Information systems. Computer sciences. Issues of information security

Информационные системы. Информатика. Проблемы информационной безопасности

UDC 004.855.5:004.622
https://doi.org/10.32362/2500-316X-2025-13-1-38-48
EDN HJHQTR

RESEARCH ARTICLE

Topic modeling in the stream of short messages in Russian

Elena S. Mozaidze @

V.G. Shukhov Belgorod State Technological University, Belgorod, 308012 Russia
@ Corresponding author, e-mail: mozaidze95@mail.ru

Abstract
Objectives. This work is devoted to the topic modeling of short messages received through social networks or in another way in the form of a series of short messages. This need arises in public relations systems in state and municipal structures, in public opinion polling centers, as well as in customer service systems and marketing departments. The aim of the work is to develop and experimentally test a set of algorithms for a thematic model for automatically determining the main topics of information exchange and typical messages illustrating these topics.
Methods. The work uses methods of variable statistical distributions applied to collocation statistics and approaches typical for resolving problems of topic modeling of short texts, but applied to successive messages. In this way, online machine learning and topic modeling are considered jointly.
Results. The work considered the construction of a thematic model in which clusters found with the presentation of their typical representatives and current weight can help decision-making in accordance with the subject of these most important messages. The proposed method was experimentally tested on a corpus of real messages. The results of topic modeling (the constructed thematic models) are consistent with the results obtained manually. The messages selected illustrate that the topics with the highest weight are seen as such from the point of view of human experts.
Conclusions. The proposed algorithm of topic modeling allows the most important topics in current communication to be automatically identified. It shows posts that serve as indicators of these topics, and thereby significantly simplifies the solution of the problem.

Keywords: topic modeling, EM-algorithm, hidden placement, streaming renormalization method

• **Submitted:** 25.03.2024 • **Revised:** 30.09.2024 • **Accepted:** 17.11.2024

For citation: Mozaidze E.S. Topic modeling in the stream of short messages in Russian. *Russian Technological Journal.* 2025;13(1):38–48. https://doi.org/10.32362/2500-316X-2025-13-1-38-48, https://elibrary.ru/HJHQTR

Financial disclosure: The author has no financial or proprietary interest in any material or method mentioned.

The author declares no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Тематическое моделирование в потоке коротких сообщений на русском языке

Е.С. Мозаидзе @

Белгородский государственный технологический университет им. В.Г. Шухова, Белгород, 308012 Россия
@ *Автор для переписки, e-mail: mozaidze95@mail.ru*

Резюме
Цели. Работа посвящена тематическому моделированию коротких сообщений, поступающих посредством социальных сетей или другим способом в виде серии. Такая задача возникает в системах работы с населением в государственных и муниципальных структурах, в центрах опроса общественного мнения, а также в системах обслуживания клиентов и маркетинговых подразделениях. Цель работы – разработка и экспериментальная проверка набора алгоритмов тематической модели для автоматического определения основных тем обмена информацией и типичных сообщений, иллюстрирующих эти темы.
Методы. Используются методы переменных статистических распределений, примененных к статистике коллокаций, и подходы, характерные для решения задач тематического моделирования коротких текстов, но в применении к следующим друг за другом сообщениям. Таким образом, задачи онлайнового машинного обучения и тематического моделирования рассматриваются в совокупности.
Результаты. Рассмотрено построение тематической модели, в которой найденные кластеры с предъявлением их типичных представителей и текущего веса могут помочь человеку в принятии решений в соответствии с тематикой этих наиболее важных сообщений. Предложенный метод был экспериментально протестирован на корпусе реальных сообщений. Результаты тематического моделирования (построенные тематические модели) согласуются с результатами, полученными вручную: выбранные сообщения, иллюстрирующие проблемные темы с наибольшим весом, являются таковыми и с точки зрения экспертов.
Выводы. Предлагаемый алгоритм тематического моделирования позволяет автоматически выявлять наиболее важные темы в текущем общении, показывает посты, служащие индикаторами этих тем, что позволяет существенно упростить решение задачи.

Ключевые слова: тематическое моделирование, EM-алгоритм, скрытое размещение, метод поточной перенормировки

• **Поступила:** 25.03.2024 • **Доработана:** 30.09.2024 • **Принята к опубликованию:** 17.11.2024

Для цитирования: Мозаидзе Е.С. Тематическое моделирование в потоке коротких сообщений на русском языке. *Russian Technological Journal.* 2025;13(1):38–48. https://doi.org/10.32362/2500-316X-2025-13-1-38-48, https://elibrary.ru/HJHQTR

Прозрачность финансовой деятельности: Автор не имеет финансовой заинтересованности в представленных материалах или методах.

Автор заявляет об отсутствии конфликта интересов.

INTRODUCTION

When working with social networks and messengers, the need almost always arises for an automated search for the most important topic in the exchange of messages. This is due to many reasons, including the need for chat moderation, identifying moments when a responsible person needs to intervene, searching for the most important topics of communication at the moment in the context of chat topics.

The case studied in this article relates to information exchange in social networks in the city of Belgorod. The reason for this choice is that the authors were able to obtain this data, however, the proposed methodology is applicable to any research subject of this kind for which there is a sufficient amount of data available.

