

Научная статья

УДК 004.912; 930.253; 93/94

DOI 10.25205/1818-7935-2025-23-3-107-122

Тематическое моделирование научных текстов с помощью BERTopic (на материале аннотаций статей о Великой Отечественной войне, опубликованных в 2014–2023 гг.)

Зинаида Николаевна Сокова¹
Валерий Михайлович Кружинов²
Анна Валерьевна Глазкова³

Тюменский государственный университет,
Тюмень, Россия

¹z.n.sokova@utmn.ru, <https://orcid.org/0000-0003-4366-3771>

²v.m.kruzhinov@utmn.ru, <https://orcid.org/0000-0002-2947-5590>

³a.v.glazkova@utmn.ru, <https://orcid.org/0000-0001-8409-6457>

Аннотация

В условиях постоянного роста объемов научной литературы разработка эффективных средств систематизации и автоматизированного анализа научных текстов становится критически важной задачей для обеспечения доступности, структурированности и упрощения информации в различных областях науки. Одним из эффективных подходов к автоматизированному анализу коллекции текстов является тематическое моделирование, которое позволяет выявлять ключевые темы в больших массивах текстов, упрощая их классификацию и ускоряя поиск релевантной информации, что особенно важно для анализа научной литературы и мониторинга исследовательских тенденций. Авторы данной работы исследуют возможности алгоритма тематического моделирования BERTopic для анализа исторических научных текстов на русском языке. В качестве материала для исследования были выбраны тексты научных статей о Великой Отечественной войне, опубликованных в российских периодических изданиях последнего десятилетия (2014–2023 гг.). Актуальность данной работы обусловлена, с одной стороны, малым количеством исследований применения алгоритмов тематического моделирования, основанных на использовании больших языковых моделей, для анализа исторических научных текстов и, с другой стороны, необходимостью развития средств поиска и отбора научных источников. Статья содержит описание процессов сбора коллекции текстов, ее предобработки и построения тематической модели. Полученные результаты иллюстрируют наиболее распространенные тематики научных работ и представляют интерес как для исследователей в области компьютерной лингвистики, так и для специалистов по истории России и отечественной историографии. Представленный в работе подход к формированию коллекции научных текстов, основанный на сочетании использования слов-маркеров, инструментов автоматического сбора текстов и экспертной проверки, может быть использован для сбора научных текстов другой тематики в похожих исследованиях. Тематическое моделирование позволило выделить основные темы, представленные в коллекции текстов, а также ключевые слова, которые их описывают, построить иерархическое и динамическое представление тем. К преимуществам BERTopic можно отнести использование векторных представлений текстов, полученных из больших языковых моделей, а также разнообразие инструментов для визуализации тем.

Ключевые слова

тематическое моделирование, компьютерная лингвистика, научные тексты, BERTopic, Великая Отечественная война, историография

Для цитирования

Сокова З. Н., Кружинов В. М., Глазкова А. В. Тематическое моделирование научных текстов с помощью BERTopic (на материале аннотаций статей о Великой Отечественной войне, опубликованных в 2014–2023 гг.) // Вестник

© Сокова З. Н., Кружинов В. М., Глазкова А. В., 2025

НГУ. Серия: Лингвистика и межкультурная коммуникация. 2025. Т. 23, № 3. С. 107–122. DOI 10.25205/1818-7935-2025-23-3-107-122

Topic Modeling of Scientific Texts Using BERTopic (Based on Scientific Abstracts on the Great Patriotic War published in 2014–2023)

Zinaida N. Sokova¹, Valery M. Kruzhinov²
Anna V. Glazkova³

Tyumen State University,
Tyumen, Russian Federation

¹z.n.sokova@utmn.ru, <https://orcid.org/0000-0003-4366-3771>

²v.m.kruzhinov@utmn.ru, <https://orcid.org/0000-0002-2947-5590>

³a.v.glazkova@utmn.ru, <https://orcid.org/0000-0001-8409-6457>

Abstract

Amid the continuous growth of the volume of scientific literature, the development of efficient tools for systematizing and automating the analysis of scientific texts has become a critical challenge. These tools are essential to ensure the accessibility, organization, and information retrieval across various scientific domains. Topic modeling is one of the most effective approaches for automated text analysis since it enables the identification of key themes within large collections of texts, facilitating classification and speeding up the search for relevant information. This is particularly important for analyzing scientific literature and tracking research trends. In this study, the authors consider the capabilities of the BERTopic algorithm for topic modeling of Russian-language scientific texts. The dataset consists of scientific articles on the Great Patriotic War, published in Russian academic journals over the past decade (2014–2023). The relevance of this work arises from the limited research on the application of topic modeling algorithms to scientific texts, especially those that utilize large language models. There is a notable gap in the studies focusing on this approach. Additionally, there is an urgent need in more advanced methods for searching and selecting scholarly sources. This further highlights the importance of the current study. The paper provides a detailed description of the data collection, preprocessing, and topic modeling processes. The results reveal the most prevalent topics in academic works, offering valuable insights for researchers in computational linguistics, as well as for historians and scholars of Russian historiography. The approach to corpus building, combining the use of keyword markers, automated data collection tools, and expert validation, can be applied to other subject areas in similar studies. Topic modeling allowed for the identification of key themes within the corpus, along with the keywords that characterize them, and facilitated the creation of both hierarchical and dynamic representations of these topics. BERTopic's strengths lie in its use of vector-based text representations from large language models, as well as its extensive tools for visualizing the resulting topics.

Keywords

topic modeling, computational linguistics, scientific texts, Great Patriotic War, historiography, methods of historical studies

For citation

Sokova Z. N., Kruzhinov V. M., Glazkova A. V. Topic Modeling of Scientific Texts Using BERTopic (Based on Scientific Abstracts on the Great Patriotic War published in 2014–2023). *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2025, vol. 23, no. 3, pp. 107–122. DOI 10.25205/1818-7935-2025-23-3-107-122

Введение

В последние десятилетия наблюдается значительное развитие компьютерных методов анализа текстов на естественном языке, что обусловлено стремительным ростом объемов публикуемых данных. Особую актуальность эти методы приобретают в области научных текстов, где точность и полнота обработки позволяют не только ускорить поиск и систематизацию информации, но и улучшить качество текстовой аналитики, что особенно важно для проведения междисциплинарных исследований и улучшения понимания научных процессов. Таким обра-

зом, внедрение компьютерных методов для анализа научных текстов становится необходимым условием для прогресса в различных научных областях.

