УДК 519.179.2 **DOI: 10.18413/2518-1092-2025-10-4-0-8**

Юрчак В.А.

ИСПОЛЬЗОВАНИЕ ТЕМАТИЧЕСКОЙ КЛАСТЕРИЗАЦИИ В МУЛЬТИМОДАЛЬНЫХ ДАННЫХ ДЛЯ ПОИСКА НЕЯВНЫХ СВЯЗЕЙ И ТРЕНДОВ В РАЗВИТИИ ТЕЗАУРУСОВ

АНО ВО «Российский новый университет (РосНОУ)»
ул. Радио, 22, г. Москва, 105005, Россия

e-mail: rabota_pres14@rambler.ru

Аннотация
С развитием искусственного интеллекта и машинного обучения [7] стало возможно использование иерархических вероятностных моделей в области обработки естественного языка. Вероятностные или «тематические модели» облегчили обнаружение базовых тем, которые формируют содержание корпусов текстов. В частности, тематические модели продемонстрировали полезность при анализе разнообразного контента, выходящего за рамки просто текстовой информации, включая изображения, биологические данные и ответы на опросы. Важным применением тематического моделирования стало выявление тенденций исследований.
Цель. Целью данного исследования является разработка и экспериментальная валидация гибридного метода определения оптимального количества тематических кластеров для автоматического обновления специализированных тезаурусов на основе анализа мультимодальных научных текстов. Метод основывается на нормализованных оценках, таких как перплексия и согласованность, что позволяет оценить качество тем и выявить неявные связи между терминами внутри каждой темы. В рамках исследования рассматривается проблема оптимизации количества тем на стыке предметных областей и фиксировании эволюции тем с выделением тренда по каждому термину внутри каждой темы.
Методы. В исследовании предложен новый подход, интегрирующий алгоритмы LDA и BERTopic с адаптивной функцией оптимизации, учитывающей одновременно метрики перплексии (P) и семантической согласованности (C). Разработана оригинальная математическая модель для выявления неявных связей между терминами через комбинирование вероятностного и контекстного сходства.
Научная новизна исследования. В рамках данного исследования представлена математическая модель выявления неявных семантических связей между терминами, сочетающая вероятность семантического и контекстуального сходства, что позволяет идентифицировать новые связи, отсутствующие в тезаурусах. Кроме того, приводится гибридный подход, сочетающий алгоритм скрытого распределения Дирихле (LDA) и BERTopic (Based on Bertopic python packaged) для определения оптимального количества тематических кластеров в мультимодальных текстах.
Результаты. Результатами исследования, описанными в данной статье, служат создание тематической модели с оптимальным количеством тем на стыке предметных областей. Использование международной базы знаний медицинских публикаций PubMed и реферативно-аналитической базы данных Dimensions AI в качестве базового набора данных позволило проследить эволюцию тем с выделением тренда по каждому термину внутри каждой темы и помогло исследователям из различных отраслей понять взаимосвязи между темами и терминами в содержании мультимодальных текстов.
Ключевые слова: LDA; BERTopic; поиск неявных связей; тренд; семантический граф; PubMed; Dimensions AI; тематическое моделирование; перплексия; согласованность
Для цитирования: Юрчак В.А. Использование тематической кластеризации в мультимодальных данных для поиска неявных связей и трендов в развитии тезаурусов // Научный результат. Информационные технологии. – Т.10, №4, 2025. – С. 88-104. DOI: 10.18413/2518-1092-2025-10-4-0-8

Yurchak V.A.

USING THEMATIC CLUSTERING IN MULTIMODAL DATA TO SEARCH FOR IMPLICIT CONNECTIONS AND TRENDS IN THESAURUS DEVELOPMENT

Autonomous Educational Institution of Higher Education "Russian New University (RosNOU)"
22 Radio St., Moscow, 105005, Russia

e-mail: rabota_pres14@rambler.ru

Abstract
With the development of artificial intelligence and machine learning, it has become possible to use hierarchical probabilistic models in the field of natural language processing. Probabilistic or "thematic models" have made it easier to discover the underlying themes that form the content of text corpora. Thematic models have demonstrated their usefulness in analyzing a variety of content that goes beyond just textual information, including images, biological data, and survey responses. An important application of thematic modeling has been the identification of research trends).
Goal. The purpose of this study is to develop and experimentally validate a hybrid method for determining the optimal number of thematic clusters for automatic updating of specialized thesauri based on the analysis of multimodal scientific texts. The method is based on normalized assessments such as perplexity and consistency, which makes it possible to assess the quality of topics and identify implicit connections between terms within each topic. The study examines the problem of optimizing the number of topics at the junction of one or more subject areas and recording the evolution of topics with highlighting the trend for each term within each topic.
Methods. The study proposes a new approach integrating the LDA and BERTopic algorithms with an adaptive optimization function that simultaneously considers the metrics of perplexity (P) and semantic consistency (C). An original mathematical model has been developed to identify implicit relationships between terms through a combination of probabilistic and contextual similarity.
Scientific novelty of the research. This study presents a mathematical model for identifying implicit semantic links between terms, combining the likelihood of semantic and contextual similarities, which makes it possible to identify new links that are missing in thesauri. In addition, a hybrid approach is presented that combines the latent Dirichlet distribution algorithm (LDA) and BERTopic (Based on Bertopic python packaged) to determine the optimal number of thematic clusters in multimodal texts.
Results. The results of the research described in this article are the creation of a thematic model with an optimal number of topics at the junction of one or more subject areas. Using the PubMed international knowledge base for medical publications and the Dimensions AI abstract and analytical database as a basic dataset, it allowed us to trace the evolution of topics with a trend for each term within each topic and helped researchers from various industries understand the interrelationships between topics and terms in the content of multimodal texts.
Keywords: LDA; BERTopic; search for implicit connections; trend; semantic graph; PubMed; Dimensions AI; thematic modeling; perplexity; consistency
For citation: Yurchak V.A. Using Thematic Clustering in Multimodal Data to Search for Implicit Connections and Trends in Thesaurus Development // Research result. Information technologies. – T.10, №4, 2025. – P. 88-104. DOI: 10.18413/2518-1092-2025-10-4-0-8

ВВЕДЕНИЕ

Тематическое моделирование выступает основным инструментом для анализа больших данных, оптимизируя поиск информации, структурируя контент и обеспечивая глубокое понимание смысловых связей в текстах. Поскольку сложность и объем данных продолжают расти, тематическое моделирование остается незаменимым решением для извлечения ценных сведений из неструктурированных текстов.

С развитием искусственного интеллекта и машинного обучения стало возможно использование иерархических вероятностных моделей в области обработки естественного языка. Вероятностные или «тематические модели» облегчили обнаружение базовых тем, которые формируют содержание корпусов текстов. В частности, тематические модели продемонстрировали

полезность при анализе разнообразного контента, выходящего за рамки просто текстовой информации, включая изображения, биологические данные и ответы на опросы. Важным применением тематического моделирования стало выявление тенденций исследований.

Главной задачей тематического моделирования стала способность выявлять закономерности в использовании слов и устанавливать связи между мультимодальными данными, которые демонстрируют сопоставимые закономерности.