Topic modeling is a way of training a machine (computer) to identify meaningful topics in texts. For example, by analyzing an array of news and journalistic texts, it is possible to identify certain topics. Of course, computers cannot understand the meaning of articles literally, but if there is a large collection of texts with different topics, then the probabilities of joint use of words enable to identify separate thematic layers.

A topic stratum filtered from a set of texts is simply a set of words characteristic of a topic. Words in such a set are sorted by importance for the topic [1–3]. In terms of cluster analysis, a topic is the result of biclustering, i.e., simultaneous clustering of both words and documents according to their semantic proximity.

In 1998, the scientists K. Papadimitriou, H. Tomaki, S. Vempala, and P. Raghavan were among the first to show interest in the topic of the probabilistic topic model [4]. Their work was devoted to latent semantic indexing (LSI), a method of information retrieval based on spectral analysis of the document database.

Further development of this topic is reflected in the works of foreign scientists.

Thomas Hofmann [5] studied probabilistic latent semantic indexing. Unlike the standard latent semantic indexing using singular value decomposition, the probabilistic variant has a strong statistical foundation and defines a proper generative model of the data. Search experiments on a number of test collections show significant performance gains over direct term matching methods as well as LSI. David Bley [6–8] considered supervised latent Dirichlet allocation (sLDA) or the statistical model of labeled documents. In his papers, he illustrates the advantages of sLDA over modern ordered regression, as well as over unsupervised latent Dirichlet allocation (LDA) analysis followed by a separate regression. Andrew Ng, an American computer scientist, Associate Professor at Stanford University, a researcher into robotics and machine learning, and one of the founders of the online learning platform “Coursera”[1], predicted long ago[2] [3] that natural language recognition would become the main method of human-computer interaction. In his work, he drew attention to reinforcement learning as one of the ways of machine learning.

Russian scientists have also contributed to the development of this topic.

Vorontsov [9] proposed in his work additive regularization of topic models (ARTM), based on maximization of the weighted sum of the logarithm of likelihood and additional criteria: regularizers. This simplifies the combination of topic models and the construction of any complex multi-objective models. Potapenko [10] considered a generalized EM-algorithm[3] with smoothing, sampling and thinning heuristics, enabling both known thematic models and new ones to be obtained at different combinations of these heuristics. Lukashevich [11] and Nokel[4] have presented the results of experiments on adding bigrams to topic models and taking into account the similarity between them and unigrams. They proposed a new algorithm PLSA-SIM, as a modification of the probabilistic latent semantic analysis (PLSA) algorithm for building topic models. The article by Korshunov and Gomzin presents a comparative review of various models, describes ways of estimating their parameters and quality of results, and gives examples of open software implementations [12].

Topic modeling software libraries such as *Mallet*[5], *Gensim*[6], and *BigArtm*[7] have been developed, thus enabling the creation of probabilistic topic modeling.

The active use of large language model (LLM) tools, including for resolving topic modeling problems, began a few years ago. Quite a large number of works in this area have appeared, a number of which are relevant to the objectives of this study. In [13], the authors study key events in news feeds. The problem of their identification and links is considered. The study is based on the use of LLM for retrieval and summarization, while actual topic modeling is done by selecting the top topic with a sliding window algorithm. Despite

[1] https://www.coursera.org. Accessed December 02, 2024.

[2] Ng A.Y. *Shaping and Policy Search in Reinforcement Learning*. Ph.D. Thesis, UC Berkley, 2003.

[3] An expectation–maximization (EM) algorithm is an iterative method used in mathematical statistics to find maximum likelihood estimates of the parameters of probabilistic models when the model depends on some hidden variables.

[4] Nokel M.A. *Methods for improving probabilistic topic models of the text collections based on lexicoterminological information*: Cand. Sci. Thesis (phys.-math.). Moscow, 2015. 20 p. (in Russ.).

[5] http://mallet.cs.umass.edu/topics.php. Accessed December 02, 2024.

[6] https://radimrehurek.com/gensim. Accessed December 02, 2024.

[7] http://bigartm.org. Accessed December 02, 2024.

good results in the stated area, the dynamics of topic distribution is not explicitly taken into account in the work and, in addition, the model is not pre-trained. The article [14] discusses the interpretation of topic model results using the large language model ChatGPT[8]. Interestingly, the result of the study showed that this LLM diverged from human interpretation in almost half of the cases. The purpose of the work was simply to demonstrate the ability of ChatGPT to describe topics and provide useful information, although the LLM did not help in topic modeling itself[9, 10]. The works focus, respectively, on topic modeling of summarized texts and the evaluation of contextualized topic coherence. The first article shows that LLM successfully translates meaning into summarized texts, but topic modeling in them is not always correct: the quality depends on the context. The automatic topic coherence evaluation proposed by the authors of the second paper works well for short documents and is not affected by meaningless but highly rated topics. This result certainly deserves to be considered and used for further research.

The aim of [1] is concentrated on statement of objective and the concept of researching the texts received by the state structures in order to categorize them into structural units of governance corresponding to the topics of the texts. The results were used in the present study.