Одним из эффективных подходов к систематизации текстов служит тематическое моделирование. Тематическое моделирование – это способ анализа текстов, который позволяет определить основную тему, к которой относится каждый из них. В ходе тематического моделирования тексты, связанные одной тематикой, объединяются в рамках одной темы [Owa, 2021; Robledo, Zuluaga; 2022]. Каждая тема характеризуется набором ключевых слов, которые описывают ее основное содержание. Текст в тематическом моделировании представляет собой смесь тем. Обычно он с разными вероятностями относится к нескольким темам одновременно. Аналогично ключевое слово, входящее в описание одной из тем, может также быть ключевым словом, описывающим другую тему. Тематическое моделирование находит применение в широком спектре задач: классификация текстов, автоматическое реферирование, тематический поиск, рубрикация и фильтрация контента, анализ новостных потоков и т. д. [Коршунов, Гомзин, 2012]. Тематика русскоязычных текстов, анализируемых в различных исследованиях с применением тематического моделирования, также очень широкая: художественные тексты [Zamiraylova, Mitrofanova, 2020; Кирина, 2022; Колмогорова, Залевская, 2023], медицинские тексты [Землянский и др., 2021], юридические документы [Митрофанова, Атугодаге, 2023], новости [Тен, 2023], тексты социальных сетей [Bodrunova et al., 2019; Maltseva et al., 2021; Попова и др., 2023], методические материалы [Тверской, Канева, 2023], научные и учебные тексты [Solovyev et al., 2023; Sakhovskiy et al., 2020], стенограммы публичных выступлений [Chizhik, Sergeyev, 2021].

В данной работе изучены возможности алгоритма BERTopic для тематического моделирования исторических научных текстов на русском языке. В качестве материала для исследования выбраны статьи, опубликованные в российских периодических изданиях по историческим наукам в течение последних десяти лет и посвященные тематике Великой Отечественной войны. Описываются этапы сбора текстов: первичный автоматический сбор статей; подход к отбору статей о Великой Отечественной войне с помощью слов-маркеров и экспертной оценки. Далее описывается процесс тематического моделирования собранной коллекции текстов с помощью библиотеки BERTopic и языка программирования Python. В завершение анализируются полученные результаты: рассматриваются темы статей, полученные тематической моделью, динамика их распространенности на протяжении десяти лет и семантические кластеры, образуемые темами.

Выбор тематики научных статей в данном исследовании обусловлен тем, что Великая Отечественная война является одной из основных составляющих исторического сознания русского народа, ее память почитается как символ национального единства и гордости. Тема Великой Отечественной войны привлекает большое внимание исследователей отечественной истории. Данные исследования являются важным материалом для сохранения и передачи знаний о прошлом, уважения к памяти погибших и извлечения уроков для будущего. Работы по истории Великой Отечественной войны охватывают широкий спектр тем и публикуются в большом количестве изданий. В преддверии юбилея Победы в 2025 г. для получения более целостной картины представления Великой Отечественной войны в российской историографии актуальным представляется использование современных методов компьютерной лингвистики и обработки естественного языка.

Таким образом, данное исследование актуально для развития методологии как обработки научных текстов, так и исторической науки. Результаты тематического моделирования для проведения исторических исследований рассматривались ранее в работах [Glazkova et al., 2019; Воронцов и др., 2022; Галушко, 2023; Колмогорова и др., 2024]. Автор статьи [Кузнецов, 2022] подчеркивает, что стремительное развитие информационных технологий изменило характер исследований в исторической науке. При этом тематическое моделирование, хотя оно и не является самостоятельным методом исследования, в сочетании с традиционными метода-

ми анализа исторических источников становится эффективным подходом к изучению научных текстов.

1. Методы

1.1. Сбор текстов

Для построения тематической модели текстов научных статей о Великой Отечественной войне были проанализированы 13 ведущих российских периодических изданий, посвященных исторической тематике. Для рассмотрения выбраны статьи, опубликованные в период с 2014 по 2023 г. в периодических изданиях, которые по состоянию на май 2024 г. входили в базу Russian Science Citation Index (RSCI) на платформе Web of Science и публиковали статьи по отечественной истории. Перечень изданий был составлен исследователями в области исторических наук и включает следующие журналы:

- Вестник Волгоградского государственного университета. Серия 4: История. Регионоведение. Международные отношения;
- Вестник Московского университета. Серия 8. История;
- Вестник Новосибирского государственного университета. Серия: История. Филология;
- Вестник Омского университета. Серия «Исторические науки»;
- Вестник Пермского университета. Серия: История;
- Вестник Православного Свято-Тихоновского гуманитарного университета. Серия 2: История. История Русской Православной Церкви;
- Вестник Санкт-Петербургского университета. История;
- Вестник Томского государственного университета. История;
- Военно-исторический журнал;
- Новейшая история России;
- Петербургский исторический журнал;
- Российская история;
- Уральский исторический вестник.

Для статей, опубликованных в перечисленных периодических изданиях, были собраны заголовки, списки ключевых слов и аннотации, размещенные в открытом доступе в библиотеке «Киберленинка» и на сайтах журналов. Для последующей обработки и анализа заголовков статьи, список ее ключевых слов и аннотация были объединены в один текст. В работе рассматриваются аннотации статей, а не их полные тексты. Аннотации статей находятся в открытом доступе, что облегчает сбор данных [Краснов и др., 2019]. Кроме того, сравнение, представленное в работе [Syed, Spruit, 2017], показало, что значение метрики когерентности, рассчитанное для тематических моделей, построенных с помощью аннотаций, сравнимо со значением данной метрики для моделей, использующих полные тексты статей.

