Terra Linguistica. 2023, Том 14, № 2. С. 77–91.
Terra Linguistica. 2023, Vol. 14, No. 2. Pp. 77–91.

Научная статья
УДК 81
DOI: https://doi.org/10.18721/JHSS.14207

ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ В ЗАДАЧЕ АВТОМАТИЧЕСКОЙ РУБРИКАЦИИ НОВОСТНЫХ ТЕКСТОВ

Л.В. Тен ✉

Санкт-Петербургский государственный университет,
Санкт-Петербург, Российская Федерация

✉ lia.ten136@gmail.com

Аннотация. Тематическое моделирование широко используется в рамках интеллектуального анализа текстов на естественном языке, в ходе которого посредством статического анализа текстов обнаруживается скрытая тематическая структура. В данной статье предлагается подход к автоматической рубрикации новостных статей с использованием методов тематического моделирования в сочетании с процедурой автоматического назначения меток тем. Тематическое моделирование осуществляется при помощи ряда алгоритмов на языке программирования Python, включая латентное размещение Дирихле (latent Dirichlet allocation, LDA), неотрицательное матричное разложение (non-negative matrix factorization, NMF) и генеративную модель битермов (biterm topic model, BTM). Для автоматического назначения меток тем применяется подход с использованием языковой модели ChatGPT. Оценка кандидатов в метки основана на результатах опроса респондентов. Проведенные эксперименты показывают, что предложенный алгоритм может служить эффективным средством в задаче автоматической рубрикации текстов. Полученные результаты представляют интерес для специалистов в области прикладной и компьютерной лингвистики, медиакоммуникаций и научной журналистики.

Ключевые слова: автоматическая рубрикация, тематическое моделирование, автоматическое назначение меток тем, корпус новостных статей, ChatGPT.

Для цитирования: Тен Л.В. Тематическое моделирование в задаче автоматической рубрикации новостных текстов // Terra Linguistica. 2023. Т. 14. № 2. С. 77–91. DOI: 10.18721/JHSS.14207

Research article
DOI: https://doi.org/10.18721/JHSS.14207

TOPIC MODELING IN AUTOMATIC CATEGORIZATION OF NEWS TEXTS

L.V. Ten ✉

St. Petersburg State University,
St. Petersburg, Russian Federation

✉ lia.ten136@gmail.com

Abstract. Topic modeling is a text mining method used for discovering underlying semantic structure in large collections of documents. In this paper, we propose a novel approach to automatic text categorization of news texts based on topic modeling techniques in combination with automatic topic label assignment. Topic modeling is performed by means of a series of algorithms including latent Dirichlet allocation (LDA), non-negative matrix factorization (NMF), and biterm topic modeling (BTM). In addition, we adopt an approach using the ChatGPT language model in order to assign topic labels. Candidate labels are evaluated by means of human assessments. The experiments carried out within our project demonstrate that the proposed algorithm can serve as an effective tool in the task of automatic text categorization. The results obtained may be of interest to experts in the field of applied and computational linguistics, media communications, and science journalism.

Keywords: text categorization, topic modeling, topic label assignment, news texts, ChatGPT.

Citation: L.V. Ten, Topic modeling in automatic categorization of news texts, Terra Linguistica, 14 (2) (2023) 77–91. DOI: 10.18721/JHSS.14207

Введение

Стремительное увеличение объемов и разнообразия текстовой информации в средствах массовой коммуникации приводит к необходимости в качественной и своевременной категоризации и систематизации текстов, в частности, новостных сообщений в электронных СМИ. Это задача автоматической рубрикации, которая является одним из наиболее распространенных способов систематизации неупорядоченной коллекции текстовых документов. Следует отметить, что в современной лингвистике нет общепринятого определения автоматической рубрикации; в нашей работе мы принимаем следующую формулировку: «рубрикация информации – отнесение порции информации к одной или нескольким категориям из ограниченного множества» [1]. В таком понимании автоматическая рубрикация сближается с задачей классификации, при которой рубрики-классы заданы заранее; тем не менее наличие рубрик предусмотрено не всегда, особенно в случае с новыми изданиями. Кроме того, с появлением новых событий и явлений может возникнуть потребность в новых рубриках. По этой причине однозначно трактовать рубрикацию как задачу классификации не представляется целесообразным.

Большинство методов рубрикации основаны на предположении, что тексты коллекции, относящиеся к одной и той же рубрике, содержат некоторые общие признаки — ключевые слова или словосочетания — наличие или отсутствие которых позволяет говорить о принадлежности или непринадлежности текста к данной рубрике [2]. В работе [1] рассматривается следующая классификация подходов к автоматической рубрикации:

1) методы, основанные на знаниях («инженерный» подход), при которых правила отнесения текста к рубрике строятся экспертами в данной предметной области;

2) методы на основе машинного обучения с учителем, при которых для обучения моделей используются предварительно отрубрицированные коллекции документов.

С реализацией первого подхода связаны определенные сложности, возникающие в силу трудоемкости и непоследовательности ручного рубрицирования, в результате которого разные эксперты могут назначать разные рубрики в зависимости от ряда объективных и субъективных факторов, например, незнания какой-либо области, выходящей за рамки компетенции эксперта, сложности ориентирования в большом классификаторе и т. п. По этой причине ручная рубрикация не снискала большой популярности. В рамках второго подхода были предприняты различные способы автоматической рубрикации: построение вектора рубрик в пространстве слов с использованием электронного тезауруса WordNet и последующего уточнения результатов при помощи метода k ближайших соседей [3]; применение алгоритмов иерархической кластеризации, при которой извлекаются не известные заранее рубрики [4]; использование многоуровневой таксономии в сочетании с мерой релевантности «строка-текст» [5]; вычисление условных вероятностей принадлежности документа рубрике на основе алгоритма PrTFIDF (англ. probabilistic TF-IDF, вероятностный вариант статистической меры TF-IDF) [6] и др. Кроме того, был предложен подход, учитывающий лингвистические особенности текстов, при котором за единицу текста принимается не словоформа (или лемма), а некоторый элемент семантического представления текста, получаемый при помощи комплексного лингвистического анализа [7]. К сожалению, практическая реализация данного алгоритма осталась нерешенной задачей.