The objectives of this study are the development and experimental validation of a set of topic modeling algorithms which will lead to the construction of a set of clusters with the presentation of their typical representatives and current weights, subject to the normal normalization and distribution conditions usual for topic modeling.

1. MATERIALS AND METHODS

1.1. Task statement

Let us assume that there is a permanent system which accepts short and medium-length messages (remarks, appeals). It can be a personal page in a social network, a web service for receiving requests, an electronic mailbox with automated text uploading, a customer relationship management system, etc.

Most often, one of two objectives arises for the information collected in this way: 1) to distribute messages into predefined groups (classes); or 2) to group messages into predefined groups (clusters) with similar semantics. Let us consider the second problem on the flow of messages, namely: to each newly arrived message we compare a vector, the coordinates of which represent the probabilities of this message belonging to the clusters formed by this moment.

The above objective is topic modeling or, in other words, soft biclustering. In this formulation, the problem is complicated by the fact that the set of messages is not bounded, and it is necessary either to determine the number of clusters each time, or to fix it and disband unnecessary clusters.

In the case of a satisfactory solution, the described objective can be applied in outreach systems for state and municipal structures, in public opinion polling centers, as well as in CRM (customer relationship management) systems and marketing services of corporations.

1.2. Solution method

Topic modeling methods are usually based on computing frequencies of words in documents, as well as words and documents in topics. The most commonly used methods for topic modeling are the EM-algorithm[11, 12], or hidden Dirichlet placement[13]. The distinguishing features of both are the need to weight all words in messages without taking meaning relations into account. Meanings are recovered from known meanings in large arrays of text. In essence, topic modeling only performs meaning benchmarking.

In the task at hand, working with meaning in the way described (similar to benchmarking) is not possible, because messages are usually short and often contain grammatical errors. In such a case, it is extremely difficult to orient them to a topic by means of benchmarking: this requires a larger block of text.

In [1], a technique of working with messages based on noun–verb pairs is proposed. The set is mapped to each message with at least one such pair. In this study, we will assume that such mapping has already been performed and each message has an identifier and a set of noun–verb pairs corresponding to it.

It is further assumed that:

- messages follow each other and each has an identifier;
- if there are no verbs or nouns in the message, it is considered irrelevant and is not taken into account, but an identifier is assigned to it;

[8] https://chat-gpt.org/. Accessed December 02, 2024.

[9] https://arxiv.org/abs/2403.15112. Accessed December 02, 2024.

[10] https://arxiv.org/abs/2305.14587. Accessed December 02, 2024.

[11] https://pythobyte.com/python-for-nlp-topic-modeling-8fb3d689/?ysclid=lgdpql4ef3963911399 (in Russ.). Accessed December 02, 2024.

[12] https://mathprofi.com/userinfo/14285/ (in Russ.). Accessed December 02, 2024.

[13] https://digitrain.ru/articles/252142/ (in Russ.). Accessed December 02, 2024.

- for the sliding message window, there is a log of noun–verb pairs with the corresponding message ID;
- topics are identified by the nouns included in the messages;
- each noun–verb pair has a probability (a number between 0 and 1) of occurring in a topic such that the sum of the probabilities for the pair across all topics is 1;
- each message has a probability of entering a topic, and the sum of the message probabilities across all topics is 1;
- the maximum number of topics is fixed. Deletion of topics is based on the topic weight indicator, calculated as the product of the average pairing probability in the topic by the number of messages related to the topic.

Such assumptions allow us to use the method of streamline renormalization: regular discarding of elements with small weights.

Next, let us consider one by one the steps of the process of analyzing messages in a moving window.

1.2.1. Preprocessing

The analyzer input receives the messages representing a text in Russian language consisting of one or several sentences. In order to prepare for modeling, the following operations are performed on the message text:

- sentence tokenization (splitting into sentences) by means of *nltk*[14];
- tokenization of words in a sentence by means of *nltk*;
- part-of-speech detection of tokens by means of *nltk*;
- lemmatization of nouns and verbs by means of *pymorphy2*[15];
- composition of noun–verb pairs according to the “nearest neighbors” rule.

All of this is implemented in the *appeals_processing.ipynb*[16] program module and can run in real time in the message stream.

1.2.2. Topic modeling

The output of the *appeals_processing.ipynb* preprocessor is a set of lemmatized nouns and verbs contained in a message box of length *A* lines, and a .csv or .json file (currently implemented to receive .csv, but reconfiguring to .json is easy), each line of which contains a list of three items:

- a string of the type “[(noun1, verb1), ..., (nounN, verbN)]”;
- file number;
- number of the message in the file.

Thus, a set of noun–verb pairs can be nested in a number set on the one hand, and on the other hand unambiguously related to the message upon which it is constructed.

Then, for each pair in each message, the following algorithm is implemented, the steps of which are discussed in more detail in the Results section:

1. If the noun is contained in topics, the pair is included in those topics as well as in newly created topics for this message with equal probabilities.
2. If the noun is not contained in topics but the verb is contained in topics, the pair is included in those topics with probabilities reduced by a multiplier μ (a given constant parameter), and a new topic on the noun is created. The pair is included in it and in the newly created topics for that message with equal probabilities.
3. If neither the noun nor the verb is contained in topics, a new topic on the noun is formed and the pair is included in it, as well as in the newly created topics for this message with equal probabilities.