Сбор текстов проводился в несколько этапов. На первом этапе были получены все тексты перечисленных периодических изданий, доступные в свободном доступе и входящие в заданный временной диапазон. Сбор текстов проводился в автоматическом режиме с помощью библиотек BeautifulSoup и Requests и языка программирования Python в мае и июне 2024 г. Количество текстов на данном этапе составило больше десяти тысяч. На втором этапе среди собранных текстов был выполнен поиск статей, посвященных Великой Отечественной войне. Для этого на основании эмпирического анализа публикаций, размещенных в научной электронной библиотеке eLIBRARY.ru, экспертами были составлены два перечня слов-маркеров: а) однозначно идентифицирующие тематику Великой Отечественной войны и б) возможно определяющие тематику Великой Отечественной войны. Перечни слов-маркеров представлены в табл. 1. В собранных текстах был произведен поиск слов-маркеров из первого списка

и производных от них, перед проведением поиска тексты и ключевые слова были лемматизированы. На основании найденных текстов была сформирована коллекция текстов статей о Великой Отечественной войне. Далее аналогично был проведен поиск текстов, включающих в себя слова-маркеры из второго списка. Тексты, включающие в себя слова-маркеры из второго списка, были прочитаны экспертами и, если они соответствовали исходной тематике, добавлялись в коллекцию текстов о Великой Отечественной войне. Наконец, на завершающем этапе собранная коллекция текстов была повторно проверена экспертами. Итоговый размер коллекции составил 544 текста, или 88 657 токенов (в соответствии с токенизатором NLTK [Bird, 2006]).

Таблица 1

Перечни слов-маркеров для поиска статей по тематике BOB

Table 1

The lists of key-phrases for finding papers on the Great Patriotic War

Ключевые слова, однозначно идентифицирующие тематику BOB	Ключевые слова, возможно идентифицирующие тематику BOB
Великая Отечественная война, Договор о ненападении с Германией, план «Барбаросса», ГКО, Ставка Верховного командования, ЦШПД, блокада Ленинграда, Дорога жизни, коренной перелом, Смоленское сражение, Московская битва, Сталинградская битва, Курская битва, форсирование Днепра, операция «Багратион», Берлинская операция, Антигитлеровская коалиция, Тегеранская конференция, Ялтинская конференция, Потсдамская конференция, 1941–1945	фашизм, война, эвакуация, И. В. Сталин, Г. К. Жуков, советские полководцы, фронты, Красная армия, вермахт, солдатская повседневность, штрафные батальоны, партизаны, тыл, соревнование, колхозы, безвозвратные потери, санитарные потери, Победа

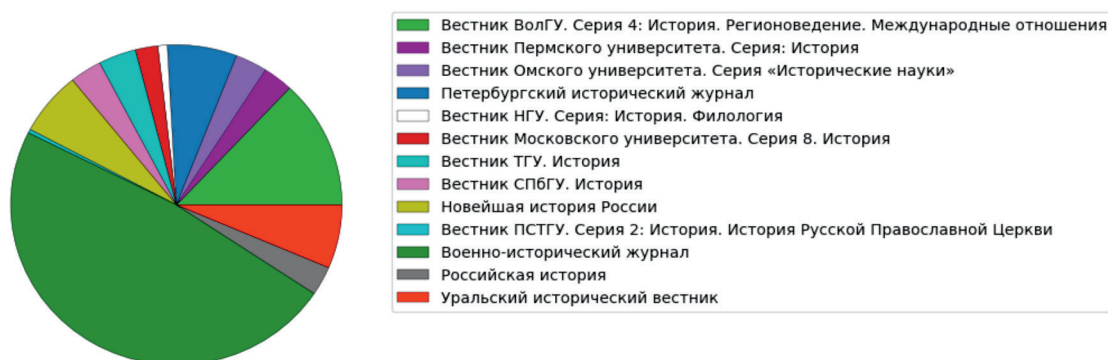


Рис. 1. Распределение количества статей по тематике BOB в периодических изданиях, выбранных для исследования

Fig. 1. The distribution of the number of papers on the Great Patriotic War in periodicals selected for the study

Распределение собранных текстов по периодическим изданиям и по годам представлено на рис. 1 и 2. Наибольшее количество статей по тематике Великой Отечественной войны в рассматриваемом временном диапазоне было опубликовано в «Военно-историческом журнале». Количество статей, опубликованных в разные годы, варьируется в собранной коллекции

текстов от 24 до 91. Средний размер текста, объединяющего в себе заголовок статьи, список ключевых слов и аннотацию, составляет 161 токен. При этом максимальный размер текста в коллекции составляет 633 токена, а минимальный – 33 токена. Среднее количество предложений в собранных текстах – 7, минимальное количество – 1, максимальное – 25. Среднее квадратическое отклонение значений длины текста равно 108 токенам, или 5 предложениям. Количество токенов и предложений определено с помощью методов `word_tokenize()` и `sent_tokenize()` из библиотеки NLTK и языка программирования Python. Распределение длины собранных текстов визуализировано на рис. 3.



Рис. 2. Распределение количества собранных статей по тематике BOB по годам

Fig. 2. The distribution of the number of papers on the Great Patriotic Warper year

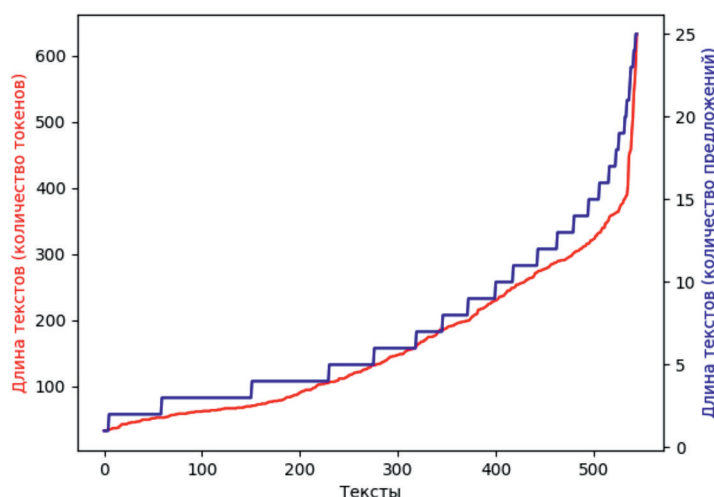


Рис. 3. Распределение длины собранных текстов

Fig. 3. The distribution of the length for collected texts

1.2. Тематическое моделирование

Собранная коллекция текстов послужила материалом для построения тематической модели. Тематическое моделирование представляет собой широко распространенный подход к анализу тематического содержания коллекций текстов на естественном языке. Тематическое моделирование является статистическим инструментом. Тема в тематическом моделировании – это результат бикластеризации, т. е. одновременной кластеризации слов и документов по их семантической близости [Коршунов, Гомзин, 2012]. Таким образом, тема объединяет в себе близкие по смыслу тексты, содержащие семантически связанные слова.