Тематические модели позволяют анализировать термины, связанные с базами знаний и тезаурусами, для выявления скрытых тематических структур в мультимодальных данных. Каждая тема в таких моделях формализуется как распределение вероятностей над словами, отражающее их семантическую связность. Модели выступают как генеративные алгоритмы: они описывают вероятностный процесс, по которому создаются текстовые данные, объединяя разнородные источники (тексты, изображения и др.) в логические тематические кластеры. Посредством взаимодействия терминов с внешними ресурсами модели выявляют паттерны, преобразуя неструктурированные данные в набор тем, где каждая тема задает вероятностные отношения между элементами данных.

Процесс моделирования мультимодальных текстов начинается с первоначального выбора распределения по темам. Затем каждое слово в тексте вероятностно распределяется по теме в соответствии с этим распределением. Наконец, слово выбирается из темы, к которой оно было отнесено.

В области автоматизированного анализа использование тематических моделей предполагает задействование международных компьютерных систем, словарей и баз знаний для решения задач распределения тем по словам в словаре. Изучение распределений приводит к сжатому представлению данных для каждого корпуса текста, фиксируя их основные характеристики. За прошедшие годы было введено несколько различных методов моделирования тем, включая неотрицательную матричную факторизацию (NMF), вероятностный латентный семантический анализ (PLSA), скрытое распределение Дирихле (LDA), иерархическое латентное распределение Дирихле (hLDA), динамические тематические модели (DTM), коррелированную тематическую модель (CTM).

Со временем темы в корпусе мультимодальных текстов могут развиваться. Игнорирование временных аспектов во время моделирования тем может затруднять их обнаружение. Решить эту проблему позволяет моделирование эволюции тем, включающее время и раскрывающее скрытые идеи в корпусе. Моделирование эволюции тем определяет темы, на которые влияют временные закономерности, что позволяет фиксировать их тренд. Семантический тренд по темам – это изменения в значении, использовании или восприятии терминов в словарных базах (тезаурусах) с течением времени. Различные области находят полезность в моделях эволюции тем, например, исследователи, изучающие прогресс тем исследования и определяющие соответствующие мультимодальные тексты.

В отличие от классических подходов к тематическому моделированию, которые трактуют мультимодальные тексты как простые наборы слов, игнорируя контекст и смысловые нюансы, текстовые эмбеддинги позволяют преодолеть это ограничение, сохраняя семантические связи между словами. Примерами текстовых эмбеддингов являются двунаправленные кодирующие представления из трансформеров (BERT) и его вариации). Такие подходы также используются для моделирования тем. Примеры включают Top2Vec, который использует представления слов и текстов для совместного внедрения векторов тем и слов. Кроме того, использующий векторные представления кластеров для представления тем, идентифицируя слова в непосредственной близости от центра кластера. Несмотря на свою привлекательность, традиционные методы тематического моделирования демонстрируют слабую эффективность в случаях, когда кластеры имеют сложную (не сферическую) форму. Для устранения этого недостатка исследователи разрабатывают алгоритмы, которые пересматривают значимость терминов в рамках кластера, учитывая не только их частоту, но и семантическую роль, анализируя их распределение и корректируя веса для повышения точности тематической группировки.

Многочисленные исследования подчеркивают эффективность использования скрытого распределения Дирихле (LDA) в различных коммуникативных контекстах, особенно с короткими текстами. Несмотря на доказательства, подтверждающие эффективность тематического моделирования с короткими текстами, по-прежнему отсутствует эмпирическая поддержка, когда дело доходит до его применения к более длинным текстам.

В то время как некоторые исследования предлагают базовое руководство по сегментации входных текстов на фрагменты по 1000 слов, наблюдается заметное отсутствие систематических исследований, подтверждающих такие утверждения. Действительно, это ограничение тематического моделирования распространяется на различные другие методы анализа текста. В то время как короткие тексты, такие как аннотации, сообщения в социальных сетях, журнальные статьи, газетные статьи имеют тенденцию кратко излагать основные темы, применение интеллектуального анализа текста к более длинным текстам выявляет более низкую плотность контента. В исследовании Сбалкьеро и Эдера рассматриваются связи между длиной текста и оптимальным количеством тем.

Определение оптимального количества тем для обнаружения в корпусе значительно влияет на результат [11]. Параметр «k» в алгоритме имеет особое значение, поскольку он влияет на процесс подгонки модели, тем самым влияя на достоверность результатов. Если количество тем слишком мало, это может привести к широким и неоднородным темам, в то время как большое значение «k» может привести к чрезмерно конкретным темам, и то и другое создает проблемы для интерпретации. Для решения этой проблемы были предложены различные подходы, одним из самых простых является метод Байеса. Этот метод включает в себя расчет логарифмического правдоподобия для всех возможных моделей в указанном диапазоне (например, от 2 до N) с использованием метода Гиббса для выборки данных из апостериорного распределения тем, определяя оптимальное количество тем для модели.

В исследовании предложен метод определения оптимального количества тем на стыке предметных областей. Метод определения оптимального количества тем функционирует на основе нормализованных оценок, включая оценки перплексии и согласованности, обеспечивающие консолидированную оценку качества тем и поиска неявных связей между терминами внутри каждой темы. Для достижения этой цели формулируется проблема оптимизации количества тем на стыке предметных областей и фиксировании эволюции тем с выделением тренда по каждому термину внутри каждой темы.

Научная новизна исследования представлена в следующих положениях:

– Разработке гибридного подхода (LDA + BERTopic) с адаптивной функцией оптимизации, учитывающей метрики перплексии (P) и семантической согласованности (C);

– Разработке оригинальной математической модели для выявления неявных связей между терминами через комбинирование вероятностного и контекстного сходства.

Исследование опирается на тщательно собранный набор данных из международной базы знаний медицинских публикаций PubMed и реферативно-аналитической базы данных Dimensions AI, включающих тысячи статей и рефератов из различных областей знаний. В рамках организации и классификации этого обширного набора данных по последовательным темам, используется алгоритм скрытого распределения Дирихле (LDA) с BERTopic. Целью исследования является предоставление метода математической оптимизации, который облегчит определение оптимального количества тем на стыке предметных областей и позволит проследить эволюцию тем с выделением тренда по каждому термину внутри каждой темы. Данный подход позволит более глубоко понять содержание, представленное в статьях об исследованиях в медицинской и иных областях знаний, предлагая понимание их эволюции с течением времени и раскрывая непреходящие важные темы. Главными задачами, рассматриваемыми в рамках данной статьи, будут:

– современные и традиционные походы к моделированию тем в мультимодальных текстах;

– механизм распределения баллов семантического сходства с учетом увеличения количества тем;

– механизм эволюции тем с выделением тренда по каждому термину внутри каждой темы.

Результатами исследования, описанными в данной статье, служат создание тематической модели с оптимальным количеством тем на стыке предметных областей. Использование международной базы знаний медицинских публикаций PubMed и реферативно-аналитической базы данных Dimensions AI в качестве базового набора данных позволит проследить эволюцию тем с выделением тренда по каждому термину внутри каждой темы и поможет исследователям из различных отраслей понять взаимосвязи между темами и терминами в содержании мультимодальных текстов, а также их эволюцию с течением времени.