В печатной журналистике рубрика определяется как некоторый содержательно-тематический и композиционный раздел издания [8]. В рамках онлайн-журналистики рубрикация обладает собственными особенностями в силу многоуровневой, а не линейной, организации чтения, благодаря чему материал может принадлежать сразу нескольким рубрикам. В работе [9] автор выделяет три способа структурирования контента в интернет-СМИ: 1) рубрикация, основанная на логическом, тематическом или жанровом делении (традиционное понимание рубрики, характерное для печатной журналистики); 2) рубрикация, построенная на объединении частей одной истории в хронологическом порядке (так называемая «длящаяся новость»), и 3) рубрикация при помощи тегов, или ключевых слов, отражающих содержание текста. Третий тип рубрикации особенно часто встречается в сайтах нового типа и опирается при классификации материалов не столько на строгие категории, сколько на множество ассоциаций, охватывающих частично пересекающиеся концепты, – эта особенность более точно отражает естественные механизмы категоризации информации, наблюдаемые в мозге человека [10]. Подобная гибкость в систематизации текстов также позволяет при необходимости извлекать новые рубрики. Таким образом, в данной статье мы придерживаемся определения рубрики как набора ключевых слов или словосочетаний.

Поскольку рубрика рассматривается как тематическая категория, представляется логичным использовать тематическое моделирование для первоначального определения тематической структуры текстов. Тематическое моделирование — одно из направлений обработки естественного языка, основанное на классе алгоритмов статистического анализа текстов. Целью тематического моделирования является выявление скрытых ассоциативно-семантических связей между терминами (словами), документами и темами в корпусах текстов. Результатом построения тематических моделей является конечное множество тем, которые, в свою очередь, образуются из конечного набора терминов, содержательно описывающих документы. Таким образом, тематическое моделирование позволяет структурировать и систематизировать информацию из больших неупорядоченных массивов текстовых коллекций.

В настоящее время наиболее популярны методы, основанные на описании распределений терминов (или единиц словаря) и документов внутри текстовой коллекции с помощью вероятностных законов. Одна из первых и наиболее известных вероятностных тематических моделей – модель латентного размещения Дирихле (англ. latent Dirichlet allocation, LDA), предложенная в 2003 г. [11]. Как и все вероятностные модели, LDA осуществляет «мягкую» кластеризацию, при

которой термин может относиться к разным темам в зависимости от своего контекстуального значения. Это позволяет решать проблемы полисемии терминов, что невозможно при обычной кластеризации [12]. Сильным недостатком модели является допущение, что все термины в документе равнозначны и независимы, в результате чего служебные слова обладают той же значимостью, что и тематические термины (полнозначные слова). Также эта модель не учитывает синтаксические связи между словами и их порядок, поскольку использует представление текста в виде «мешка слов» – неупорядоченного набора токенов. Наконец, LDA существенно ограничивается зависимостью от гиперпараметров, которые необходимо указывать заранее, например, число тем.

Наряду с вероятностными моделями выделяют алгебраические, например, неотрицательное матричное разложение (англ. non-negative matrix factorization, NMF). NMF – алгоритм мультивариантного анализа, использующий схему TF-IDF для оценки весов терминов. Для каждого документа выбираются термины с наибольшим весом; они же и составляют тему для данного документа. Благодаря присваиванию весов терминам достигается высокая интерпретируемость тем, что является существенным достоинством модели.

В случае с длинными текстами традиционные тематические модели типа LDA напрямую извлекают из них шаблоны совместной встречаемости слов, поскольку имеют достаточное количество информации о корреляции терминов. Однако, когда речь идет о коротких текстах, например, твитах, возникает проблема разреженности данных, для решения которой была предложена генеративная модель битермов (англ. biterm topic model, BTM), где битермом считается неупорядоченная пара слов, встречающихся вместе в минимально информативном контексте [13]. Данная модель явным образом генерирует примеры совместной встречаемости терминов на уровне целого корпуса, а не отдельного документа. В результате BTM способна не только корректно разрешать проблему недостаточной частотности терминов в коротких текстах, но и извлекать темы более высокого качества в текстах среднего и большого размера (например, в научных статьях) по сравнению с LDA, что подтверждается экспериментами на корпусах новостных текстов.

В связи с тем, что в результате тематического моделирования автоматически создается список слов-тематизаторов, этому списку можно в дальнейшем присвоить метку – некоторое обобщающее слово или словосочетание, которое бы охватило содержание данной темы. Мы предполагаем, что такие метки тем должны соответствовать потенциальным рубрикам.

Исследовательский корпус русскоязычных новостных текстов

Для проведения процедур тематического моделирования и автоматического назначения меток тем был собран корпус новостных текстов, составляющих один из разделов научно-популярного журнала «Наука и жизнь»[1] и освещающих различные события в сфере науки. Особенностью этого издания является то, что в нем отсутствуют привычные тематические разделы, однако под каждой новостью встречаются теги, по которым можно осуществлять тематический поиск. Итого в корпус вошло 7609 статей, опубликованных с 2006 по 2023 гг. На рис. 1 представлена динамика изменения количества публикаций в указанный период. Размер новостного сообщения был в среднем равен 239 словоупотреблениям. Объем корпуса составил 5,1 млн. словоупотреблений.

Результаты тематического моделирования напрямую зависят от того, каким образом была проведена предварительная обработка текстов [14]. Под последней традиционно понимают такие процедуры, как токенизация, морфологический анализ, в том числе лемматизация, удаление наиболее частотной общеупотребительной лексики (так называемых стоп-слов), удаление редких слов, выделение ключевых выражений, извлечение коллокаций, распознавание именованных сущностей [15]. В результате разрешается проблема «мешка слов», при которой все термины считаются равнозначными независимо от частеречной принадлежности. Также при помощи

[1] https://www.nkj.ru/

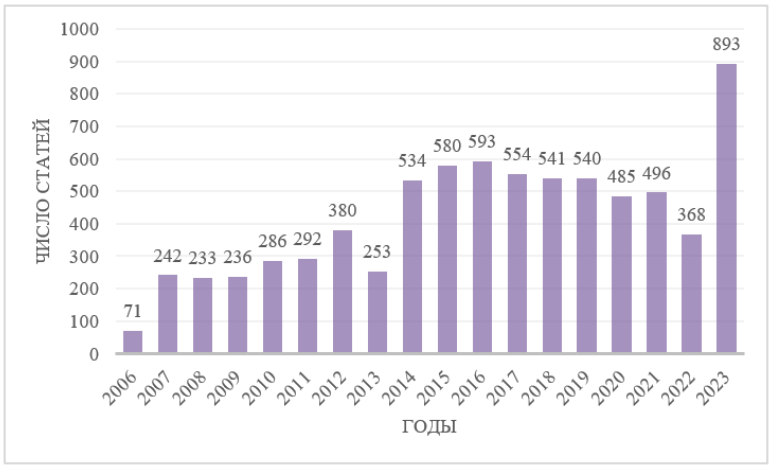

Рис. 1. Распределение текстов исследовательского корпуса по годам

Fig. 1. Distribution of texts in the research corpus by years

предварительной обработки возможен учет синтаксических особенностей текстов, например, порядок слов, а также контекста.