After all pairs of all messages have been processed by this algorithm, a .json file of the topic model is created. Each topic key (noun) is matched to a .json object with keys (noun–verb pairs), and values—probabilities of occurrence of the pair in the topic.

The result of the topic modeling is a convolved .json object of the topic model. In it, each topic key corresponds to a numerical value of the weight of this topic. Sorted by descending values, this object will give the desired clusters with their weights.

1.3. Materials

The data for this study was prepared from incoming messages which came to the administration of the city of Belgorod. Proper names and other attributes which de-mask personal data were removed. The dataset processed in this way was a .csv file with a header of the form “Date; Time; Question,” any message can be reconstructed by date and time of receipt. The questions were the messages (requests) processed for the following purposes: to build a tag cloud (which in this case was a topic cloud) and to identify the most important questions.

The dataset included 3621 messages, some of which were repetitive (people copied the question and resent it). Some of which did not contain noun–verb pairs or were misspelled in such a way that the words were not in the underlying dictionaries.

The dataset review was organized by a sliding window of 300 messages. This is the average volume of messages for a month. Thus, the system analyzed the content of messages for the past month and provided the topics and the most important issues in messages for

[14] https://www.nltk.org/. Accessed December 02, 2024.

[15] https://github.com/pymorphy2/pymorphy2. Accessed December 02, 2024.

[16] https://disk.yandex.ru/d/8LPWy3ZP-7V30Q (in Russ.). Accessed December 02, 2024.

this period. Since the window was sliding, the result of the work could change every day, even with each new message.

Currently, the administration specialists prioritize the consideration of appeals manually. The proposed program system will be able to do this automatically. The study proposes a metric to measure the quality of sorting of appeals by a machine in comparison with sorting by humans.

2. RESULTS

2.1. Computing probabilities in a topic model

There are the following items as inputs to the topic modeling algorithm:

- .csv-file or other referral source containing the date and time the referral was received and the text of the reference;
- dictionaries of Russian nouns and verbs;
- dictionary T_m of the topic model built by this moment (by message number m): the key (noun) finds a dictionary with a key (a noun–verb pair) and a value in the form of the pair's probability of belonging to the topic. At the beginning of the work this dictionary does not yet exist, it is initiated by the first essential (containing a noun and a verb) reference.

In the dictionary of the topic model, the sum of the probability values (hereafter simply values) across all topics for a given pair must equal 1. That is:

$$T_m = \{t : \{s : p_{st}\}\}, \sum_t p_{st} = 1 \; \forall s \in S_m,$$

wherein t is the noun (topic), s are the noun–verb pairs constructed from the set S_m containing m references, p_{st} is the probability that the pair s belongs to topic t.

The output of the algorithm is a new T_{m+1} topic model dictionary which satisfies the same requirement.

Possible cases for each pair of s from reference $m + 1$:

1. If there is already a pair s in T_m, all its values are replaced by the values of $\hat{p}(1 - p_0 n_0)$, where $\hat{p}$ is the current value and p_0, n_0 are the value for the pair s in the new topics and their number, respectively;
2. If there is no pair s in T_m, but there is a topic that matches a noun from the pair s, then s is entered into the dictionary of that topic with a value equal to p_1; in addition, the pair s is entered into all newly created topics with probabilities p_4;
3. If there is no topic matching the noun from s in T_m, but some topics have a pair with the verb from s, then s is entered into the dictionaries of these topics with a value equal to p_2. In addition, the pair s is entered into all newly created topics with probabilities p_5;
4. If neither the noun nor the verb from the pair s is present in T_m, then a new topic (based on the noun from s) is created in T_{m+1}, and s is entered into it with probability p_3. In all other newly created topics, s is entered with the probability p_6.

Thus, we have non-intersecting possibilities for a pair s, and finally the following relations must be satisfied:

Case 1. The sum over all topics containing s, its new values, must be equal to 1:

$$p_0 n_0 + \sum_s \hat{p}_s (1 - p_0 n_0) = 1,$$

which is fulfilled identically. But since p_0 has the sense of probability, then:

$$p_0 \le \frac{1}{n_0},$$

Case 2. The sum of all old and new topics is equal to 1:

$$p_1 n_{\text{noun}} + n_0 p_4 = 1,$$

wherein n_{noun} is the number of available (old) topics with the noun from the pair s. Hence we obtain:

$$p_1 = \frac{1 - n_0 p_4}{n_{\text{noun}}}, \quad p_4 \le \frac{1}{n_0}.$$

Case 3. The sum of all old and new topics is equal to 1:

$$p_2 n_{\text{verb}} + n_0 p_5 = 1,$$

wherein n_{verb} is the number of old topics with a verb from the pair s. Hence we obtain:

$$p_2 = \frac{1 - n_0 p_5}{n_{\text{verb}}}, \quad p_5 \le \frac{1}{n_0}.$$

Case 4. The sum of all new topics is equal to 1:

$$p_3 n_{\text{new}} + (n_0 - n_{\text{new}}) p_6 = 1,$$

wherein n_{new} is the number of new topics with the noun pair s. Hence we obtain:

$$p_3 = \frac{1 - (n_0 - n_{\text{new}}) p_6}{n_{\text{new}}}, \; p_6 \le \frac{1}{n_0 - n_{\text{new}}}.$$

Since all probabilities p_4, p_5, p_6 correspond to the inclusion of a pair in the topic only on the basis of its presence in one reference, we will assume that they are equal to:

$$p_4 = p_5 = p_6 \equiv q \le \frac{1}{n_0}.$$

We will consider the probabilities p_1 and p_5 to be equal, since they correspond to inclusion in the topic by noun.