Статья организована следующим образом. В разделе 1 содержится информация о процедуре сбора и подготовки данных по исследованию.

В разделе 2 содержится информация о подходах к тематическому моделированию.

В разделе 3 содержится информация о настройке тематических моделей LDA и Bertopic.

В разделе 4 содержится информация о механизме распределения оценочных баллов семантического сходства с учетом увеличения количества тем.

В разделе 5 содержится информация о механизме поиска неявных связей между терминами внутри каждой предметной области.

В разделе 6 содержится информация о результатах исследования.

Выводы по исследованию и их обсуждение приводятся в разделе 7 Заключение.

1 ПРОЦЕДУРА СБОРА И ПОДГОТОВКИ ДАННЫХ ПО ИССЛЕДОВАНИЮ

В рамках исследования извлекаются мультимодальные текстовые данные, соответствующие научным статьям, рефератам об исследованиях в медицинской и иных областях знаний, из международной базы знаний медицинских публикаций PubMed и реферативно-аналитической базы данных Dimensions AI. В целом статьи охватывают период с 1998 по 2024 год и взяты из 7016 журналов. Эта цифра также подчеркивает значимость исследований в медицинской и иных областях знаний благодаря увеличению с каждым годом количества публикаций. После исключения дубликатов и статей на неанглоязычных языках были получены 5169 уникальных статей из Dimensions AI и 2000 из PubMed. Всего статьи, извлеченные из этих баз данных, составили коллекцию из 7169 уникальных статей. Нерелевантные метаданные, такие как имена авторов, язык, тип документа, время цитирования, информация об издателе, аббревиатура журнала, том, выпуск, категории были исключены из статей. Впоследствии было произведено объединение соответствующих метаданных, включая название статьи, ключевые слова, аннотацию, дату публикации и название источника, которые содержат информацию для этого исследования, для дальнейшей предварительной обработки и анализа.

В области обработки естественного языка и анализа текста предварительная обработка играет важную роль в уточнении необработанных мультимодальных текстовых данных для последующих операций анализа и моделирования. Первоначально текст преобразуется в нижний регистр. Затем процесс переходит к удалению знаков препинания, при котором все небуквенно-цифровые символы, такие как запятые, точки и круглые скобки, удаляются, тем самым акцентируя внимание на значимых словах, которые в противном случае могли бы внести шум в анализ. Следующий шаг включает токенизацию, разбивку текста на отдельные токены или слова, что обеспечивает более детальный уровень анализа. Кроме того, был включен анализ биграмм и триграмм, чтобы учитывать частые сочетания слов. Биграммы относятся к парам слов, которые часто встречаются в тексте, тогда как триграммы состоят из трех слов, которые часто встречаются вместе. Впоследствии акцент смещается на удаление стоп-слов. Кроме того, исключаются слова длиной менее 3 букв, поскольку этот шаг помогает устранить многие сокращения, встречающиеся в нашем наборе данных с длиной в 2 буквы, что дополнительно уточняет текст для анализа. На следующем этапе выполняется процедура лемматизации, служащая для преобразования слов к их корневой форме. В совокупности эти этапы предварительной обработки повышают качество и связность текстовых данных, делая их более пригодными для преобразования необработанного текста в управляемый и информативный формат для различных задач обработки естественного языка.

2 ПОДХОДЫ К ТЕМАТИЧЕСКОМУ МОДЕЛИРОВАНИЮ

В основе подходов к тематическому моделированию лежат методы кластеризации мультимодальных текстовых данных. Под кластеризацией мультимодальных текстовых данных понимают процесс группировки информации по их семантической схожести. Любое сходство может быть оценено численно, например, расстояние между терминами из двух слов. Кластеры представляют собой не только набор значений объектов с числовым сходством, но и группу объектов, которые представляют одно и то же понятие друг для друга. Высокая эффективность и точность кластеров данных – две основные и важные цели кластеризации [3].

В данном исследовании рассматриваются две тематические модели анализа: LDA и BERTopic. Количество скрытых тем, обозначенных как K, является свободным параметром в обеих моделях. В LDA параметры β и α управляют тематической структурой: β отвечает за распределение слов по темам, а α — за распределение тем по документам. Две модели отличаются методологией, используемой для определения этих скрытых параметров.

Скрытое распределение Дирихле (LDA) – это обобщающая модель для текстов, в которой пропорции смеси θd рассматриваются как случайные величины. Метод LDA моделирует каждый из текстов как комбинацию скрытых тем, описывая распределение слов по темам и документам [7].

В процессе моделирования мультимодальных текстов используется словарь, определяющийся следующим образом. Создается корпус, который включает в себя все анализируемые мультимодальные тексты, где D обозначает общее количество данных, содержащихся в тексте [7].

Метод LDA работает как вероятностная генеративная модель, предполагающая, что каждый мультимодальный текст содержит сочетание различных тем, при этом каждое слово в тексте относится к одной из этих тем. Модель определяет вероятность появления каждого слова в данной теме и вероятность появления каждой темы в тексте. Метод демонстрирует масштабируемость и способен обрабатывать большие массивы данных. LDA эффективно управляет мультимодальными текстами различной длины благодаря своей вероятностной природе, обеспечивая гибкость структуры текстов. Однако он чувствителен к гиперпараметрам, таким как количество тем и критерии Дирихле, что затрудняет поиск оптимальных значений [7].

С другой стороны, BERTopic использует модели на основе transformer для создания векторных представлений слов, учитывающих контекст в мультимодальных текстах. Он использует алгоритмы кластеризации этих вложений для определения тем, предоставляя более контекстуально насыщенные представления. BERTopic может предложить несколько меньшую возможность прямой интерпретации по сравнению с LDA из-за своей зависимости от сложных моделей на основе трансформеров, это компенсируется контекстуально насыщенными темами и значимыми представлениями. BERTopic может быть ресурсоемким с точки зрения вычислений, особенно для больших наборов данных. Как и LDA, BERTopic зависит от настройки гиперпараметров, особенно тех, что связаны с алгоритмами кластеризации, применяемыми совместно с векторными представлениями. Модель BERTopic эффективно обрабатывает мультимодальные тексты, выявляя и детализируя распределение тем [7].

3 НАСТРОЙКА ТЕМАТИЧЕСКИХ МОДЕЛЕЙ LDA И BERTOPIC

Важной характеристикой тематических моделей, таких как LDA и BERTopic, является их способность самостоятельно определять количество тем во время моделирования или требовать их определенного количества на этапе обучения. Определение оптимального количества тем [11] остается серьезной проблемой, не имеющей однозначного решения. Для решения текущей проблемы были применены механизмы по ограничению количества тем во время обучения модели LDA. Для оценки обобщающих тематических представлений каждой модели используются оценочные баллы, чтобы определить наиболее согласованное представление [7].

