,37	0,02	0,02	0,00	0,08	0,15	0,01	**0,11**
	9	0,13	0,03	0,23	0,03	0,03	0,01	0,35	0,13	0,00	0,06
	10	0,01	0,03	0,05	0,00	0,00	0,00	0,01	0,13	**0,77**	0,00

Table 2b. Correspondence of expert and LDA topics for messages corpus D_2

		LDA topics (z)							
		2	**7**	**8**	**1**	**4**	**3**	**6**	**5**
Expert topics (t)	**1**	**0,66**	**0,26**	0,08	0,00	0,00	0,00	0,00	0,00
	2	0,03	0,05	**0,90**	0,00	0,00	0,00	0,01	0,01
	3	0,01	0,03	0,84	0,01	0,00	0,00	0,10	0,01
	4	0,02	0,05	0,09	**0,81**	0,01	0,00	0,01	0,01
	5	0,02	0,04	0,03	0,00	**0,90**	0,00	0,00	0,00
	6	0,00	0,01	0,01	0,00	0,00	**0,79**	**0,19**	0,00
	7	0,14	0,16	0,29	0,04	0,01	0,06	0,08	0,23
	8	0,00	0,04	0,04	0,01	0,00	0,36	0,11	**0,43**

Table 5 provides the estimates for Naive Bayes classification based on θ_i vector representation for messages. Number of messages sampled from the corpus D_1 for learning and testing the classifier was about 10000. The accuracy value was 60%.

It can be seen that with the chosen LDA dimension $\boldsymbol{K}$=10, relatively good indicators of accuracy and recall are obtained for topics #3, #5 and #10. Messages for topics #2, #4, #7, #8 are erroneously recognized by the trained classifier as belonging to other topics. It can be expected that in order to improve the quality of the classifier, an increase of the K dimension will be required.

Table 3a. Correlation of expert topics in the corpus D_1.

		Expert topic									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Expert topic	**1**	1,00	0,30	0,10	0,72	-0,30	0,11	0,86	0,45	0,39	-0,13
	2	0,30	1,00	-0,18	0,27	-0,17	0,79	0,41	-0,02	0,44	-0,22
	3	0,10	-0,18	1,00	-0,09	-0,44	-0,25	0,20	0,30	-0,17	-0,09
	4	0,72	0,27	-0,09	1,00	-0,21	0,23	0,40	0,82	0,56	-0,17
	5	-0,30	-0,17	-0,44	-0,21	1,00	-0,27	-0,42	-0,40	-0,40	-0,41
	6	0,11	0,79	-0,25	0,23	-0,27	1,00	0,15	0,06	0,81	-0,24
	7	0,86	0,41	0,20	0,40	-0,42	0,15	1,00	0,21	0,27	-0,01
	8	0,45	-0,02	0,30	0,82	-0,40	0,06	0,21	1,00	0,48	-0,20
	9	0,39	0,44	-0,17	0,56	-0,40	0,81	0,27	0,48	1,00	-0,26
	10	-0,13	-0,22	-0,09	-0,17	-0,41	-0,24	-0,01	-0,20	-0,26	1,00

Table 3b. Correlation of expert topics in the corpus D_2.

		Expert topic							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Expert topic	**1**	1,00	-0,05	-0,09	-0,20	-0,19	-0,27	0,18	-0,39
	2	-0,05	1,00	0,99	-0,06	-0,13	-0,19	0,71	-0,21
	3	-0,09	0,99	1,00	-0,06	-0,15	-0,18	0,68	-0,21
	4	-0,20	-0,06	-0,06	1,00	-0,18	-0,22	-0,29	-0,31
	5	-0,19	-0,13	-0,15	-0,18	1,00	-0,20	-0,44	-0,31
	6	-0,27	-0,19	-0,18	-0,22	-0,20	1,00	-0,31	0,55
	7	0,18	0,71	0,68	-0,29	-0,44	-0,31	1,00	0,20
	8	-0,39	-0,21	-0,21	-0,31	-0,31	0,55	0,20	1,00

Table 4a. Top 10 terms from the lists corresponding to the z-topics identified by the LDA for corpus D_1.

N	LDA topic									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	пропасть	буран	спектакль	животное	пожар	музей	коронавирус	дтп	штраф	убийство
2	поиск	матч	артист	собака	мчс	выставка	вырасти	сбить	нарушение	злоумышленник
3	волонтёр	факел	музыкант	рассказывать	пожарный	художник	цена	авария	прокуратура	подозревать
4	вести	чемпионат	актёр	почему	огонь	акция	пациент	водитель	административный	стража
5	уйти	команда	песня	говорить	спасатель	мероприятие	средний	трасса	закон	задержать
6	искать	игра	сцена	ребёнок	этаж	культура	сутки	ехать	деятельность	преступление
7	родственник	клуб	концерт	мама	поступить	праздник	регион	пассажир	право	судить
8	родный	сезон	музыка	помочь	сообщение	желать	число	травма	млн	возбудить
9	смотреть	спорт	фильм	делать	вода	военный	уровень	автобус	размер	уголовный
10	тело	победа	театр	родитель	окно	участник	послед след ний	дорога	миллион	лишение

Table 4b. Skew coefficient values for top 10 terms taken from z-topics identified by the LDA for corpus D_1.