We will relate the probability of inclusion in the topic by noun to the probability of inclusion in the topic by verb:

$$p_2 = kp_1.$$

Let us denote $p_1 = p_3 = p$, and now the unit sums are rewritten as:

$$pn_{\text{noun}} + n_0 q = 1,$$

$$kpn_{\text{verb}} + n_0 q = 1,$$

$$pn_{\text{new}} + (n_0 - n_{\text{new}})q = 1.$$

Let us solve this system of equations and determine p, q, k:

$$p = \frac{n_{\text{new}}}{n_{\text{new}}n_0 - n_{\text{noun}}n_0 + n_{\text{noun}}n_{\text{new}}},$$

$$q = \frac{n_{\text{new}} - n_{\text{noun}}}{n_{\text{new}}n_0 - n_{\text{noun}}n_0 + n_{\text{noun}}n_{\text{new}}},$$

$$k = \frac{n_{\text{noun}}}{n_{\text{verb}}},$$

if all n_{verb} and $n_{\text{new}} \cdot n_0 - n_{\text{noun}} \cdot n_0 + n_{\text{noun}} \cdot n_{\text{new}}$ differ from 0.

Let us now consider the cases where one or both of these values are equal to 0.

If $n_{\text{verb}} = 0$, then there are no pairs in the message corresponding to Case 3, and hence there are only two equations. The unknowns are only p and q. The solutions for them have already been found above.

If

$$n_{\text{new}} \cdot n_0 - n_{\text{noun}} \cdot n_0 + n_{\text{noun}} \cdot n_{\text{new}} = 0,$$

then it follows from the equations that $n_{\text{new}} = n_{\text{noun}} = 0$, i.e., this case corresponds to the described Case 1, when the pair s is already contained in T_m, and the parameter p_0 can have any value.

Since the new topics in this case definitely do not contain a noun from s, the pairing probability in them must be less than or equal to the probability of any pairing from a given reference in the old dictionary:

$$\hat{p}_s(1 - p_0 n_0) \le \min_s \hat{p}_s \Big|_{\hat{p}_s \ne 0}.$$

As a result, in order to find a new dictionary T_{m+1} for each pair from the circulation, we need to find the numbers:

$$n_{\text{noun}}, n_0, n_{\text{new}}, n_{\text{verb}}, p_0,$$

and n_0 is the same for all pairs s from a given circulation.

When building a new dictionary T_{m+1}, the following values of probabilities should be used:

$$p_0 = \frac{1}{n_0}\left(1 - \frac{1}{\hat{p}} \min_1 \hat{p}_1 \Big|_{\hat{p}_1 \ne 0}\right),$$

$$p_1 = p_3 = \frac{n_{\text{new}}}{n_{\text{new}}n_0 - n_{\text{noun}}n_0 + n_{\text{noun}}n_{\text{new}}},$$

$$p_2 = \frac{1}{n_{\text{verb}}} \cdot \frac{n_{\text{new}}n_{\text{s}}}{n_{\text{new}}n_0 - n_{\text{noun}}n_0 + n_{\text{noun}}n_{\text{new}}},$$

$$p_4 = p_5 = p_6 = \frac{n_{\text{new}} - n_{\text{noun}}}{n_{\text{new}}n_0 - n_{\text{noun}}n_0 + n_{\text{noun}}n_{\text{new}}}.$$

2.2. Algorithm of topic modeling of appeals represented by sets of "noun–verb" pairs

Input: T_m topic-model dictionary (can be empty) containing at most w different noun–verb pairs (w is the size of the "window"); a reference represented as a list of ordered noun–verb pairs (noun comes first, tokens are lemmatized).

Output: T_{m+1} dictionary containing at most w distinct pairs, in which all pairs from $m + 1$ reference were given probabilities of belonging to topics expressed by nouns.