Существует большое количество методов тематического моделирования, однако не все они являются широко распространенными. К часто используемым относятся алгебраические (LSA, NMF), вероятностные (LDA) и нейросетевые методы (LDA2Vec, BERTopic) [Abdelrazek et al., 2023]. Современные нейросетевые методы основаны на применении предварительно обученных больших языковых моделей [Devlin et al., 2019], которые используют контекстуализированные векторные представления слов, построенные на материале больших текстовых корпусов. Как показывают обзоры статей, представленные в работах [Bodrunova, 2021; Babina, 2024], большая часть существующих исследований тематического моделирования для русскоязычных текстов используют классический метод LDA.

В данной работе используется модель BERTopic [Grootendorst, 2022]. При построении тематической модели BERTopic сначала строит векторные представления текстов. Уменьшив размерность полученных векторов с помощью метода UMAP, модель формирует кластеры из полученных векторов, после чего для каждого кластера выделяются ключевые слова с помощью модифицированной меры TF-IDF. Преимуществами BERTopic является семантическая обоснованность, обусловленная использованием больших языковых моделей, а также широкие возможности для визуализации полученных тем.

Поскольку построение тематической модели требует предварительной обработки коллекции текстов, после сбора и статистического анализа текстов были выполнены следующие действия:

- 1) удаление специальных символов и знаков препинания;
- 2) приведение текста к нижнему регистру;
- 3) удаление слов, не несущих в себе значимой информации, с помощью списка стоп-слов для русского языка из библиотеки NLTK [Bird, 2006], а также слов, не включающих в себя кириллические символы;
- 4) приведение слов в нормальную форму (лемматизация) с использованием библиотеки `ru morphology2` [Korobov, 2015].

Предварительно обработанные тексты использовались для построения тематической модели с помощью библиотеки BERTopic. Для построения тематической модели использовались следующие параметры метода UMAP: количество ближайших соседей – 6, размерность итоговых данных – 5, минимальное расстояние для точек в итоговом представлении меньшей размерности – 0, метрика – косинусное расстояние. Параметры кластеризации: минимальный размер кластера – 10 текстов, метрика – евклидово расстояние, метод выбора кластеров – Excess of Mass (EoM). На основании исходных текстов были построены n-граммы размером [1;3]. При построении были исключены самые распространенные n-граммы, встретившиеся в 85 % текстов и чаще, а также редкие n-граммы, которые встретились меньше чем в двух текстах. В качестве базовой лингвистической модели используется мультязычная модель `paraphrase-multilingual-MiniLM-L12-v2` [Reimers, Gurevych, 2019].

2. Результаты

В результате тематического моделирования коллекции текстов научных статей о Великой Отечественной войне в ней были выделены 16 тем. Для оценки качества тематической моде-

ли использовался метод экспертной оценки, описанный в работе [Нокель, Лукашевич, 2015]. Экспертам-историкам были представлены полученные темы в виде списков описывающих их ключевых слов, упорядоченных по убыванию значимости. Экспертам предлагалось определить, является ли каждая тема интерпретируемой, что подразумевает возможность присвоения ей обобщенного названия. Анализ результатов показал, что всем 16 темам удалось присвоить обобщающие названия, что свидетельствует о высокой интерпретируемости и объяснимости выделенных тем экспертами.

В табл. 2 представлен список полученных тем, соответствующих им ключевых слов и количество текстов, в которых эта тема является основной. Столбец «Ключевые слова» содержит существительные и именные словосочетания, описывающие тему и выявленные тематической моделью. В столбце «Тема» представлены обобщающие названия тем, подобранные экспертами на основании списков ключевых слов и текстов статей, относящихся к данной теме.

Как видно из данных табл. 2, самый крупный кластер образуют статьи, связанные с темой сохранения исторической памяти о Великой Отечественной войне. К данной теме относятся статьи, описывающие образ Великой Отечественной войны в художественной литературе и периодической печати, анализирующие мемуарную литературу, а также статьи, посвященные проблемам фальсификации исторических событий в научной и учебной литературе. Второй по распространенности является тема, объединяющая в себя разноплановые исследования истории блокады Ленинграда. Третья по распространенности тема описывает жизнь в тылу в период Великой Отечественной войны: вопросы жизнеобеспеченности населения, особенностей быта разных слоев населения, административного управления. Распространены темы, связанные с анализом различных аспектов ведения боевых действий, в частности, описанием событий, связанных с различными видами войск (темы 4, 8), вопросами обеспечения фронта и командования (темы 5, 9), анализом действий и сил противника (тема 11), партизанского движения (тема 14). Значительная часть исследований посвящена Сталинградской битве (тема 7). Часть тем описывает демографическую ситуацию в период Великой Отечественной войны и проблемы социальной сферы (темы 6, 10).

Стоит подчеркнуть, что в тематическом моделировании текст, как правило, представляет собой смесь тем. То есть одна статья может включать несколько из перечисленных тем, а также множество других терминов и аспектов, которые в силу более редкой встречаемости в коллекции не были объединены моделью в отдельную тему. Например, статья может описывать демографические последствия Сталинградской битвы (сочетая темы 6 и 7) [Кринко, Медведев, 2018], организацию санитарно-эпидемиологических мероприятий в Сталинграде (темы 6 и 15) [Епифанов, Красноженова, 2023] или представлять собой смесь нескольких семантически близких тем (в частности, тем 3, 6, 10) [Снегирева, 2016; Романов, 2016]. Значение в столбце «Количество текстов» в табл. 2 определяется как количество статей в коллекции текстов, для которых данная тема является наиболее выраженной. Такое представление дает общее понимание о тематике статей в статистическом смысле, однако для многих статей из собранной коллекции ни одна из 16 выделенных тем не является преобладающей. Это говорит о достаточно большом тематическом разнообразии анализируемых работ.