Увеличение количества проходов может улучшить качество тем, позволяя модели уточнять задания тем. В то время как размер блока определяет количество мультимодальных текстов,

обрабатываемых в каждом обучающем блоке. Это влияет на скорость и требования к памяти во время обучения. Изменение этого параметра может повлиять на скорость конвергенции, согласованность и сложность. В рамках данного исследования поиск оптимального количества тем начинается с диапазона от 4 до 32 с увеличением на 4. Кроме того, данные каждой модели LDA и BERTopic подвергаются двум процессам кодирования: один из них с использованием представления Bagof-Words (BoW), а другой – с использованием частотности терминов (TF–IDF). Bagof-Words (BoW) представляет текст в виде векторов, фокусируясь на количестве слов и игнорируя порядок слов. TF–IDF учитывает важность слов в тексте в зависимости от их встречаемости во всем корпусе. В результате в данном исследовании было получено в общей сложности 150 примеров тем с разбивкой по корпусам текстов [7].

4 МЕХАНИЗМ РАСПРЕДЕЛЕНИЯ ОЦЕНОЧНЫХ БАЛЛОВ СЕМАНТИЧЕСКОГО СХОДСТВА С УЧЕТОМ УВЕЛИЧЕНИЯ КОЛИЧЕСТВА ТЕМ

Несмотря на распространенное мнение о том, что тематическое моделирование раскрывает значимые и ценные скрытые концепции, проверка этого предположения является сложной задачей из-за отсутствия заранее заданных тем для анализа. Отсутствие четкого набора тем для каждого корпуса требует включения внешних оценок для оценки определенных тем в скрытых структурах данных. В целом, оценка моделей может проводиться с помощью двух подходов: внешней оценки и внутренней оценки. Внешняя оценка требует тестирования моделей на реальных задачах и анализа конечной точности, что является лучшим способом понять, как различные модели влияют на задачу. Однако это может быть медленным и дорогостоящим с точки зрения вычислений. С другой стороны, внутренняя оценка, предполагающая поиск показателя для оценки эффективности модели на основе ее внутренних характеристик, без учета конкретной задачи, для решения которой она будет использоваться [7].

В рамках данного исследования использовались различные оценки, такие как перплексия (мера неопределенности модели) и согласованность, чтобы определить наиболее подходящее количество тем для отлаженных моделей LDA и BERTopic. Для того, чтобы продемонстрировать использование оценок запутанности и согласованности приведем математическую модель оптимизации для определения оптимального количества тем моделей LDA и BERTopic [7].

Оптимизация для определения оптимального количества тем моделей LDA и BERTopic будет нацелена на уменьшение запутанности тем и максимизацию согласованности, что позволит улучшить интерпретируемость и точность моделей. Рассмотрим, возможность построения математической модели для этой оптимизации [7].

Перплексия или запутанность (P) – это мера, показывающая, насколько хорошо модель предсказывает вероятностное распределение слов в документах. Чем ниже значение перплексии, тем лучше модель соответствует данным.

Для метода LDA, перплексия для коллекции документов $D = \{d_1, \dots, d_N\}$ определяется как:

$$P(D) = exp\left(-\frac{1}{\sum_{d\in D} N_d}\sum_{d\in D}\log p(d)\right), \quad (1)$$

где N_d – общее количество слов в документе d, а $p(d)$ – вероятность документа d, определяемая моделью.

Для BERTopic, перплексия рассчитывается на основе распределения эмбеддингов и вероятности тем, но принцип остаётся схожим.

Согласованность (C) – метрика, показывающая, насколько семантически связаны слова внутри темы. Чаще всего используют согласованность на основе метрики PMI [8], измеряющей согласованность по частоте встречаемости слов в документах.

Для темы t, содержащей множество слов $\{\omega_1, \omega_2, \dots, \omega_N\}$, согласованность может быть рассчитана как:

$$C(t) = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\log\frac{P(\omega_i,\omega_j)+\epsilon}{P(\omega_i)+P(\omega_j)}, \quad (2)$$

где $P(\omega_1, \omega_2)$ – вероятность совместного появления слов $P(\omega_i)$ и $P(\omega_j)$ – их индивидуальные вероятности, а ϵ – небольшое положительное число для предотвращения деления на ноль.

Общая согласованность для модели – это среднее значение согласованности всех тем:

$$C = \frac{1}{T}\sum_{t=1}^{T} C(t), \quad (3)$$

где T – общее количество тем.

Цель – минимизировать перплексию и максимизировать согласованность.

Пусть:

– $P(\theta)$ – функция, вычисляющая перплексию модели с параметрами θ,
– $C(\theta)$ – функция, вычисляющая согласованность модели с параметрами θ,
– θ – параметры модели (количество тем).

Тогда задача оптимизации формулируется как задача многокритериальной оптимизации:

$$\frac{min}{\theta} P(\theta), \frac{max}{\theta} C(\theta), \quad (4)$$

Это можно переписать как задачу минимизации с учётом весовых коэффициентов:

$$\frac{min}{\theta} \lambda_1 P(\theta), \frac{max}{\theta} \lambda_2 C(\theta), \quad (5)$$

где λ_1 и λ_2 – коэффициенты, управляющие вкладом каждой из метрик в функцию оптимизации.

Альтернативный способ формулировки – использовать для построения математической модели оптимизации регуляризацию:

$$\frac{min}{\theta} P(\theta) + a^* \frac{1}{C(\theta)}, \quad (6)$$

где a – регуляризационный параметр, который позволяет настроить баланс между перплексией и согласованностью.

Высокие значения гиперпараметров α и γ совместно усиливают согласованность.

5 МЕХАНИЗМ ПОИСКА НЕЯВНЫХ СВЯЗЕЙ МЕЖДУ ТЕРМИНАМИ ВНУТРИ КАЖДОЙ ПРЕДМЕТНОЙ ОБЛАСТИ

Неявные семантические связи между терминами в предметных областях – это скрытые взаимосвязи, которые невозможно увидеть напрямую, но их можно выявить через лингвистические, статистические и онтологические методы. Применяя такие методы к тематическим моделям и семантическим представлениям, можно обнаруживать отношения, отражающие сходство, причинно-следственные связи или функциональные зависимости между терминами.

Для того, чтобы продемонстрировать поиск неявных связей между терминами внутри одного или нескольких тематических кластеров приведем математическую модель.

Пусть T – множество терминов в рассматриваемых предметных областях.

$T_k \subset T$ – термины, принадлежащие теме *k*.

O – онтология предметной области, представляющая отношения между терминами (например, иерархические, синонимические или ассоциативные связи) [9].

E – множество эмбеддингов, представляющих термины в скрытом пространстве признаков.

$Sim(\omega_i, \omega_j)$ – функция, измеряющая семантическое сходство между терминами ω_i и ω_j (например, косинусное сходство их эмбеддингов).

Неявные связи $R_{implict}$ – это скрытые связи, которые можно выявить через анализ вероятностных распределений, контекстных эмбеддингов и анализ графов.

Для двух терминов ω_i и ω_j, принадлежащих одной теме T_k, вероятностная связь может быть определена через их совместное распределение в теме.