На первом этапе обработки текстов была проведена двухступенчатая процедура токенизации и распознавания именованных сущностей, которые включают имена людей, названия организаций, топонимы и другие имена собственные, а потому часто представляют собой неоднословные выражения, которые необходимо рассматривать как одно целое. Данная процедура была выполнена с помощью библиотеки spaCy для обработки естественного языка[2]: если термины входили в состав какой-либо одной сущности, все они объединялись в один токен при помощи символа нижнего подчеркивания. Всего было получено более 23 тыс. сущностей трех видов: персоны (PER; «академик_фёдоров», «карл_маркс»), организации (ORG; «вестминстерский_аббатство», «оао_газпром», «цукубский_университет») и топонимы (GPE, геополитические сущности (города и государства) и LOC, прочие географические объекты; «ханты-мансийский_округ», «выборгский_район», «босния_и_герцеговина»). Далее документы были очищены от знаков препинания, нетекстовых элементов (иллюстраций, графиков, гиперссылок), цифр, иноязычных слов и наиболее частотной нетематической лексики, в том числе слов служебных частей речи (предлоги, союзы, частицы), числительных, местоимений и междометий. На этапе лемматизации в корпусе сохранялись только существительные и прилагательные — такой подход позволяет существенным образом сократить время обработки данных без ущерба для качества моделей [16]. В ходе завершающей процедуры обработки корпуса были выделены биграммы, или коллокации, представляющие собой статистически устойчивые сочетания из двух слов. В результате было получено 5042 биграммы, многие из которых были образованы именными группами с зависимым прилагательным или существительным в родительном падеже. Объем корпуса сократился до 1,8 млн. лемм, что составило 35% от первоначального размера.

Результаты тематического моделирования

Перед обучением тематических моделей предусматривается возможность удаления из корпуса наиболее частотных и наиболее редких слов в соответствии с абсолютной частотой их встречаемости. Этот шаг оправдан тем, что часто встречающаяся общеупотребительная лексика редко является тематической, равно как и лексика, характерная только для небольшого числа текстов. Для каждой модели были установлены собственные значения: каждый термин должен входить не

[2] https://spacy.io/

менее чем в 50 (LDA, BTM) или 65 (NMF) документов и не более чем в 10–20% текстов от всего объема корпуса в зависимости от модели. Таким образом, словарь уникальных терминов сокращался в среднем на 96%.

Для оценки тематических моделей вводятся понятия интерпретируемости и когерентности. Считается, что тема высоко интерпретируема, если при восприятии составляющих ее слов носитель языка без дополнительных усилий понимает ее содержание и способен дать ей осмысленное название. Для автоматической оценки интерпретируемости используется мера когерентности, которая указывает на степень семантической близости между терминами внутри одной темы: высокие показатели когерентности указывают на то, что термины встречаются неслучайно.

LDA

Как правило, непосредственно перед обучением модели LDA требуется выбрать значения гиперпараметров, или исходных предположений касательно вероятностного распределения, – как показали некоторые недавние публикации [17], эта процедура напрямую влияет на качество работы моделей. К таким гиперпараметрам относятся прежде всего число тем и значения α и β. Параметр α отвечает за плотность тем в документе – чем выше его значение, тем больше тем содержится в одном документе; параметр β представляет собой плотность терминов в теме – при более высоком коэффициенте предполагается, что темы состоят из большего количества терминов. Чтобы определить оптимальные значения данных гиперпараметров, мы перебрали их возможные значения до достижения наивысшего показателя когерентности, при котором показатели α и β составили 0,99. Число тем было выбрано эмпирически и равнялось 13. Выделенные темы приведены в табл. 1.

NMF

Выбор в пользу данной модели объясняется тем, что в ряде случаев она демонстрирует способность выявлять более связные темы с точки зрения как формальной оценки меры когерентности [18], так и экспертной оценки [19], особенно если речь идет о коротких текстах [20]. Кроме того, утверждается, что в то время как LDA выделяет более общие темы, NMF объединяет тексты по более слабо выраженным темам и способна разграничивать темы, схожие ассоциативно или по смыслу [21]. Другое отличие от LDA заключается в том, что NMF, как правило, не требует точной настройки гиперпараметров: единственный параметр, который необходимо указать заранее, — это число тем. Итого модель сгенерировала 15 тем; это число было также выбрано эмпирически. В табл. 2 указан полный перечень тем, полученных при помощи модели NMF.

Таблица 1. Перечень тем, выделенных при помощи модели LDA
Table 1. Topics obtained through LDA

№ темы	Слова темы
1	*объект, звезда, глаз, галактика, экзопланета, планета, масса, солнце, точка, волна, цвет, красный, расстояние, телескоп, гравитационный*
2	*кость, древний, хвост, останки, углеводород, неандерталец, предок, пещера, след, кислород, архея, кожа, смерть, динозавр, микроб*
3	*аппарат, космический, поверхность, атмосфера, комета, станция, луна, планета, спутник, марс, проект, солнечный, орбита, лёд, самолёт*
4	*стресс, сердце, еда, диабет, жир, пища, гормон, депрессия, микрофлора, час, диета, чувство, рецептор, вес, сердечный*
5	*наука, россия, проект, технология, программа, страна, российский, фото, образование, премия, компания, ран, вуз, студент, москва*
6	*город, археолог, находка, раскопка, экспедиция, культура, история, памятник, древний, территория, житель, погребение, римский, период, война*