N	LDA topic									
	1	2	3	4	5	6	7	8	9	10
1	3,15	3,16	3,16	3,09	3,16	3,13	3,15	3,16	3,05	3,15
2	2,96	3,16	3,15	3,09	3,16	3,12	3,12	3,15	2,99	3,15
3	2,96	3,16	3,15	2,80	3,15	2,83	3,07	3,13	2,91	3,12
4	2,69	3,15	3,15	2,73	3,14	2,83	3,05	3,13	2,76	3,09
5	2,52	3,08	3,15	2,70	3,10	2,73	3,03	3,12	2,72	3,08
6	2,37	2,93	3,13	2,66	2,68	2,57	2,98	3,07	2,62	3,05
7	2,17	2,54	3,13	2,64	2,49	2,49	2,96	3,06	2,50	2,96
8	2,00	2,36	3,12	2,57	2,06	2,47	2,90	3,04	2,48	2,92
9	1,80	2,22	3,09	2,53	1,91	2,24	2,88	2,97	2,39	2,81
10	1,66	2,06	3,07	2,47	1,80	2,23	2,84	2,89	2,34	2,77

Table 4c. Top 10 terms from the lists corresponding to the z-topics identified by the LDA for corpus D_2.

N	LDA topic							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	covid	выставка	кража	буран	мчс	прокуратура	говорить	млн
2	инфекция	театр	полицейский	матч	пожар	нарушение	очень	работа
3	коронавирус	фестиваль	мужчина	факел	авария	штраф	это	рубль
4	пациент	музей	уголовный	турнир	пожарный	получение	такой	проект
5	пандемия	художник	возбудить	спортсмен	дтп	особо	ребёнок	год
6	врач	искусство	ук	соревнование	водитель	долг	свой	регион
7	случай	спектакль	ст	команда	пострадать	административный	время	тыс
8	сутки	конкурс	рф	игра	автомобиль	мошенничество	наш	область
9	борьба	культура	дело	кг	произойти	размер	человек	район
10	число	концерт	летний	клуб	улица	крупный	который	воронежский

Table 4d. Skew coefficient values for top 10 terms taken from z-topics identified by the LDA for corpus D_2.

N	LDA topic							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	2,83	2,82	2,83	2,83	2,83	2,33	2,57	2,36
2	2,83	2,80	2,69	2,83	2,83	2,32	1,84	1,99
3	2,83	2,80	2,61	2,83	2,82	2,31	1,77	1,96
4	2,83	2,79	2,57	2,83	2,80	2,08	1,57	1,67
5	2,81	2,79	2,57	2,83	2,80	2,08	1,24	1,42
6	2,73	2,79	2,52	2,83	2,67	1,87	0,90	1,27
7	2,66	2,75	2,41	2,82	2,64	1,66	0,78	0,98
8	2,63	2,03	2,21	2,80	1,77	1,54	0,69	0,84
9	1,65	1,64	2,19	2,79	1,51	1,07	0,36	0,67
10	1,64	1,62	1,68	2,74	0,96	0,57	-0,17	0,18

On the basis of the matrix θ_i, a clustering of a random sample of messages using the *K*-means method was also carried out. The sample size was 1974 messages, the number of clusters was 10. Table 6 shows the distribution of the set of sampled messages from the corpus over the formed clusters. Table rows correspond to expert topics. The table also shows the skewness values calculated from the rows and columns of the table.

Analysis of the data from this table shows that some of the expert topics, for example #3, #7, #10 differ in that most of the messages from them were grouped within single clusters, while messages from other topics were dispersed over several clusters. On the other hand, among the formed clusters there are those in which there are strongly dominant expert topics, for example clusters 1,3,7,9 and those in which the topics are very blurred, for example clusters 2 and 5.

It is noteworthy that expert topic #5, for which the classifier provides relatively high values of recall and precision. On the other hand, during clustering, the bulk of the messages of this set are distributed among three dominant and three subdominant subsets.

Interpretation of the results shown in Figures 2a, 2b and in tables 2-3, 5-6 allows us to draw the following conclusions.

1. As result of the analysis of the distribution of messages in corpus the following types of expert topics could be distinguished:

- topics that mostly retain their "identity", for example #3 ("culture"), #10 ("sports") in corpus D_1, and #1 ("culture"), #4 ("health"), #5 ("sport") in corpus D_2.
- topics that are clearly divided into component subtopics, for example, topic #5 ("incidents") is divided into well-defined LDA z-topics t_5, t_8, t_{10} in corpus D_1, and topics #1 ("culture"), #7 ("society"), #8 ("accidents") are subjects for possible splitting in corpus D_2 as well.