Функционал библиотеки BERTopic дает возможность построить иерархическое представление тематики коллекции текстов на основании семантической близости выявленных тем (рис. 4). Темы образуют древовидную структуру, где на более низких уровнях дерева (слева на рисунке) находятся отдельные темы, которые объединяются в более крупные кластеры. В рассматриваемой коллекции текстов выделяются два крупных тематических кластера. Один из них преимущественно связан с непосредственным ведением боевых действий (выделен красным цветом на рис. 4). Другой кластер объединяет темы, описывающие повседневную жизнь народа, бытовые условия, сохранение исторической памяти (выделен зеленым цветом).

Таблица 2

Список наиболее частотных тем

Table 2

The list of the most common topics

№ темы	Тема	Ключевые слова	Количество текстов
1	Сохранение памяти о ВОВ	память, Победа, фальсификация, Вторая мировая, история Великой Отечественной, историческая память	34
2	Блокада Ленинграда	Ленинград, блокада, блокада Ленинграда, город, голод, карточка, снабжение	32
3	Жизнь в тылу	социальная политика, экономика, изучение, предприятие, власть, население, пропаганда, советский тыл	30
4	Авиация	авиация, самолет, истребитель, оборона, противовоздушная оборона, снаряд, летчик, истребительная авиация	29
5	Обеспечение фронта	связь, войско, трибунал, управление войском, правительство, обеспечение, деньги, финансы	23
6	Демография	ребенок, социальная структура, население, эвакуация, демография, село, ленинградец, регион, положение, Ленинград, житель, молодежь, смертность	22
7	Сталинградская битва	битва, Сталинградская битва, Сталинград, войско, советское войско, потеря, боевые действия, бой, сражение	21
8	Артиллерия	артиллерия, производство, вооружение, завод, стрелковая дивизия, танк, техника, конструкторское бюро, артиллерийское вооружение	21
9	Военачальники	генерал, оборона, битва, руководство страны, советский военный, офицер, военная миссия, служба, маршал	20
10	Социальная сфера	соотечественник, дети, детдом, детский дом, беспризорность, публикация, процедура, человек, материальные условия	17
11	Германия	войско, Германия, операция, вермахт, Гитлер, немец, Ленинград, восточный коммунист, агентура, пропаганда, противник, коллаборационист	17
12	Регионы России	историческая память, казачество, кинотеатр, миф, исследователь, север, Мурманск, пространство, Юг России, Москва	14
13	Церковь	церковь, православие, духовенство, религия, Русская православная церковь, государство, епархия, советское государство, политика, курс	13
14	Партизаны	партизан, Москва, партизанское движение, Карельский фронт, декабрь, операция, ополчение, народ, боевые действия	12
15	Медицина	медицинский госпиталь, госпиталь, медицина, военная медицина, раненый, эвакогоспиталь, эвакуация, обеспечение	12
16	Лагеря	военнопленный, лагерь, НКВД, побег, МВД, лагерь для военнопленных, охрана, иностранный военнопленный, преступление, содержание, агентура	11

Иерархическое представление тем

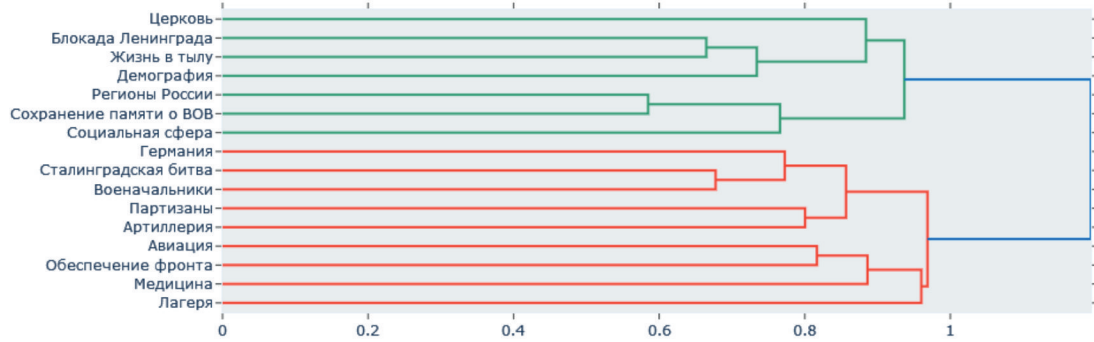


Рис. 4. Иерархия выделенных тем
Fig. 4. The hierarchy of identified topics

Динамика тем



Рис. 5. Динамика частоты тем
Fig. 5. The dynamics of topic frequency

На рис. 5 визуализирована динамика тем в течение десяти рассмотренных лет. Для большей иллюстративности темы сгруппированы по частотности, динамика каждой группы представлена на отдельном графике. Несмотря на то что в разные годы количество статей по конкретным темам менялось, распределение частоты тем по годам в целом выглядит достаточно равномерно. Это показывает, что тематика исследований в разные годы оставалась примерно в одном тематическом диапазоне. Исключением являются даты, связанные с годовщинами значимых событий. Так, например, виден рост интереса к теме «Сталинградская битва» в 2017–2018 гг. и в 2023 г., что, в частности, объясняется тем, что в указанные годы «Вестник ВолГУ. Серия 4: История. Регионоведение. Международные отношения» выпускал специальные выпуски, посвященные годовщине Победы в Сталинградской битве.

Заключение

Результаты исследования демонстрируют возможности BERTopic для анализа исторических научных текстов на русском языке. Тематическое моделирование позволило выделить основные темы, представленные в коллекции текстов, а также ключевые слова, которые их описывают, построить иерархическое и динамическое представление тем. BERTopic обладает рядом преимуществ, включая семантическую обоснованность благодаря использованию больших языковых моделей, а также расширенные возможности для визуализации обнаруженных тем. Представленный в работе подход к формированию коллекции научных текстов, основанный на сочетании использования слов-маркеров, инструментов автоматического сбора текстов и экспертной проверки, может быть использован для сбора научных текстов другой тематики в похожих исследованиях.