Procedure:

1. We set $n_0 = n_{\text{new}} = 0$.
2. For all s pairs in a reference.
3. We look through the T_m dictionary and find:

   a) values $\hat{p}_s$ that s has in T_m, and the corresponding topics $t_{sp}^1, ..., t_{sp}^n$;

   b) topics $t_s^1, ..., t_s^{n_{\text{noun}}}$ matching the noun from s, but not containing s in its entirety;

   c) topics $t_{\text{verb}}^1, ..., t_{\text{verb}}^{n_{\text{verb}}}$ not matching the ones in (a) and (b) that have at least one verb from s;
4. If step 3 did not yield any matches, we increase n_0 by 1; if, in addition, there is no noun from s among the new topics already created (or no new topics yet), we increase n_{new} на 1 from it by copying and then work with it. Topic: $t_{\text{new}}^{n_{\text{new}}}$;

5. After viewing of the dictionary T_m is completed, we create a dictionary T_{m+1} from it by copying and then work with it.
6. We find

$$p_0 = \frac{1}{n_0}\left(1 - \frac{1}{\hat{p}}\min_1 \hat{p}_1\Big|_{\hat{p}_1 \neq 0}\right);$$

7. We write into the topics $t_{sp}^1, ..., t_{sp}^n$ with probabilities p_0 the pairs s found at step 3a).
8. In the topics $t_s^1, ..., t_s^{n_\text{noun}}$ we write with probabilities $p = \dfrac{n_\text{new}}{n_\text{new} n_0 - n_\text{noun} n_0 + n_\text{noun} n_\text{new}}$ the pairs s found at step 3b).
9. In the topics $t_\text{verb}^1, ..., t_\text{verb}^{n_\text{verb}}$ we write with probabilities $p_2 = \dfrac{1}{n_\text{verb}} p$ the pairs s found at step 3c).
10. In the topics $t_\text{new}^1, ..., t_\text{new}^{n_\text{new}}$ we write:
    a) with probabilities p, which nouns match the topic;
    b) with probabilities $q = \dfrac{n_\text{new} - n_\text{noun}}{n_\text{new} n_0 - n_\text{noun} n_0 + n_\text{noun} n_\text{new}}$ all other pairs of the reference.
11. If the number of pairs W in the dictionary is greater than w, we find the topic with the lowest weight and remove all or $(W - w)$ pairs (whichever is less) from it. We do this until all $(W - w)$ redundant pairs have been removed from the dictionary.
12. We compute the weights of the remaining topics and sort them.

Following this algorithm, we will obtain in each iteration the most weighted topics sorted by weight.

2.3. Results of topic modeling

In order to test the algorithm, the previously mentioned dataset was selected. Repetitive messages and messages that did not contain nouns with verbs were removed. The working dataset contained about 2700 messages sent to the mayor of Belgorod and city departments over a year. The data was anonymized. The window size (in pairs simultaneously in the dictionary) was assumed to be $w = 300$.

The figure shows a graph of the change in the weight of the top vocabulary topics throughout the year.

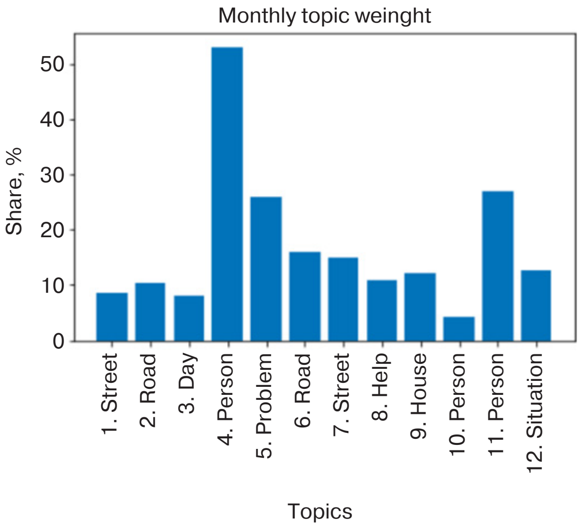

Figure. The most popular topics in references to the Mayor's Office of the city of Belgorod

In general, the topics selected by the model correspond to the most important issues of concern to citizens over a certain period of time. This is not seen directly in the topics, because the topic in the context of this paper is a single word (noun). It is difficult to assess the importance of messages. However, a topic is a key to a set of messages and if that set is such, and changes in such a way that the weight of the topic increases, then the messages in the set are worthy of attention. The administration worked on the posts grouped around the topic leaders, and then other topics came to the top. This means that the people (administration) and the machine (proposed system) reacted correctly to the mood in the city. Otherwise there would have been topics that remained at the top for a long time. According to the assessment of the city administration specialists, the topic modeling was carried out correctly.

Metrics can be used to quantify the quality of topic modeling, such as the BCubed metric[17] [15], the use of which is justified by the construction of the algorithm. If we denote by $p_{ts.mon}$ the probability that a message s related to month mon belongs to topic t $\left(\sum_t p_{ts.mon} = 1\right)$, then BC precision ($BCP$) and BC recall ($BCR$) for the topic t in month mon will be defined by the formulae:

$$BCP(t, mon) = \frac{\sum_s p_{ts.mon}}{\sum_{s,k} p_{tsk}},$$

$$BCR(t, mon) = \frac{\sum_s p_{ts.mon}}{\sum_{r,k} p_{rsk}}.$$

For the dataset from the experiment, the first metric of top topics always exceeded 55%, while the second metric ranged from 27% to 83% (respectively the weight of topics shown in the figure). Given that the number of

[17] The BCubed family of metrics is implemented in the library https://pypi.org/project/bcubed-metrics/. Accessed December 02, 2024.

topics per month was never less than 70, the quality of the clustering can be recognized as high for the metric as well: random selection would result in a rate of about 1.5%.