– "non-self-sufficient" topics that are combined with other expert topics as part of LDA z-topics, for example topics #1 and #7, #4 and #8 in corpus D_1, and topics #2 and #3, #2 and #7, #3 and #7 8 in corpus D_2. Tables 3a, 3b shows a clear correlation for such pairs of topics.

Table 5. Performance evaluation for message classification based on the θ_i vector.

True/Prediction	True topic 1	True topic 2	True topic 3	True topic 4	True topic 5	True topic 6	True topic 7	True topic 8	True topic 9	True topic 10	class precision, %
Prediction topic 1	45	12	36	10	2	7	16	15	36	11	23,7
Prediction topic 2	0	0	0	0	1	0	0	0	0	0	0,0
Prediction topic 3	3	0	123	0	0	1	1	22	6	6	75,9
Prediction topic 4	0	0	0	0	0	0	0	0	0	0	0,0
Prediction topic 5	30	18	1	3	645	18	1	20	79	0	79,1
Prediction topic 6	8	28	1	5	33	38	1	4	47	0	23,0
Prediction topic 7	0	0	0	0	0	0	0	0	0	0	0,0
Prediction topic 8	0	0	1	0	0	0	0	0	1	0	0,0
Prediction topic 9	26	27	6	22	44	95	6	50	238	7	45,7
Prediction topic 10	0	1	2	0	3	0	1	1	5	82	86,3
class recall, %	40,2	0,0	72,4	0,0	88,6	23,9	0,0	0,0	57,8	77,4	

Table 6. Distribution of messages by clusters.

Expert topic #	**Cluster Id**										**Skew**
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
1	11	3	41	4	0	11	26	15	0	0	*1,4*
2	13	0	22	1	5	38	5	2	0	0	*1,8*
3	1	114	46	0	0	2	5	0	1	0	*2,5*
4	5	0	8	0	0	6	19	2	0	0	*2,0*
5	6	1	4	173	204	48	65	144	83	0	*0,7*
6	64	0	13	0	10	54	17	0	0	0	*1,5*
7	2	2	17	0	0	3	1	1	0	0	*2,9*
8	11	21	23	0	1	10	34	4	6	2	*1,0*
9	164	9	36	4	10	45	109	10	23	1	*1,7*
10	4	3	10	0	0	3	4	0	0	82	*3,1*
Skew	*2,6*	*3,0*	*0,6*	*3,2*	*3,1*	*0,5*	*1,8*	*3,1*	*2,8*	*3,2*	

The identification of different types of expert topics as result of presented in corpus of messages topic model diagnostics can be used for the subsequent reorganization of the topic model.

2. LDA allows us to represent a collection of texts in a more compact and structured form (like matrices $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$), convenient for classifier training or clustering. As can be seen from table 5, even with a relatively small dimensionality of the feature space formed on the basis of implicit LDA z-topics, relatively good accuracy and recall rates are achieved for some expert topics.

6 Conclusion

The results of applying LDA to the corpus of texts could be used to diagnose the existing expert topic model in the corpus of text messages labeled up by experts. Vague ideas about the “orthogonality” of topics can be translated into quite measurable numerical metrics through LDA based representation of the corpus.

The discussed in the paper approach to the definition of the "orthogonality" metric for expert topics was tested using a computer experiment for two large collections of text messages. The experiment demonstrated the possibility of applying the proposed approach to the task of diagnosing an expert topic model. The approach could be implemented involving simplified calculations, for example, based on formula (1), or by carrying out classification or clustering procedures.

The influence of the choice of the value of the parameter K of the LDA-model on the change in accuracy and recall rates for all expert topics of the corpus of messages requires a separate study. This will allow evaluating the effectiveness of LDA as a tool for reducing the dimensionality of space for representing textual data in data mining.

References

1. Sychev, A.V. Diagnostics of the Topic Model for a Collection of Text Messages Based on Hierarchical Clustering of Terms. Lobachevskii J Math 44, 219–226. (2023). https://doi.org/10.1134/S1995080223010390
2. Steuber, F., Schoenfeld, M., Rodosek, G.D.: Topic Modeling of Short Texts Using Anchor Words. WIMS 2020: Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics, June 2020, pp. 210–219. (2020) https://doi.org/10.1145/3405962.3405968.
3. Hofmann, T.: Probabilistic latent semantic analysis. Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 50–57. (1999) https://doi.org/10.48550/arXiv.1301.6705
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research. Vol. 3, 993–1022. (2003).
5. Dieng, A.B, Ruiz, F.J.R., Blei, D.M.: Topic Modeling in Embedding Spaces. Transactions of the Association for Computational Linguistics, 8, 439–453 (2020). https://doi.org/10.1162/tacl_a_00325

6. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, USA, 3111-3119 (2013). https://doi.org/10.48550/arXiv.1310.4546
7. Liu, T., Zhang, N.L., Chen, P.: Hierarchical Latent Tree Analysis for Topic Detection. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD. Lecture Notes in Computer Science, vol 8725, Springer, Berlin, Heidelberg. (2014) https://doi.org/10.1007/978-3-662-44851-9_17.