Результаты также показывают, что в ведущих российских изданиях по историческим наукам каждый год публикуются десятки исследований по тематике Великой Отечественной войны. С одной стороны, можно отметить, что значительная часть этих работ посвящена темам, выделенным в рамках данного исследования. При этом ряд значимых событий, произошедших в 1941–1945 гг., освещается исследователями не столь широко. Такие события могут стать перспективными направлениями дальнейших исследований по тематике Великой Отечественной войны и содержательно расширить отечественную историографию. С другой стороны, в собранной коллекции текстов присутствует большое количество статей, для которых ни одна из выделенных 16 тем не является преобладающей. Это говорит о значительном тематическом разнообразии проанализированных публикаций.

Список литературы

- Воронцов К. В., Карабулатова И. С., Катцина Т. А., Бородина Н. В., Синицын М. В. Проблемы разработки информационно-аналитической системы по истории социальной помощи в Енисейской губернии (XIX – первая четверть XX в.) // Вопросы истории. 2022. № 12–3. С. 150–161. DOI 10.31166/VoprosyIstorii202212Statyi91
- Галушко И. Н. Применение тематического моделирования для оптимизации процесса поиска релевантных исторических документов (на примере биржевой прессы начала XX в.) // Историческая информатика. 2023. № 2. С. 129–144. DOI 10.7256/2585-7797.2023.2.43466
- Епифанов А. Е., Красноженова Е. Е. Организация и проведение санитарно-эпидемиологических мероприятий в Сталинграде и отдельных районах Сталинградской области (1942–1943 гг.) // Вестник Волгоград. гос. ун-та. Серия 4: История. Регионоведение. Международные отношения. 2023. № 1. С. 162–174. DOI 10.15688/jvolsu4.2023.1.15
- Землянский С. А., Аксёнов С. В., Лызин И. А., Берестнева О. Г. Тематическое моделирование в контексте медицинских текстов // Доклады ТУСУР. 2021. № 4. С. 58–64. DOI 10.21293/1818-0442-2021-24-4-58-64

- Кирина М. А.** Сравнение тематических моделей на основе LDA, STM и NMF для качественного анализа русской художественной прозы малой формы // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2022. № 2. С. 84–101. DOI 10.25205/1818-7935-2022-20-2-93-109
- Колмогорова А. В., Залевская Е. Д.** Компьютерное моделирование как инструмент анализа художественного текста // Филологический класс. 2023. № 2. С. 22–33.
- Колмогорова А. В., Колмогорова П. А., Куликова Е. Р.** О прошлом, но в разное время: компьютерный анализ текстов учебников по истории СССР/России для шести поколений студентов // Вестник Томского гос. ун-та. Филология. 2024. №. 89. С. 73–103. DOI 10.17223/19986645/89/4
- Краснов Ф. В., Диментов А. В., Шварцман М. Е.** Использование тематических моделей для парного сравнения коллекций научных статей // Информатика и ее применения. 2020. Т. 14, № 3. С. 129–135. DOI 10.14357/19922264200318
- Коршунов А., Гомзин А.** Тематическое моделирование текстов на естественном языке // Тр. ИСП РАН. 2012. № 23. С. 215–244.
- Кринко Е. Ф., Медведев М. В.** Демографические последствия Сталинградской битвы // Вестник ВолГУ. Серия 4: История. Регионоведение. Международные отношения. 2018. № 1. С. 91–104. DOI 10.15688/jvolsu4.2018.1.9
- Кузнецов А. В.** Тематическое моделирование в исторической науке // Социосфера. 2022. № 4. С. 9–12.
- Митрофанова О. А., Атугодаге М. М.** Динамическое тематическое моделирование русскоязычного корпуса юридических документов // Terra Linguistica. 2023. № 1. С. 70–87. DOI 10.18721/JHSS.14107
- Нокель М. А., Лукашевич Н. В.** Тематические модели: добавление биграмм и учет сходства между униграммами и биграммami // Вычислительные методы и программирование. 2015. Т. 16. С. 215–234. DOI 10.26089/NumMet.v16r222
- Попова О. В., Гришин Н. В., Погодина М. Я.** Коммуникация молодежи с главами исполнительной власти регионов Российской Федерации во «ВКонтакте» в 2022 году // Полис. Политические исследования. 2023. № 4. С. 122–137. DOI 10.17976/jpps/2023.04.09
- Романов Р. Е.** Социально-трудовые перемещения эваконаселения в западносибирском тылу (1941–1945): ограничения и стимулы // Вестник НГУ. Серия: История, филология. 2016. № 8. С. 146–156.
- Снегирева Л. И.** Резэвакуация детей и детских учреждений из Западно-Сибирского региона (1942–1945 годы) // Вестник НГУ. Серия: История, филология. 2016. № 8. С. 157–169.
- Тверской О. Ю., Канева О. Н.** Применение тематического моделирования для декомпозиции текстовой формулировки образовательной компетенции // Прикладная математика и фундаментальная информатика. 2023. № 1. С. 33–46. DOI 10.25206/2311-4908-2023-10-1-33-43
- Тен Л. В.** Тематическое моделирование в задаче автоматической рубрикации новостных текстов // Terra Linguistica. 2023. № 2. С. 77–91. DOI 10.18721/JHSS.14207
- Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., Hassan, A.** Topic modeling algorithms and applications: A survey // Information Systems. 2023. No. 112. P. 102131. DOI 10.1016/j.is.2022.102131
- Babina O. I.** Topic Modeling for Mining Opinion Aspects from a Customer Feedback Corpus // Automatic Documentation and Mathematical Linguistics. 2024. Vol. 58. No. 1. P. 63–79. DOI 10.3103/S0005105524010060
- Bird S.** NLTK: the natural language toolkit // Proceedings of the COLING/ACL. 2006. P. 69–72.
- Bodrunova S. S., Blekanov I. S., Kukarkin M.** Topic modeling for Twitter discussions: Model selection and quality assessment // 6th SWS International Scientific Conferences on social sciences. 2019. P. 207–214. DOI 10.5593/SWS.ISCSS.2019.5/S16.026