CONCLUSIONS

In this study, we proposed a topic modeling method and algorithm for short messages. Clustering of short messages is a challenging task because such messages are very difficult to map to any context, i.e., large corpora of text on which the model can be trained do not usually provide satisfactory descriptions of topics.

The work is based on a previously proposed methodology of message meaning contour extraction in Russian texts based on noun–verb pairs.

Using the proposed methodology, a topic model was built and simulated on real data. The constructed clustering showed a relatively high level of quality by BCubed metric. At the same time, the result is also visible in qualitative evaluation. If a topic is calculated as a top topic in a particular period, the issues raised in it deserve to be prioritized and acted upon by people. In the example dataset used in the experiment, such a correlation (between the suggested set of messages for prioritized response and the messages people selected) was greater than 70%. However, even without quantification, the employees of the organization owning the dataset expressed their willingness to use the software built on the proposed algorithm as a decision support system (more precisely, an advising system for priority response to messages). In their opinion, this would greatly reduce message processing time.

ACKNOWLEDGMENTS

This work was realized in the framework of the Priority 2030 Program using the equipment of High Technology Center at the V.G. Shukhov Belgorod State Technological University.

REFERENCES

1. Brusentsev A.G., Zueva E.S. Thematic models and tools for processing the natural language in application to the problems of municipal structures. In: *Actual Theoretical and Applied Issues of the Socio-Economic Systems Management: Proc. Second International Scientific and Practical Conference.* Moscow; 2020. V. 2. P. 262–269 (in Russ.). https://elibrary.ru/fkgyxn
2. Zueva E.S. Probabilistic classification of incoming calls based on a controlled recurrent neurons algorithm. In: *Proc. International Scientific and Technical Conference of Young Scientists of V.G. Shukhov BSTU.* Belgorod: V.G. Shukhov BSTU; 2021. P. 3564–3575 (in Russ.). https://www.elibrary.ru/nhlzpv
3. Polyakov V.M., Mozaidze E.S. Collaborative filtering algorithm as a possible tool for detecting a dangerous tweet (short message) in social networks of a representative office of the government of the Belgorod region. In: *Modern Issues of Sustainable Development of Society in the Era of Transformation Processes*: *Collection of Materials of the 4th International Scientific and Practical conference*. Moscow; 2022. P. 136–148 (in Russ.). https://doi.org/10.34755/IROK.2022.14.90.027, https://www.elibrary.ru/mzrsgm
4. Papadimitriou C.H., Tamaki H., Raghavan P., Vempala S. Latent semantic indexing: A probabilistic analysis. In: *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems.* ACM; 1998. P. 159–168. https://doi.org/10.1145/275487.275505
5. Hofmann T. Probabilistic latent semantic indexing. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM; 1999. P. 50–57. https://doi.org/10.1145/312624.312649
6. Blei D., McAuliffe J. Supervised topic models. In: *Advances in Neural Information Processing Systems 20* (*NIPS 2007*). 2008. P. 121–128.
7. Blei D.M., Lafferty J.D. Dynamic topic models. In: *Proceedings of the 23rd International Conference on Machine learning* (*ICML '06*). ACM; 2006. P. 113–120. https://doi.org/10.1145/1143844.1143859
8. Blei D.M. Probabilistic topic models. *Communications of the ACM*. 2012;55(4):77–84. https://doi.org/10.1145/2133806.2133826
9. Vorontsov K.V. Additive regularization for topic models of text collections. *Dokl. Math.* 2014;89(3):301–304. https://doi.org/10.1134/S1064562414020185
[Original Russian Text: Additive regularization for topic models of text collections. *Doklady Akademii Nauk*. 2014;456(3): 268–271 (in Russ.). https://doi.org/10.7868/S0869565214090096]
10. Vorontsov K.V., Potapenko A.A. EM-like algorithms for probabilistic topic modeling*. Mashinnoe obuchenie i analiz dannykh = Machine Learning and Data Analysis*. 20131(6):657–686 (in Russ).
11. Nokel M.A., Lukashevich N.V. Topic Models: Adding Bigrams and Taking Account of the Similarity between Unigramsand Bigrams*. Vychislitel'nye metody i programmirovanie = Numerical Methods and Programming*. 2015;16(2):215–234 (in Russ.). https://doi.org/10.26089/NumMet.v16r222
12. Korshunov A., Gomzin A. Topic modeling in natural language texts. *Trudy Instituta sistemnogo programmirovaniya RAN* (*Trudy ISP RAN*) = *Proceedings of the Institute for System Programming of the RAS* (*Proceedings of ISP RAS*). 2012;23: 215–240 (in Russ.). https://doi.org/10.15514/ISPRAS-2012-23-13

13. Nakshatri N., Liu S., Chen S., Roth D., Goldwasser D., Hopkins D. Using LLM for Improving Key Event Discovery: Temporal-Guided News Stream Clustering with Event Summaries. *Findings of the Association for Computational Linguistics: EMNLP*. 2023:4162–4173. https://doi.org/10.18653/v1/2023.findings-emnlp.274
14. Rijcken E., Scheepers F., Zervanou K., Spruit M., Mosteiro P., Kaymak U. Towards Interpreting Topic Models with ChatGPT. *2023. Paper presented at The 20th World Congress of the International Fuzzy Systems Association*, Daegu, Republic of Korea. 2023. V. 5. Available from URL: https://pure.tue.nl/ws/portalfiles/portal/300364784/IFSA_InterpretingTopicModelsWithChatGPT.pdf
15. Amigo E., Gonzalo J., Artiles J., Verdejo F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*. 2009;12(4):461486.