Окончание таблицы 1

7	*опухоль, иммунный, рак, иммунитет, воспаление, раковый, мутация, кровь, лекарство, больной, печень, заболевание, стволовой, злокачественный, препарат*
8	*вирус, инфекция, вирусный, слон, грипп, бактериальный, вероятность, антибиотик, иммунитет, сон, антитело, геном, признак, белка, штамм*
9	*энергия, частица, комплекс, физика, структура, электрон, модель, химический, основа, элемент, взаимодействие, фотон, теория, излучение, атом*
10	*нейрон, геном, память, нервный, движение, мутация, кора, примат, спинной, последовательность, имплантат, хромосома, мышца, белка, функция*
11	*растение, кошка, собака, дерево, почва, климатический, климат, температура, гриб, территория, экологический, природа, лес, корень, домашний*
12	*ребёнок, поведение, социальный, самка, самец, язык, звук, женщина, взрослый, нейрон, чужой, эмоция, детёныш, материнский, зона*
13	*птица, сон, рыба, быстрый, устройство, воздух, голова, размер, температура, химический, слой, соединение, поверхность, медленный, электрический*

Таблица 2. Перечень тем, выделенных при помощи модели NMF
Table 2. Topics obtained through NMF

№ темы	Слова темы
1	*иммунный, воспаление, иммунитет, воспалительный, лимфоцит, болезнь, кишечник, кожа, стволовой, кишечный, инфекция, бактерия, сигнальный, печень, реакция*
2	*наука, россия, учёный, научный, проект, российский, ран, премия, страна, технология, академик, программа, международный, мир, компания*
3	*мозг, нейрон, спинной, память, кора, имплантат, нервный, движение, сигнал, активность, нейронный, головной, голова, мышца, импульс*
4	*ген, днк, геном, генетический, мутация, белок, примат, хромосома, обезьяна, последовательность, человеческий, активность, вариант, фермент, белка*
5	*опухоль, рак, раковый, мутация, злокачественный, метастаз, первичный, печень, лекарство, вторичный, иммунный, генетический, опухолевый, метод, иммунитет*
6	*археолог, век, древний, кость, находка, город, пещера, раскопка, погребение, могила, захоронение, останки, возраст, памятник, территория*
7	*сон, инфекция, птица, быстрый, мозг, медленный, фаза, вероятность, статистический, плохой, признак, частый, птичий, грипп, час*
8	*растение, бактерия, гриб, дерево, земля, почва, корень, микрофлора, климатический, температура, условие, экологический, семя, территория, вещество*
9	*экзопланета, звезда, планета, галактика, солнце, орбита, космический, солнечный, вода, гравитационный, земля, астроном, объект, телескоп, атмосфера*
10	*стресс, мышь, жирный, удовольствие, нейрон, подкрепление, калория, центр, еда, сладкое, сладкий, психологический, обычный, чувство, лишний*
11	*вирус, кошка, вирусный, коронавирус, фолликул, яйцеклетка, грипп, белок, инфекция, бактерия, клеточный, антитело, уровень, гормон, частица*
12	*слон, хвост, кожа, кость, кожный, ящерица, зверь, мышь, музейный, пластина, конец, хищник, движение, тело, голова*
13	*комплекс, фотон, энергия, свет, фотосинтез, электрон, химический, возбуждение, пигмент, молекула, белок, частица, излучение, центр, перенос*
14	*самка, самец, ребёнок, детёныш, материнский, поведение, потомство, нейрон, чужой, любовь, социальный, сигнал, ядро, агрессия, собственный*
15	*углеводород, архея, микроб, кислород, бактерия, метан, фермент, вода, углекислый_газ, электрон, энергия, нефть, температура, кислота, длинный*

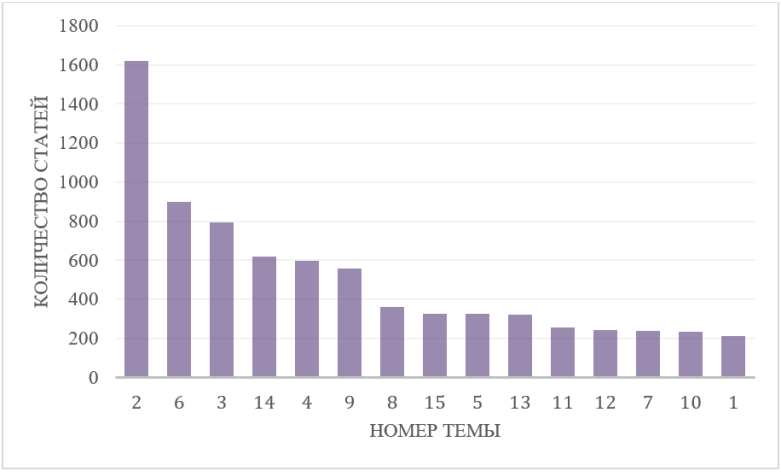

Рис. 2. Распределение текстов исследовательского корпуса по темам согласно модели NMF
Fig. 2. Dictribution of texts in the research corpus by topic according to the NMF model

Согласно модели NMF, четыре самые частотные темы охватывают более 50% текстов корпуса. Распределение статей по темам представлено на рис. 2.

BTM

Модель BTM генерирует шаблоны совместной встречаемости слов (так называемые битермы) во всем корпусе, что делает эту модель наиболее эффективной при анализе коротких текстов, поскольку такие шаблоны в них редки и ненадежны. Вместе с тем, как утверждают авторы данной модели, BTM превосходит LDA с точки зрения когерентности даже в случае с длинными текстами, например, научными статьями. Для проверки этого утверждения мы построили вариант данной модели, реализованный в библиотеке bitermplus[3]. Всего BTM выделила 12 тем, которые представлены в табл. 3.

При знакомстве с выделенными темами можно заметить, что для некоторых из них характерна смешанная лексика, когда одна тема содержит слова, относящиеся по смыслу к двум и более темам; особенно явно это заметно в случае с моделью LDA. Также в результате фильтрации словаря терминов в темы не попали неоднословные выражения, за исключением биграммы «углекислый газ» в модели NMF, что может говорить о недостаточно высокой эффективности алгоритмов автоматического извлечения коллокаций и распознавания именованных сущностей. С другой стороны, моделям удалось избежать многих других проблем, связанных с низким качеством тематических моделей по классификации [22]; например, в нашем случае отсутствовали темы, состоящие исключительно из частотной общеупотребительной лексики либо, напротив, из узкоспециальных терминов, а также темы, образованные случайным набором слов или дублирующие друг друга.