- Chizhik A. V., Sergeyev D. A.** Exploring the Parliamentary Discourse of the Russian Federation Using Topic Modeling Approach // Digital Transformation and Global Society (DTGS 2021). 2021. P. 403–416. DOI 10.1007/978-3-030-93715-7_29
- Devlin J., Chang M.-W., Lee K., Toutanova K.** BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL. 2019. P. 4171–4186. DOI 10.18653/v1/N19-1423
- Glazkova A. V., Kruzhinov V. M., Sokova Z. N.** Dynamic Topic Models for Retrospective Event Detection: A Study on Soviet Opposition-Leaning Media // Lecture Notes in Computer Science. 2019. No. 11832. P. 145–154. DOI 10.1007/978-3-030-37334-4_13
- Grootendorst M.** BERTopic: Neural topic modeling with a class-based TF-IDF procedure // ArXiv preprint arXiv:2203.05794. 2022.
- Korobov M.** Morphological analyzer and generator for Russian and Ukrainian languages // AIST: 4th International Conference, Revised Selected Papers. 2015. P. 320–332. DOI 10.1007/978-3-319-26123-2_31
- Maltseva A., Shilkina N., Evseev E., Matveev M., Makhnytkin O.** Topic modeling of Russian-language texts using the parts-of-speech composition of topics (on the example of volunteer movement semantics in social media) // FRUCT. 2021. P. 247–253. DOI 10.23919/FRUCT52173.2021.9435475
- Owa D. L. M.** Identification of topics from scientific papers through topic modeling // Open Journal of Applied Sciences. 2021. Vol. 10, No. 04. P. 541. DOI 10.4236/ojapps.2021.104038
- Reimers N., Gurevych I.** Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of EMNLP-IJCNLP. 2019. P. 3982–3992. DOI 10.18653/v1/D19-1410
- Robledo S., Zuluaga M.** Topic modeling: Perspectives from a literature review // IEEE Access. 2022. No. 11. P. 4066–4078. DOI 10.1109/ACCESS.2022.3232939
- Sakhovskiy A., Solovyev V., Solnyshkina M.** Topic modeling for assessment of text complexity in Russian textbooks // 2020 Ivannikov Ispras Open Conference (ISPRAS). 2020. P. 102–108. DOI 10.1109/ISPRAS51486.2020.00022
- Syed S., Spruit M.** Full-text or abstract? examining topic coherence scores using latent dirichlet allocation // 2017 IEEE International conference on data science and advanced analytics (DSAA). 2017. P. 165–174. DOI 10.1109/DSAA.2017.61
- Solovyev V., Solnyshkina M., Tutubalina E.** Topic Modeling for Text Structure Assessment: The case of Russian Academic Texts // Journal of Language and Education. 2023. Vol. 9. No. 3 (35). P. 143–158. DOI 10.17323/jle.2023.16604
- Zamiraylova E., Mitrofanova O.** Dynamic topic modeling of Russian prose of the first third of the XXth century by means of non-negative matrix factorization // Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019). 2020. P. 321–339.

References

- Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., Hassan, A.** Topic modeling algorithms and applications: A survey. *Information Systems*, 2023, no. 112, p. 102131. DOI 10.1016/j.is.2022.102131
- Babina O. I.** Topic Modeling for Mining Opinion Aspects from a Customer Feedback Corpus. *Automatic Documentation and Mathematical Linguistics*, 2024, vol. 58, no. 1, pp. 63–79. DOI 10.3103/S0005105524010060
- Bird S.** NLTK: the natural language toolkit. *Proceedings of the COLING/ACL*, 2006, pp. 69–72.
- Bodrunova S. S., Blekanov I. S., Kukarkin M.** Topic modeling for Twitter discussions: Model selection and quality assessment. *6th SWS International Scientific Conferences on social sciences*, 2019, pp. 207–214. DOI 10.5593/SWS.ISCSS.2019.5/S16.026

- Chizhik A. V., Sergeyev D. A.** Exploring the Parliamentary Discourse of the Russian Federation Using Topic Modeling Approach. *Digital Transformation and Global Society (DTGS 2021)*, 2021, pp. 403–416. DOI 10.1007/978-3-030-93715-7_29
- Devlin J., Chang M.-W., Lee K., Toutanova K.** BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL*, 2019, pp. 4171–4186. DOI 10.18653/v1/N19-1423
- Epifanov A. E., Krasnozhenova E. E.** Organization and holding sanitary and epidemiological measures in Stalingrad and separate districts of the Stalingrad region (1942-1943). *Science journal of Volgograd state university. History. Area Studies. International Relations*, 2023, iss. 1, pp. 162–174. (in Russ.) DOI 10.15688/jvolsu4.2023.1.15
- Galushko I. N.** The use of topic modeling to optimize the process of searching for relevant historical documents (on the example of the stock exchange press of the early 20th century). *Historical informatics*, 2023, iss. 2, pp. 129–144. (in Russ.) DOI 10.7256/2585-7797.2023.2.43466
- Glazkova A. V., Kruzhinov V. M., Sokova Z. N.** Dynamic Topic Models for Retrospective Event Detection: A Study on Soviet Opposition-Leaning Media. *Lecture Notes in Computer Science*, 2019, no. 11832, pp. 145–154. DOI 10.1007/978-3-030-37334-4_13
- Grootendorst M.** BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv preprint arXiv:2203.05794*, 2022.
- Kirina M. A.** A comparison of topic models based on LDA, STM and NMF for qualitative studies of Russian short prose. *NSU Vestnik. Series: Linguistics and Intercultural Communication*, 2022, iss. 2, pp. 84–101. (in Russ.) DOI 10.25205/1818-7935-2022-20-2-93-109
- Kolmogorova A. V., Kolmogorova P. A., Kulikova E. R.** About the past, but at different times: computer analysis of texts in textbooks on the history of the USSR/Russia for six generations of students. *Tomsk State University Journal of Philology*, 2024, iss. 89, pp. 73–103. (in Russ.) DOI 10.17223/19986645/89/4
- Kolmogorova A. V., Zalevskaya E. D.** Computer-assisted modeling as an instrument for fiction text analysis. *Philological class*, 2023, iss. 2, pp. 22–33. (in Russ.)
- Korobov M.** Morphological analyzer and generator for Russian and Ukrainian languages. *AIST: 4th International Conference, Revised Selected Papers 4*, 2015, pp. 320–332. DOI 10.1007/978-3-319-26123-2_31
- Korshunov A., Gomzin A.** Topic modeling of texts in natural language. *Proceedings of ISP RAS*, 2012, iss. 23, pp. 215–244. (in Russ.)
- Krasnov F. V., Dimentov A. V., Shvartsman M. E.** Using topic models for pairwise comparison of collections of scientific papers. *Informatics and Applications*, 2020, vol. 14, iss. 3, pp. 129–135. (in Russ.) DOI 10.14357/19922264200318
- Krinko E. F., Medvedev M. V.** Demographic consequences of the Stalingrad battle. *Science journal of Volgograd state university. History. Area Studies. International Relations*, 2018, iss. 1, pp. 91–104. (in Russ.) DOI 10.15688/jvolsu4.2018.1.9
- Kuznetsov A. V.** Topic modeling in historical sciences. *Sociosfera [Sociosphere]*, 2022, iss. 4, pp. 9–12. (in Russ.)
- Maltseva A., Shilkina N., Evseev E., Matveev M., Makhnytkin O.** Topic modeling of Russian-language texts using the parts-of-speech composition of topics (on the example of volunteer movement semantics in social media). *FRUCT*, 2021, pp. 247–253. DOI 10.23919/FRUCT52173.2021.9435475
- Mitrofanova O. A., Athogodage M. M.** Dynamic topic modelling of the Russian legal text corpus. *Terra Linguistica*. 2023, iss. 1, pp. 70–87. (in Russ.) DOI 10.18721/JHSS.14107
- Nokel M. A., Loukachevitch N. V.** Topic Models: Adding Bigrams and Taking Account of the Similarity between Unigrams and Bigrams. *Numerical Methods and Programming*, 2015, vol. 16, pp. 215–234. (in Russ.) DOI 10.26089/NumMet.v16r222