СПИСОК ЛИТЕРАТУРЫ

1. Брусенцев А.Г., Зуева Е.С. Тематические модели и инструменты обработки естественного языка в применении к задачам муниципальных структур. В сб.: *Актуальные теоретические и прикладные вопросы управления социально-экономическими системами: материалы II Международной научно-практической конференции.* М.: Институт развития дополнительного профессионального образования; 2020. Т. 2. С. 262–269. https://elibrary.ru/fkgyxn
2. Зуева Е.С. Вероятностная классификация входящих обращений на основе алгоритма управляемых рекуррентных нейронов. В сб.: *Материалы Международной научно-технической конференции молодых ученых БГТУ им. В.Г. Шухова.* Белгород: БГТУ им. В.Г. Шухова; 2021. С. 3564–3575. https://www.elibrary.ru/nhlzpv
3. Поляков В.М., Мозаидзе Е.С. Алгоритм коллаборативной фильтрации как возможный инструмент выявления опасного твита (короткого сообщения) в социальных сетях представительства органа государственной власти Белгородской области. В сб.: Современные вопросы устойчивого развития общества в эпоху трансформационных процессов: *материалы IV международной научно-практической конференции.* М.: ООО «ИРОК»; 2022. С. 136–148. https://doi.org/10.34755/IROK.2022.14.90.027, https://www.elibrary.ru/mzrsgm
4. Papadimitriou C.H., Tamaki H., Raghavan P., Vempala S. Latent semantic indexing: A probabilistic analysis. In: *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems.* ACM; 1998. P. 159–168. https://doi.org/10.1145/275487.275505
5. Hofmann T. Probabilistic latent semantic indexing. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.* ACM; 1999. P. 50–57. https://doi.org/10.1145/312624.312649
6. Blei D., McAuliffe J. Supervised topic models. In: *Advances in Neural Information Processing Systems 20* (*NIPS 2007*). 2008. P. 121–128.
7. Blei D.M., Lafferty J.D. Dynamic topic models. In: *Proceedings of the 23rd International Conference on Machine learning* (*ICML '06*). ACM; 2006. P. 113–120. https://doi.org/10.1145/1143844.1143859
8. Blei D.M. Probabilistic topic models. *Communications of the ACM*. 2012;55(4):77–84. https://doi.org/10.1145/2133806.2133826
9. Воронцов К.В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. *Доклады академии наук*. 2014;456(3):268–271. https://doi.org/10.7868/S0869565214090096
10. Воронцов К.В., Потапенко А.А. Модификации EM-алгоритма для вероятностного тематического моделирования. *Машинное обучение и анализ данных*. 2013;1(6):657–686.
11. Нокель М.А., Лукашевич Н.В. Тематические модели: добавление биграмм и учет сходства между униграммами и биграммами. *Вычислительные методы и программирование*. 2015;16(2):215–234. https://doi.org/10.26089/NumMet.v16r222
12. Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке. *Труды Института системного программирования РАН* (*Труды ИСП РАН*). 2012;23:215–242. https://doi.org/10.15514/ISPRAS-2012-23-13
13. Nakshatri N., Liu S., Chen S., Roth D., Goldwasser D., Hopkins D. Using LLM for Improving Key Event Discovery: Temporal-Guided News Stream Clustering with Event Summaries. *Findings of the Association for Computational Linguistics: EMNLP*. 2023:4162–4173. https://doi.org/10.18653/v1/2023.findings-emnlp.274
14. Rijcken E., Scheepers F., Zervanou K., Spruit M., Mosteiro P., Kaymak U. Towards Interpreting Topic Models with ChatGPT. *2023. Paper presented at The 20th World Congress of the International Fuzzy Systems Association*, Daegu, Republic of Korea. 2023. V. 5. URL: https://pure.tue.nl/ws/portalfiles/portal/300364784/IFSA_InterpretingTopicModelsWithChatGPT.pdf
15. Amigo E., Gonzalo J., Artiles J., Verdejo F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*. 2009;12(4):461486.

About the author

Elena S. Mozaidze, Postgraduate Student, Department of Computer Software and Automated Systems, V.G. Shukhov Belgorod State Technological University (46, Kostyukova ul., Belgorod, 308012 Russia). E-mail: mozaidze95@mail.ru. https://orcid.org/0000-0002-7919-7963

Об авторе

Мозаидзе Елена Сергеевна, аспирант, кафедра программного обеспечения вычислительной техники и автоматизированных систем, ФГБОУ ВО «Белгородский государственный технологический университет им. В.Г. Шухова» (308012, Россия, Белгород, ул. Костюкова, д. 46). E-mail: mozaidze95@mail.ru. https://orcid.org/0000-0002-7919-7963

Translated from Russian into English by Lyudmila O. Bychkova
Edited for English language and spelling by Dr. David Mossop