Пусть $P(\omega_i|T_k)$ и $P(\omega_j|T_k)$ – вероятности терминов ω_i и ω_j в теме T_k:

$$Sim_{prob}(\omega_i, \omega_j) = \frac{P(\omega_i|T_k) * P(\omega_j|T_k)}{P(\omega_i,\omega_j|T_k)+\epsilon}, \quad (7)$$

где ϵ – это малое значение для предотвращения деления на ноль.

Данная мера показывает степень вероятностной ассоциации, если термины часто встречаются вместе.

Эмбеддинги (вектора слов) позволяют оценивать семантическую близость терминов, не отражённую напрямую в онтологии. Косинусное сходство между векторами двух терминов ω_i и ω_j оценивается следующим образом:

$$Sim_{cos}(\omega_i, \omega_j) = \frac{v(\omega_i) * v(\omega_j)}{\|v(\omega_i)\| \|v(\omega_j)\|}, \quad (8)$$

где $v(\omega_i)$ и $v(\omega_j)$ – эмбеддинги терминов ω_i и ω_j, полученные из тематической модели или предобученной модели (Word2Vec, BERT).

Онтология O предметной области представляет явные семантические связи (например, синонимию). Если для пары терминов не задано явного отношения, мы оцениваем их неявную связь по взвешенной комбинации вероятностного и контекстного сходства:

$$Implict\_Sim_{(\omega_i,\omega_j)} = \alpha * Sim_{prob}(\omega_i, \omega_j) + \beta * Sim_{cos}(\omega_i, \omega_j), \quad (9)$$

где α и β – коэффициенты, которые можно оптимизировать для каждой предметной области.

Для выделения тематических кластеров, отражающих неявные связи, применяются методы, такие как:

– Спектральная кластеризация или разбиение на сообщества в графе [14].

– Поиск «коротких путей»: определяет степень связности, если неявные связи создают замкнутые подграфы (треугольники, циклы) между терминами.

Спектральная кластеризация или разбиение на сообщества в графе позволяет выделить новые тематические кластеры, отразив посредством графа все неявные связи представляющие подмножества терминов [14]. Пусть $C = \{C_1, C_2, \dots, C_m\}$ – набор кластеров, полученных из графа G. Тогда каждый кластер C_i можно рассматривать как группу связанных терминов, которая потенциально представляет новый тематический кластер.

6 МЕТОД РЕШЕНИЯ ПРОБЛЕМЫ ПОИСКА НЕЯВНЫХ СВЯЗЕЙ С УЧЕТОМ УВЕЛИЧЕНИЯ КОЛИЧЕСТВА ТЕМ В ПРЕДМЕТНЫХ ОБЛАСТЯХ

В рамках данного исследования предлагается метод решения проблемы поиска новых неявных связей с учетом увеличения количества тем в предметных областях на примере отлаженных моделей LDA и BERTopic.

На основе вышеописанных механизмов выводится объединенная математическая модель решения проблемы поиска неявных связей с учетом увеличения количества тем в предметных областях. Пусть

$Sim(t_i, c_j)$ – семантическое сходство между терминами.

$F(t_i, c_j)$ – частота термина t_i в теме c_j.

(c_j) – общее количество терминов в теме c_j.

α и β – веса для регулирования вклада.

Объединенная математическая модель на основе двух механизмов примет следующий вид:

$$Sim_{final}(t_i, c_j) = \frac{\alpha * Sim(t_i,c_j) + \beta * F(t_i,c_j) * F(t_i,c_j)}{N(c_j)}, \quad (10)$$

где посредством $Sim_{final}(t_i, c_j)$ происходит выделение новых тем в предметных областях, основываясь на сочетании оценок сходства и частот.

Объединенная математическая модель позволяет также находить неявные связи между терминами, что приводит к более глубокому пониманию семантической структуры данных. Данная модель обеспечивает синергию между механизмами оценки тем и выявления скрытых взаимосвязей, позволяя более эффективно анализировать данные в различных предметных областях. Использование метода решения проблемы поиска неявных связей с учетом увеличения количества тем в предметных областях приводится в разделе 6, содержащем информацию о результатах исследования.

7 РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

В этом разделе представляются результаты исследования, использующие метод решения проблемы поиска неявных связей с учетом увеличения количества тем в предметных областях.

В рамках первичного анализа было приведено тематическое моделирование, которое не накладывает ограничений на количество тем для выбора. В частности, были использованы предварительно обработанные данные в качестве входных данных как для отлаженных моделей LDA, так и для моделей BERTopic. В результате LDA сгенерировал 100 тем, в то время как BERTopic сформировал 1101 тему. Также построенный семантический граф, основанный на иерархической структуре и содержащий в своей основе синонимичные пары терминов, найденных посредством механизма поиска неявных связей между терминами внутри каждой предметной области, позволил найти новые группы связанных терминов и новый тематический кластер. Результаты первичного анализа представлены на рисунке 1. в виде семантического графа на примере набора из международной базы знаний медицинских публикаций PubMed, реферативно-аналитической базы данных Dimensions AI и международного словаря медицинских терминов UMLS.

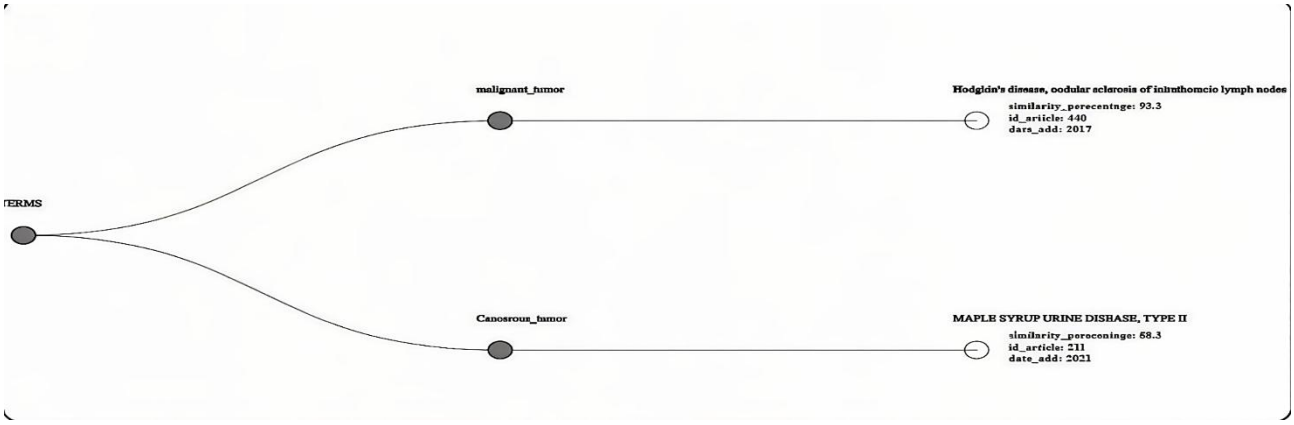

Рис. 1. Семантический граф поиска неявных связей между терминами
Fig. 1. Semantic graph of the search for implicit relationships between terms