Что касается формальной оценки моделей с точки зрения когерентности, мы использовали метрику C_{UMass}, которая вычисляет, насколько часто два слова w_i и w_j встречаются вместе в корпусе, по формуле:

$$C_{UMass}\left(w_i, w_j\right) = \log \frac{D\left(w_i, w_j\right)+1}{D\left(w_i\right)},$$

где $D(w_i, w_j)$ равняется числу совместной встречаемости w_i и w_j в корпусе, а $D(w_i)$ — числу вхождений w_i без w_j. Как правило, чем больше значение этой меры, тем лучше показатель интерпре-

[3] https://pypi.org/project/bitermplus/

тируемости [23]. Показатель когерентности для модели LDA составил –2,33 на интервале [–14, 14], где оптимальными считаются значения, близкие к нулю. Для BTM когерентность была равна –1,82, что говорит о более высоком качестве модели. В случае с реализацией модели NMF, использованной в данной работе, вычисление меры когерентности не предусмотрено вовсе, что делает суждение исследователя единственным возможным вариантом при оценке данной модели. Отметим только, что поскольку от NMF требовалось выделить пригодные к обобщению осмысленные темы, можно считать, что модель успешно справилась с поставленной задачей.

Таблица 3. Перечень тем, выделенных при помощи модели BTM
Table 3. Topics obtained through BTM

№ темы	Слова темы
1	*звезда, частица, свет, экзопланета, галактика, излучение, объект, физика, волна, структура, энергия, процесс, масса, солнце, свойство*
2	*опухоль, рак, раковый, мутация, иммунный, лекарство, иммунитет, злокачественный, генетический, днк, бактерия, вирус, молекулярный, больной, препарат*
3	*иммунный, воспаление, болезнь, иммунитет, инфекция, стресс, бактерия, кровь, кишечник, жир, воспалительный, реакция, вирус, ткань, кишечный*
4	*ребёнок, язык, социальный, поведение, память, женщина, способность, возраст, участник, связь, звук, вывод, состояние, внимание, собака*
5	*нейрон, сон, сигнал, поведение, быстрый, самка, зона, детёныш, нервный, самец, материнский, птица, область, нейронный, медленный*
6	*днк, геном, генетический, вирус, мутация, человеческий, примат, развитие, хромосома, особенность, белка, последовательность, яйцеклетка, вариант, клеточный*
7	*наука, россия, проект, учёный, технология, программа, российский, страна, развитие, мир, ран, премия, образование, фото, вуз*
8	*земля, аппарат, космический, планета, поверхность, комета, луна, вода, станция, марс, спутник, орбита, атмосфера, солнечный, лёд*
9	*кожа, птица, хвост, слон, кошка, кость, голова, вода, рыба, тело, зверь, хищник, самец, конец, самка*
10	*век, археолог, древний, находка, город, кость, останки, раскопка, ранний, территория, погребение, современный, последний, возраст, экспедиция*
11	*бактерия, растение, энергия, вода, комплекс, молекула, кислород, температура, химический, углеводород, свет, фотон, условие, архея, земля*
12	*нейрон, движение, сигнал, спинной, мышца, имплантат, импульс, нервный, информация, сеть, тело, нога, головной, нейронный, кора*

Автоматическое назначение меток тем

Тематическое моделирование в традиционном понимании не предполагает назначения меток тем в качестве обязательной операции. В общем случае тема представляется в виде номера или списка топ-слов — слов с наибольшими весами или вероятностями. Для облегчения интерпретации тем могут использоваться метки (англ. label), представляющие собой некоторое слово или словосочетание, охватывающее общее содержание темы как набора ключевых слов. Как правило, метки присваиваются самими исследователями на основании субъективных факторов. Тем не менее, возможно автоматическое назначение меток, что значительно упрощает интерпретацию тем, а также экономит время и усилия экспертов.

Автоматическое назначение тем опирается либо на внутренние источники, в роли которого выступают исследовательские корпуса, или внешние источники, включающие справочные корпуса, базы данных и онтологии, результаты поисковых систем и пр. В первом случае извле-

чение меток может осуществляться несколькими способами, например, при помощи мультимодальных моделей на основе нейронных сетей [24], алгоритмов суммаризации [25], модели глубокого обучения seq2seq [26] и др. Существует и большое разнообразие среди подходов к назначению меток с использованием внешних источников. Например, в работе [27] кандидаты в метки генерировались на основе англоязычной Википедии с последующим ранжированием при помощи косинусного сходства. В статье [28] была предложена модель LDA, дополненная понятиями из формальной онтологии; для поиска наиболее значимых меток использовался семантический граф. Модель ранжирования при помощи графов была также использована в других работах [29]. В статье [30] авторы предложили два подхода к назначению меток тем в корпусах русскоязычной малой прозы, а именно извлечение меток-кандидатов из поисковой системы Яндекс и извлечение меток-кандидатов из Википедии при помощи процедуры эксплицитного семантического анализа (англ. explicit semantic analysis). Процедура оценки показала, что первый алгоритм в большинстве случаев предсказывает правильные метки, но не всегда корректно связывает их с ключевыми словами в теме, в то время как второй алгоритм выделяет метки с широким содержанием. Таким образом, комбинация этих методов для назначения меток тем представляется оптимальным решением.

Для семантической компрессии выделенных тем мы использовали ChatGPT — чат-бота с искусственным интеллектом, разработанного компанией OpenAI[4], который имеет хороший потенциал в качестве инструмента автоматического присвоения меток (англ. labeling) [31]. ChatGPT представляет собой нейронную сеть типа трансформер, способную генерировать и суммаризировать тексты, отвечать на вопросы, относящиеся к широкому кругу предметных областей, производить семантический и тематический поиск, решать математические задачи, создавать программный код и мн. др. В настоящее время для пользователей из России доступны различные версии на основе ChatGPT, адаптированные под русский язык в виде чат-ботов, которые распространятся преимущественно через мессенджер Telegram. Все они базируются на другой языковой модели от OpenAI — GPT-3.5, которая была обучена на наборах данных, собранных до сентября 2021 г., а поэтому модель не способна работать с информацией, относящейся к более поздним датам. К другим ограничениям модели можно отнести возможность выдавать правдоподобные, но бессмысленные или неправильные ответы, а также чувствительность к формулировкам запросов, от которых зависит качество результатов. Поскольку ChatGPT работает в диалоговом режиме, получение необходимой информации происходит посредством запросов на естественном языке — в нашем случае русском. А именно, боту было поручено подобрать одно обобщающее слово, которое охватило бы значение данной темы; при этом формулировка запроса была постоянной и включала в себя все слова темы, разделенные пробелом. Примеры запросов и ответов модели приведены на рис. 3.