- Owa D. L. M.** Identification of topics from scientific papers through topic modeling. *Open Journal of Applied Sciences*, 2021, vol. 10, no. 04, p. 541. DOI 10.4236/ojapps.2021.104038
- Popova O. V., Grishin N. V., Pogodina M. Ya.** Communication of youth with the highest officials of the subjects of the Russian Federation on the social network “VKontakte” in 2022. *Polis. Political Studies*, 2023, iss. 4, pp. 122–137. (in Russ.) DOI 10.17976/jpps/2023.04.09
- Reimers N., Gurevych I.** Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of EMNLP-IJCNLP*, 2019, pp. 3982–3992. DOI 10.18653/v1/D19-1410
- Robledo S., Zuluaga M.** Topic modeling: Perspectives from a literature review. *IEEE Access*, 2022, no. 11, pp. 4066–4078. DOI 10.1109/ACCESS.2022.3232939
- Romanov R. E. Social and labor relocations of evacuated people in west Siberian rear (1941-1945): constraints and incentives. *Vestnik NSU. Series: History and Philology*, 2016, iss. 8, pp. 146–156. (in Russ.)
- Sakhovskiy A., Solovyev V., Solnyshkina M.** Topic modeling for assessment of text complexity in Russian textbooks. *2020 Ivannikov Ispras Open Conference (ISPRAS)*, 2020, pp. 102–108. DOI 10.1109/ISPRAS51486.2020.00022
- Solovyev V., Solnyshkina M., Tutubalina E.** Topic Modeling for Text Structure Assessment: The case of Russian Academic Texts. *Journal of Language and Education*, 2023, vol. 9, no. 3 (35), pp. 143–158. DOI 10.17323/jle.2023.16604
- Snegireva L. I.** Re-evacuation of children and children institutions form West-Siberian region (1942-1945). *Vestnik NSU. Series: History and Philology*, 2016, iss. 8, pp. 157–169. (in Russ.)
- Syed S., Spruit M.** Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. *2017 IEEE International conference on data science and advanced analytics (DSAA)*, 2017, pp. 165–174. DOI 10.1109/DSAA.2017.61
- Ten L. V.** Topic modeling in automatic categorization of news texts. *Terra Linguistica*, 2023, iss. 2, pp. 77–91. (in Russ.) DOI 10.18721/JHSS.14207
- Tverskoi O. I., Kaneva O. N.** Application of topic modeling for the decomposition of the textual formulation of educational competence. *Applied mathematics and fundamental computer science*, 2023, iss. 1, pp. 33–46. (in Russ.) DOI 10.25206/2311-4908-2023-10-1-33-43
- Vorontsov K. V., Karabulatova I. S., Kattsina T. A., Borodina N. V., Sinitsyn M. V.** Problems of developing an analytical information system on the history of social assistance in the Yenisei province (19th - first quarter of the 20th century). *Issues of history*, 2022, iss. 12-3, pp. 150–161. (in Russ.) DOI 10.31166/VoprosyIstorii202212Statyi91
- Zamiraylova E., Mitrofanova O.** Dynamic topic modeling of Russian prose of the first third of the XXth century by means of non-negative matrix factorization. *Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019)*, 2020, pp. 321–339.
- Zemlyansky S. A., Axyonov S. V., Lyzin I. A., Berestneva O. G.** Topic modeling in the context of medical texts. *Proceedings of TUSUR University*, 2021, iss. 4, pp. 58–64. (in Russ.) DOI 10.21293/1818-0442-2021-24-4-58-64

Информация об авторах

Сокова Зинаида Николаевна, доктор исторических наук, профессор

Кружинов Валерий Михайлович, доктор исторических наук, профессор

Глазкова Анна Валерьевна, кандидат технических наук, доцент

Information about the Authors

Zinaida N. Sokova, Doctor of Sciences (History), Professor

Valery M. Kruzhinov, Doctor of Sciences (History), Professor

Anna V. Glazkova, Candidate of Sciences (Technology), Associate Professor

*Статья поступила в редакцию 16.09.2024;
одобрена после рецензирования 17.01.2025; принята к публикации 17.01.2025*

*The article was submitted 16.09.2024;
approved after reviewing 17.01.2025; accepted for publication 17.01.2025*