Построенный семантический граф, основанный на иерархической структуре и содержащий в своей основе синонимичные пары терминов, представленный на рисунке 1 задействует в своей основе объединенную математическую модель поиска неявных связей между терминами. Связи между диагнозами на рисунке 1 отображаются стрелками. Диагнозы заболеваний подтягиваются из международного словаря медицинских терминов UMLS и отображаются в интерфейсе в виде закрашенных серых кругов. Белые круги на рисунке 1 отображают данные по заболеваниям из международной базы знаний медицинских публикаций PubMed и реферативно-аналитической базы данных Dimensions AI. Каждому заболеванию отображенному в белом круге соответствует id-номер статьи (id_article), дата публикации статьи (date_add), % соответствия статьи диагнозу заболевания в сером круге (similarity_percentage) из базы данных Dimensions AI и PubMed. В представленном на рисунке 1 семантическом графе можно рассмотреть основной кластер, ассоциированный с онкологическими заболеваниями (серые круги), и подкластер, связанный с аутоиммунными

процессами (белые круги). Белые круги, представляющие статьи из PubMed/Dimensions AI, часто находятся на стыке кластеров, что подтверждает их роль в установлении междисциплинарных связей.

Карты меж тематических расстояний на основе диаграмм Венна иллюстрируют связи и различия между темами в рамках нескольких предметных областей. Эти визуальные представления отображают степень корреляции между темами. Расстояния между точками на этих картах символизируют сходство и различие между соответствующими темами. Результаты карт меж темных расстояний на основе диаграмм Венна представлены на рисунках 2 и 3.

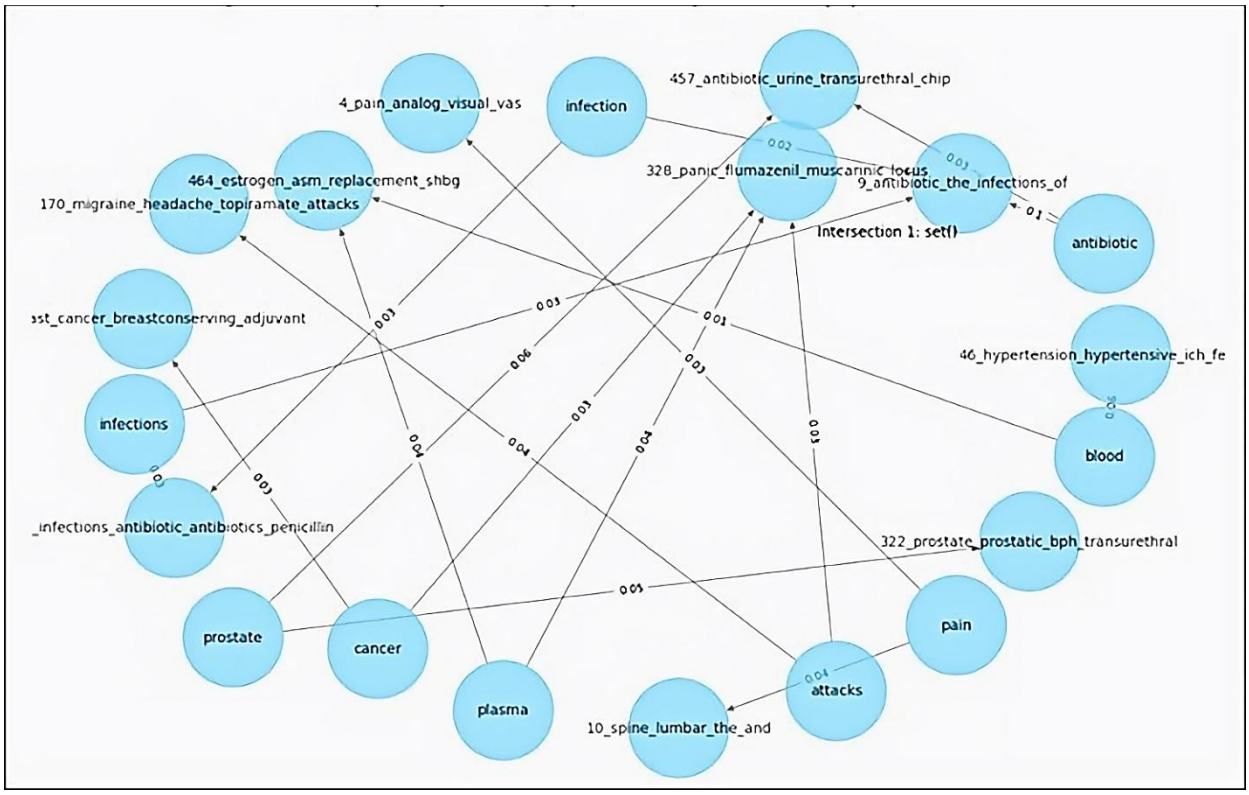

Рис. 2. Результаты карты меж темных расстояний на основе диаграммы Венна
Fig. 2. Results of the inter-dark distance map based on the Venn diagram

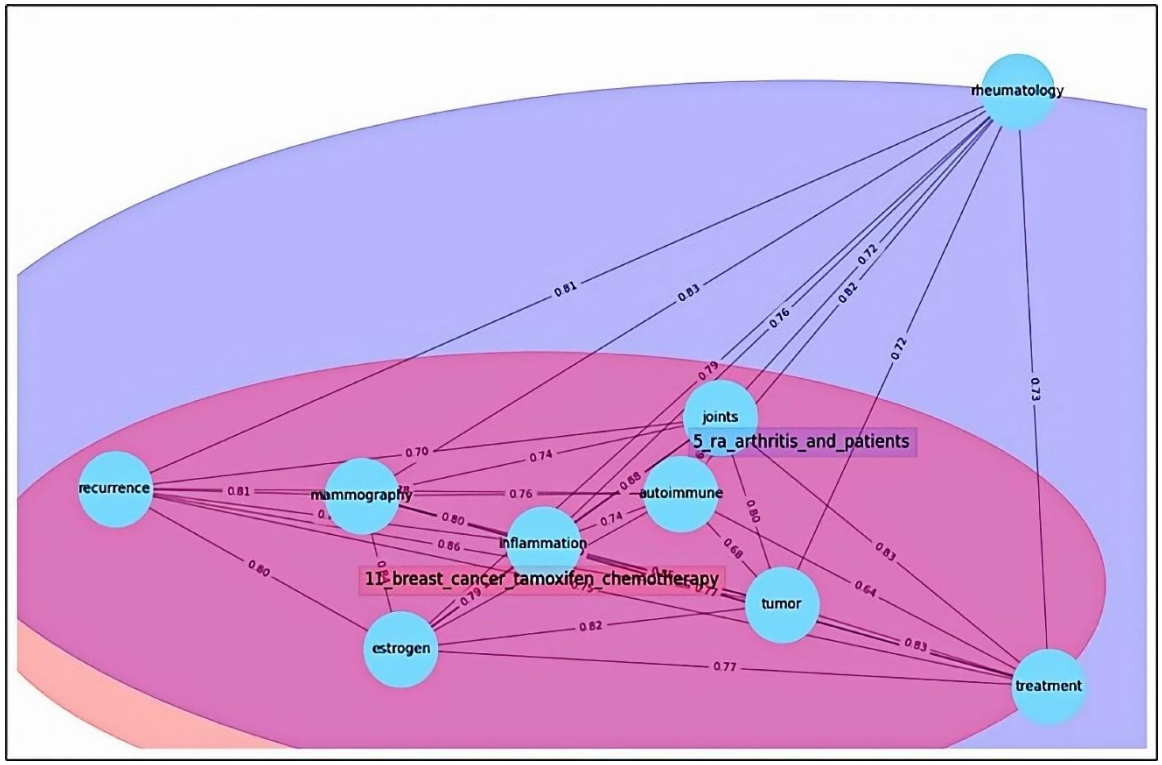

Рис. 3. Результаты карты меж темных расстояний на основе диаграммы Венна с семантическим схождением терминов по 2-ум тематическим кластерам