Несмотря на единообразие запросов, в некоторых случаях бот выдавал словосочетание вместо слова, например, «научные исследования» или «квантовая физика». Итого боту было предложено 40 тем, полученные в ходе тематического моделирования на основе моделей LDA, BTM и NMF. Некоторые из примеров представлены в табл. 4. Как можно видеть, потенциальной меткой часто становилось название предметной области, к которой относится статья.

Из 40 назначенных меток только 15 оказались уникальными, что свидетельствует о схожести тем в различных моделях. Например, метки «онкология», «иммунология» и «археология» встречались по три раза, поскольку соответствующие им темы были выделены в каждой из моделей. При этом наблюдались метки, синонимичные друг другу, например, «нейробиология» и «нейронаука», «астрономия» и «астрофизика», «космология» и «космические исследования». Все это говорит, с одной стороны, о схожести тематических моделей, выдающих сопоставимые результаты, а с другой, о достаточно последовательном выборе меток со стороны самого бота, который, как правило, отдавал предпочтение общенаучным терминам при обобщении тем.

[4] https://openai.com/blog/chatgpt/

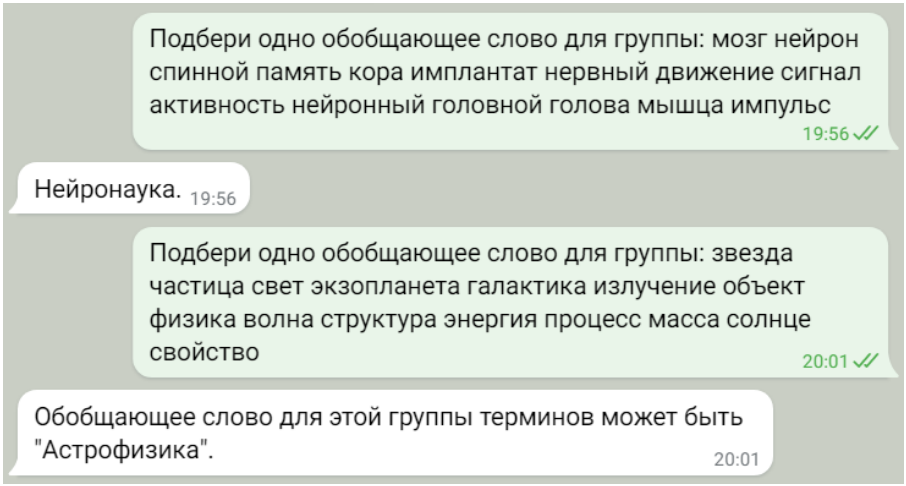

Рис. 3. Примеры запросов к модели ChatGPT и ее ответов

Fig. 3. Examples of prompts given to the ChatGPT model and its responses

Таблица 4. Примеры меток тем, полученных при помощи ChatGPT
Table 4. Examples of topic labels obtained through ChatGPT

Тема	Метка ChatGPT
мозг, нейрон, спинной, память, кора, имплантат, нервный, движение, сигнал, активность, нейронный, головной, голова, мышца, импульс	нейронаука
вирус, кошка, вирусный, коронавирус, фолликул, яйцеклетка, грипп, белок, инфекция, бактерия, клеточный, антитело, уровень, гормон, частица	инфекция
век, археолог, древний, находка, город, кость, останки, раскопка, ранний, территория, погребение, современный, последний, возраст, экспедиция	археология
ребёнок, поведение, социальный, самка, самец, язык, звук, женщина, взрослый, нейрон, чужой, эмоция, детёныш, материнский, зона	психология
комплекс, фотон, энергия, свет, фотосинтез, электрон, химический, возбуждение, пигмент, молекула, белок, частица, излучение, центр, перенос	фотохимия
наука, россия, проект, учёный, технология, программа, российский, страна, развитие, мир, ран, премия, образование, фото, вуз	научные исследования
опухоль, иммунный, рак, иммунитет, воспаление, раковый, мутация, кровь, лекарство, больной, печень, заболевание, стволовой, злокачественный, препарат	онкология
экзопланета, звезда, планета, галактика, солнце, орбита, космический, солнечный, вода, гравитационный, земля, астроном, объект, телескоп, атмосфера	астрономия

Для проверки результатов мы использовали процедуру оценки, основанную на [25], попросив 6 человек, трое из которых имеют филологическое образование, оценить шесть случайно выбранных меток в каждой модели по шкале от 0 до 2, где 0 означает, что метка не охватывает содержание темы, 1 указывает что метка частично покрывает содержание темы, а 2 указывает, что метка полностью описывает содержимое темы. Для каждой группы меток рассчитывалось среднее; при этом метки со средней оценкой $\geq 1{,}5$ считались хорошими, а $\leq 0{,}5$ – плохими. Нами был получен результат 1,6 что указывает на высокое качество меток.

Заключение

В данной статье была предпринята попытка автоматической рубрикации новостных статей на основе комбинированного подхода с использованием тематического моделирования и автоматического назначения меток тем при помощи языковой модели ChatGPT. Были построены три

тематические модели, включая латентное размещение Дирихле (LDA), неотрицательное матричное разложение (NMF) и генеративную модель битермов (BTM). Семантическая компрессия тем осуществлялась при помощи языковой модели ChatGPT, которая назначала каждой теме некоторое общее название. В результате экспериментов была оценена эффективность данного алгоритма в качестве инструмента автоматической рубрикации текстов. Согласно оценкам респондентов, предложенный подход может быть использован для автоматической генерации рубрик и структурирования больших массивов текстовых данных.

СПИСОК ИСТОЧНИКОВ

1. **Агеев М.С., Добров Б.В., Лукашевич Н.В.** Автоматическая рубрикация текстов: методы и проблемы // Ученые записки Казанского государственного университета. 2008. Т. 150, № 4. С. 25–40.

2. **Агеев В.Н.** Задача классификации и рубрикации текстов // Вестник МГУП. 2011. № 1. С. 15–22.

3. **Гуляев О.В., Лукашевич Н.В.** Автоматическая классификация текстов на основе заголовка рубрики // Новые информационные технологии в автоматизированных системах. 2013. № 16. С. 238–244.

4. **Сорокин Д.И., Нужный А.С., Савельева Е.А.** Иерархическая рубрикация текстовых документов // Труды ИСП РАН. 2020. № 6. С. 127–136. DOI: 10.15514/ISPRAS–2020–32(6)–10

5. **Черняк Е.Л., Миркин Б.Г.** Использование мер релевантности строка-текст для автоматизации рубрикации научных статей // Бизнес-информатика. 2014. № 2 (28). С. 51–62.

6. **Дунаев Е.В., Шелестов А.А.** Автоматическая рубрикация web-страниц в интернет-каталоге с иерархической структурой // Интернет-математика 2005: автоматическая обработка веб-данных. 2005. С. 382–398.

7. **Добров А.В.** Комплексный лингвистический подход к автоматической рубрикации новостных сообщений // Политическая лингвистика. 2011. № 3. С. 202–209.

8. **Волокитин Д.Ю.** Рубрики и тематические разделы как пространство содержательного развития развернутой журналистской истории // Знак: проблемное поле медиаобразования. 2019. № 2 (32). С. 149–157.

9. **Устюжанина Д.А.** Интернет-журналистика: учебное пособие. Красноярск: СФУ, 2019. 120 с.

10. **O'Reilly T.** What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software // Communications & Strategies. 2007. No. 1. Pp. 17–37.

11. **Blei D.M., Ng A.Y., Jordan M.I.** Latent Dirichlet Allocation // The Journal of Machine Learning Research. 2003. Vol. 3. Pp. 993–1022.

12. **Steyvers M., Griffiths T.** Probabilistic Topic Models // Handbook of Latent Semantic Analysis. 2007. Pp. 439–460.

13. **Yan X., Guo J., Lan Y., Cheng X.** A Biterm Topic Model for Short Texts // Proceedings of the 22nd International Conference on World Wide Web. 2013. Pp. 1445–1456. DOI: 10.1145/2488388.2488514

14. **Denny M.J., Spirling A.** Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It // Political Analysis. 2018. Vol. 26, No. 2. Pp. 168–189. DOI: 10.1017/pan.2017.44

15. **Воронцов К.В.** Вероятностное тематическое моделирование: теория, модели и проект BigARTM. М., 2020. 95 с.

16. **Martin F., Johnson M.** More Efficient Topic Modelling through a Noun Only Approach // Proceedings of the Australasian Language Technology Association Workshop. 2015. Pp. 111–115.

17. **Agrawal A., Fu W., Menzies T.** What Is Wrong with Topic Modeling? (and How to Fix it Using Search-based SE) // Information and Software Technology. 2018. Vol. 98. Pp. 74–88. DOI: 10.1016/j.infsof.2018.02.005

18. **O'Callaghan D., Greene D., Carthy J., Cunningham P.** An Analysis of the Coherence of Descriptors in Topic Modeling // Expert Systems with Applications. 2015. Vol. 42, No. 13. Pp. 5645–5657. DOI: 10.1016/j.eswa.2015.02.055

19. **Egger R., Yu J.** A Topic Modeling Comparison between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts // Frontiers in Sociology. 2022. Vol. 7. Pp. 1–18. DOI: 10.3389/fsoc.2022.886498

20. **Chen Y., Hui Z., Rui L., Zhiwen Y., Jianying L.** Experimental Explorations on Short Text Topic Mining between LDA and NMF Based Schemes // Knowledge-Based Systems. 2019. Vol. 163. Pp. 1–13. DOI: 10.1016/j.knosys.2018.08.011

21. **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. Т. 20, № 2. С. 93–109. DOI: 10.25205/1818-7935-2022-20-2-93-109

22. **Boyd-Graber J., Mimno D., Newman D.** Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements // Handbook of Mixed Membership Models and Their Applications. 2014.

23. **Mimno D., Wallach H., Talley E., Leenders M., McCallum A.** Optimizing Semantic Coherence in Topic Models // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011. Pp. 262–272.

24. **Sorodoc I., Lau J. H., Aletras N., Baldwin T.** Multimodal Topic Labelling // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017. Vol. 2. Pp. 701–706.

25. **Amparo C. B., Xu R.** Automatic Labelling of Topic Models Learned from Twitter by Summarisation // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014. Pp. 618–624. DOI: 10.3115/v1/P14-2101

26. **Alokaili A., Aletras N., Stevenson M.** Automatic Generation of Topic Labels // Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020. Pp. 1965–1968. DOI: 10.1145/3397271.3401185

27. **Bhatia S., Lau J. H., Baldwin T.** Automatic Labelling of Topics with Neural Embeddings // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016. Pp. 953–963.

28. **Allahyari M. Pouriyeh S. Kochut K., Arabnia H.** A Knowledge-based Topic Modeling Approach for Automatic Topic Labeling // International Journal of Advanced Computer Science and Applications. 2017. Vol. 8. Pp. 335–349. DOI: 10.14569/IJACSA.2017.080947

29. **He D., Ren Y., Khattak M., Liu X., Tao S., Gao W.** Automatic Topic Labeling using Graph-based Pre-trained Neural Embedding // Neurocomputing. 2021. Vol. 463. Pp. 131593–131608. DOI: 10.1016/j.neucom.2021.08.078. DOI: 10.1016/j.neucom.2021.08.078

30. **Mitrofanova O., Kriukova A., Shulginov V., Shulginov V.** E-hypertext Media Topic Model with Automatic Label Assignment // Recent Trends in Analysis of Images, Social Networks and Texts. 2021. Vol. 1357. Pp. 102–114. DOI: 10.1007/978-3-030-71214-3_9

31. **Zhu Y., Zhang P., Haq EU., Hui P., Tyson G.** Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks. 2023.

REFERENCES

[1] **M.S. Ageyev, B.V. Dobrov, N.V. Lukashevich,** Automatic Text Categorization: Methods and Problems, Kazan. Gos. Univ. Uchen. Zap. Ser. Fiz.-Mat. Nauki. 150 (4) (2008) 25–40.

[2] **V.N. Ageyev,** Zadacha klassifikatsii i rubrikatsii tekstov [The task of classifying and categorizing texts], Vestnik MGUP. 1 (2011) 15–22.

[3] **O.V. Gulyayev, N.V. Lukashevich,** Avtomaticheskaya klassifikatsiya tekstov na osnove zagolovka rubriki, Novyye informatsionnyye tekhnologii v avtomatizirovannykh sistemakh. 16 (2013) 238–244.

[4] **D.I. Sorokin, A.S. Nuzhnyy, Ye.A. Savelyeva,** Hierarchical rubrication of text documents, Proceedings of ISP RAS. 6 (2020) 127–136. DOI: 10.15514/ISPRAS–2020–32(6)–10

[5] **Ye.L. Chernyak, B.G. Mirkin,** Using phrase-to-text relevance score to annotate research publications, Business Informatics. 2 (28) (2014) 51–62.