Fig. 3. Results of a map of inter-dark distances based on a Venn diagram with semantic convergence of terms across 2 thematic clusters

Представленные выше рисунки 2 и 3 карт меж тематических расстояний на диаграмме Венна иллюстрируют связи различных тематических кластеров, построенных на основе методов LDA и BERTopic на примере набора из международной базы знаний медицинских публикаций PubMed, реферативно-аналитической базы данных Dimensions AI и международного словаря медицинских терминов UMLS. Пересечение тематических кластеров происходит в узловых точках терминов. При наложении терминов в рамках 2 тематических кластеров отображается числовая мера приближения, выраженная в вероятности семантического схождения термина из 1-го тематического кластера (пример: «11_breast_cancer_tamoxiten_chemotherapy») к термину из 2-го тематического кластера (пример: «5_ra_arthritis_and_patients») (рисунок 3). Кластер «11_breast_cancer_tamoxiten_chemotherapy» демонстрирует значительное перекрытие (25%) с кластером «5_ra_arthritis_and_patients», что указывает на обширную семантическую близость в области исследования воспалительных процессов. Узловые точки в пересечении тематических кластеров, такие как «inflammation» и «treatment», являются семантическими «мостами» между этими предметными областями. Результаты по семантическому схождению терминов представлены на рисунке 4. Исходя из представленных на рисунке 4 результатов, можно сделать следующие выводы:

– Термин «tumor» демонстрирует высокую семантическую близость с терминами первого кластера (значения до 0,9). Это может указывать на общие патологические механизмы (например, воспаление при раке и аутоиммунных заболеваниях).

– Термин «treatment» во втором кластере имеет низкую связь (0,76) с некоторыми терминами первого кластера, что объясняется разницей в контекстах (например, химиотерапия и противовоспалительная терапия).

– Термин «recurrence» (рецидив) демонстрирует умеренные значения (~0,85), что связано с универсальностью понятия для разных заболеваний.

– Методы LDA и BERTopic выделяют значимые семантические связи между разнородными

медицинскими кластерами, особенно в областях диагностики (mammography) и патогенеза (tumor, inflammation).

Результаты семантического схождения терминов, представленные на рисунке 4. количественно подтверждают наблюдения с семантического графа (рисунок 1) и диаграмм Венна (рисунки 2-3.). Высокие значения схождения для термина «tumor» (~0.9) с терминами кластера «5_ra_arthritis_and_patients» свидетельствуют о его центральной роли в обоих направлениях исследований. Низкие значения для термина «treatment» (0.76) объясняются различной спецификой терапии в онкологии и ревматологии.

Index	Термины	Термины из тематического кластера "Arthritis_and_patients"					Термины из тематического кластера "Breast_cancer_tamoxifen_chemotherapy"					Термины
		inflammation	joints	autoimmune	rheumatology	treatment	mammography	tumor	estogen	treatment	recurrence	
2							0,87	0,9	0,84	0,88	0,88	inflammation
3							0,84	0,88	0,83	0,87	0,85	joints
4							0,79	0,83	0,86	0,76	0,86	autoimmune
5							0,91	0,86	0,87	0,87	0,87	rheumatology
6							0,87	0,86	0,8	1	0,84	treatment
7	mammography	0,87	0,84	0,79	0,91	0,87						
8	tumor	0,9	0,88	0,83	0,86	0,86						
9	estogen	0,84	0,83	0,86	0,87	0,8						
10	treatment	0,88	0,87	0,76	0,87	1						
11	recurrence	0,88	0,85	0,86	0,87	0,84						

Рис. 4. Результаты по семантическому схождению терминов 2-ух тематических кластеров
Fig. 4. Results on the semantic convergence of terms in 2 thematic clusters

Проведенный анализ выявил фундаментальную проблему тематического моделирования: формирования новых неявных семантических связей и поиска оптимального количества тем. В рамках исследования, основанного на анализе данных из международных ресурсов (PubMed, Dimensions AI, UMLS), предложенный метод оптимизации с четырьмя оценочными критериями продемонстрировал следующие результаты: критерий сложности достиг значения 4, а показатель согласованности — 32. Эти данные подтверждают, что использование параметра сложности на уровне 4 эффективно для выявления семантических взаимосвязей между кластерами. Значение согласованности в 32 балла указывает на перспективность метода, которая может быть усилена за счет оптимизации алгоритмов и интеграции дополнительных источников информации. Все критерии были успешно интегрированы в математическую модель, предназначенную для обнаружения скрытых связей между медицинскими терминами.

Использование математической модели поиска неявных связей между терминами позволило рассматривать статическую модель с учетом времени. Теперь формирование оптимального количества тематических кластеров на основе моделей LDA и BERTopic позволило выделять тренд по каждому термину внутри каждой темы. Результаты в виде гистограммы построения модели тренда по каждому термину внутри каждой темы представлены на рисунке 5.