[6] **Ye.V. Dunayev, A.A. Shelestov,** Avtomaticheskaya rubrikatsiya web-stranits v internet-kataloge s iyerarkhicheskoy strukturoy [Automatic categorization of web pages in an Internet catalog with a hierarchical structure], Internet-matematika 2005: avtomaticheskaya obrabotka veb-dannykh [Internet Mathematics 2005: automatic processing of web data]. (2005) 382–398.

[7] **A.V. Dobrov,** Kompleksnyy lingvisticheskiy podkhod k avtomaticheskoy rubrikatsii novostnykh soobshcheniy [A comprehensive linguistic approach to automatic categorization of news reports], Political Linguistics. 3 (2011) 202–209.

[8] **D.Yu. Volokitin,** Rubriki i tematicheskiye razdely kak prostranstvo soderzhatelnogo razvitiya razvernutoy zhurnalistskoy istorii [Headings and thematic sections as a space for the meaningful development of a detailed journalistic history], Znak: problemnoye pole mediaobrazovaniya [Sign: a problematic field of media education]. (32) (2019) 149–157.

[9] **D.A. Ustyuzhanina,** Internet-zhurnalistika: uchebnoye posobiye [Online journalism: a textbook]. Krasnoyarsk: SFU, 2019. 120 c.

[10] **T. O'Reilly,** What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software // Communications & Strategies. 1 (2007) 17–37.

[11] **D.M. Blei, A.Y. Ng, M.I. Jordan,** Latent Dirichlet Allocation, The Journal of Machine Learning Research. 3 (2003) 993–1022.

[12] **M. Steyvers, T. Griffiths,** Probabilistic Topic Models, Handbook of Latent Semantic Analysis. 2007. Pp. 439–460.

[13] **X. Yan, J. Guo, Y. Lan, X. Cheng,** A Biterm Topic Model for Short Texts, Proceedings of the 22nd International Conference on World Wide Web. 2013. Pp. 1445–1456. DOI: 10.1145/2488388.2488514

[14] **M.J. Denny, A. Spirling,** Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It, Political Analysis. 2018. Vol. 26 (2) (2018) 168–189. DOI: 10.1017/pan.2017.44

[15] **K.V. Vorontsov,** Veroyatnostnoye tematicheskoye modelirovaniye: teoriya, modeli i proyekt BigARTM [Probabilistic thematic modeling: theory, models and the BigARTM project]. M., 2020.

[16] **F. Martin, M. Johnson,** More Efficient Topic Modelling through a Noun Only Approach, Proceedings of the Australasian Language Technology Association Workshop. 2015. Pp. 111–115.

[17] **A. Agrawal, W. Fu, T. Menzies,** What Is Wrong with Topic Modeling? (and How to Fix it Using Search-based SE), Information and Software Technology. 98 (2018) 74–88. DOI: 10.1016/j.infsof.2018.02.005

[18] **D. O'Callaghan, D. Greene, J. Carthy, P. Cunningham,** An Analysis of the Coherence of Descriptors in Topic Modeling, Expert Systems with Applications. 42 (13) (2015) 5645–5657. DOI: 10.1016/j.eswa.2015.02.055

[19] **R. Egger, J. Yu,** A Topic Modeling Comparison between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts, Frontiers in Sociology. 7 (2022) 1–18. DOI: 10.3389/fsoc.2022.886498

[20] **Y. Chen, Z. Hui, L. Rui, Y. Zhiwen, L. Jianying,** Experimental Explorations on Short Text Topic Mining between LDA and NMF Based Schemes, Knowledge-Based Systems. 163 (2019) 1–13. DOI: 10.1016/j.knosys.2018.08.011

[21] **M.A. Kirina,** 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. 20 (2) (2022) 93–109. DOI: 10.25205/1818-7935-2022-20-2-93-109

[22] **J. Boyd-Graber, D. Mimno, D. Newman,** Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements, Handbook of Mixed Membership Models and Their Applications, 2014.

[23] **D. Mimno, H. Wallach, E. Talley, M. Leenders, A. McCallum,** Optimizing Semantic Coherence in Topic Models, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011. Pp. 262–272.

[24] **I. Sorodoc, J.H. Lau, N. Aletras, T. Baldwin,** Multimodal Topic Labelling, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2 (2017) 701–706.

[25] **C.B. Amparo, R. Xu,** Automatic Labelling of Topic Models Learned from Twitter by Summarisation, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014. Pp. 618–624. DOI: 10.3115/v1/P14-2101

[26] **A. Alokaili, N. Aletras, M. Stevenson,** Automatic Generation of Topic Labels, Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020. Pp. 1965–1968. DOI: 10.1145/3397271.3401185

[27] **S. Bhatia, J.H. Lau, T. Baldwin,** Automatic Labelling of Topics with Neural Embeddings, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016. Pp. 953–963.

[28] **M. Allahyari, S. Pouriyeh, K. Kochut, H. Arabnia,** A Knowledge-based Topic Modeling Approach for Automatic Topic Labeling, International Journal of Advanced Computer Science and Applications. 8 (2017) 335–349. DOI: 10.14569/IJACSA.2017.080947

[29] **D. He, Y. Ren, M. Khattak, X. Liu, S. Tao, W. Gao,** Automatic Topic Labeling using Graph-based Pre-trained Neural Embedding, Neurocomputing. 463 (2021) 131593–131608. DOI: 10.1016/j.neucom.2021.08.078. DOI: 10.1016/j.neucom.2021.08.078

[30] **O. Mitrofanova, A. Kriukova, V. Shulginov, V. Shulginov,** E-hypertext Media Topic Model with Automatic Label Assignment, Recent Trends in Analysis of Images, Social Networks and Texts. 1357 (2021) 102–114. DOI: 10.1007/978-3-030-71214-3_9

[31] **Y. Zhu, P. Zhang, EU. Haq, P. Hui, G. Tyson,** Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks. 2023.

СВЕДЕНИЯ ОБ АВТОРЕ / INFORMATION ABOUT AUTHOR

Тен Лия Валериевна
Lia V. Ten
E-mail: lia.ten136@gmail.com

Поступила: 18.05.2023; Одобрена: 27.06.2023; Принята: 29.06.2023.
Submitted: 18.05.2023; Approved: 27.06.2023; Accepted: 29.06.2023.