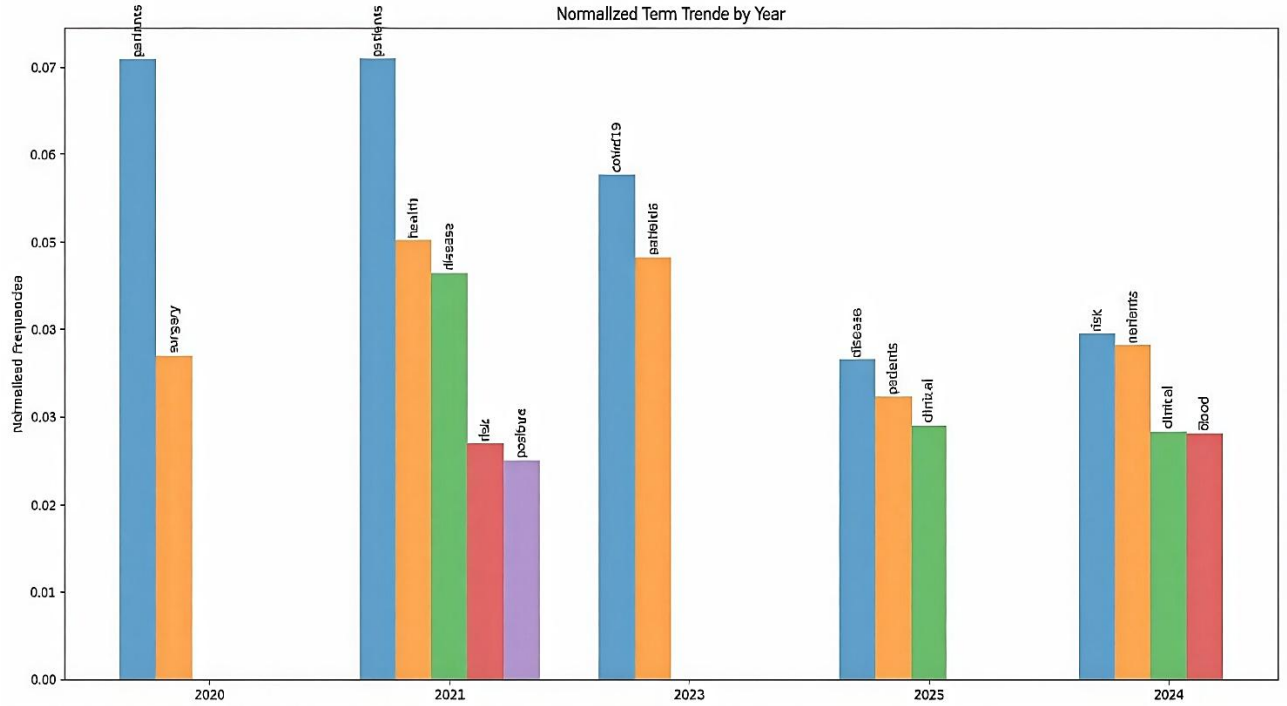

Рис. 5. Результаты построения модели тренда по каждому термину внутри каждой темы
Fig. 5. The results of building a trend model for each term within each topic

Результаты в виде гистограммы построения модели тренда на основе набора данных из международной базы знаний медицинских публикаций PubMed и реферативно-аналитической базы данных Dimensions AI, представленные на рисунке 5 позволили проследить эволюцию тем с выделением тренда по каждому термину внутри каждой темы и помогли исследователям из различных отраслей понять взаимосвязи между темами и терминами в содержании мультимодальных текстов, а также их эволюцию с течением времени.

Анализ тренда по каждому термину внутри каждой темы публикаций набора данных из международной базы знаний медицинских публикаций PubMed и реферативно-аналитической базы данных Dimensions AI демонстрируют возросший научный интерес к области Covid19, охватывающей период 2023 г [5]. Из гистограммы на рисунке 5 очевиден резкий рост частоты встречаемости терминов в публикациях области Covid19, по-видимому, благодаря результатам деятельности Международного консорциума по тяжёлым острым респираторным и новым инфекционным заболеваниям (International Severe Acute Respiratory and emerging Infection Consortium, ISARIC) в части противодействия заболеванию Covid19 [5].

Анализ тренда по каждому термину проводился на корпусе из 7169 медицинских публикаций (PubMed: 2000, Dimensions AI: 5169). Экспериментально подтверждено преимущество предложенной математической модели поиска неявных связей между терминами, позволяющей рассматривать статическую модель с учетом времени. Результатами использования математической модели служит выявление 147 новых семантических связей между терминами, не представленных в исходном тезаурусе UMLS.

Полученные в рамках исследования результаты могут быть использованы для автоматизации процесса обновления и расширения тезаурусов, а также для улучшения качества поиска информации. Дальнейшие исследования могут быть направлены на разработку интерактивных инструментов, позволяющих пользователям исследовать полученные кластеры и самостоятельно выявлять новые связи и тренды в развитии тезаурусов.

ЗАКЛЮЧЕНИЕ

Тематическое моделирование остается основным инструментом в управлении огромными объемами данных, помогая улучшать поиск информации и понимание контекста. Поскольку сложность и объем данных продолжают расти, тематическое моделирование остается незаменимым решением для извлечения информации из неструктурированных текстов.

В рамках данного исследования предложен метод определения оптимального количества тематических кластеров на пересечении предметных областей, основанный на интеграции алгоритмов LDA и BERTopic с математической оптимизацией. Основной акцент сделан на решении двух взаимосвязанных задач:

1. Оптимизации количества тем через комбинирование метрик перплексии и семантической связности, что позволило минимизировать шум и максимизировать интерпретируемость моделей.
2. Анализа эволюции тем с выделением трендов для каждого термина, что обеспечило возможность отслеживания динамики научных интересов и семантических сдвигов во времени.

Применение предложенного подхода к данным из PubMed и Dimensions AI продемонстрировало его эффективность. Семантические графы и диаграммы Венна визуализировали неявные связи между терминами, например, между патологическими процессами (воспаление, опухоль) и методами диагностики. Анализ трендов выявил резкий рост упоминаний терминов, связанных с Covid-19 в 2023 году, что коррелирует с активностью научного сообщества в этой области. Гистограммы эволюции тем подтвердили возможность автоматизации выявления долгосрочных и краткосрочных научных трендов. Кроме того, благодаря математической модели стало возможно более глубоко понимать содержание, представленное в статьях об исследованиях в медицинской и иных областях знаний.

Практическая значимость исследования заключалась в автоматизации обновления тезаурусов за счет динамического анализа семантических связей и улучшении точности поисковых систем через уточнение тематических кластеров.

Список литературы

1. Юрчак В.А. Инструменты решения проблем распознавания и кластеризации данных из документов методами машинного обучения / Золотарев О.В., Юрчак В.А. // ИВД. – 2023. – № 2 (98). – С. 156-164.
2. Корней А.О. Семантико-статистический алгоритм определения категорий аспектов в задачах сентимент-анализа / Корней А.О., Крючкова Е.Н. // Известия ЮФУ. Технические науки. – 2020. – №6 (216). – С. 66-74.
3. Клименко С.В. Использование онтологического подхода для анализа текстов естественного языка / Клименко С.В., Золотарев О.В., Шарин М.М. // Вестник российского нового университета. Серия: сложные системы: модели, анализ и управление. – 2017. – С. 67-71.
4. Хакимова А.Х. Подходы к созданию многоязычного лексического ресурса для семантометрической оценки межъязыкового семантического подобия текстов / Хакимова А.Х., Золотарёв О.В., Шарнин М.М. // Нижегородский государственный архитектурно-строительный университет, Научно-исследовательский центр физико-технической информатики. – 2019. – С. 319-324.
5. Золотарев О.В., Хакимова А.Х., Шарнин М.М. Разработка методов интеллектуального анализа научных публикаций для мониторинга приоритетных направлений развития превентивной и персонализированной медицины / О.В. Золотарев, А.Х. Хакимова, М.М. Шарнин // Вестник Российского нового университета. Серия «Сложные системы». – 2019. – С. 110-117.
6. Методика построения ассоциативно-иерархического портрета предметной области: иерархия категорий / Клименко С.В. и др. // Автономная некоммерческая организация «Институт физико-технической информатики». – 2017. – С. 251-260.
7. Модель и технология извлечения новых терминов из медицинских текстов / Золотарев О.В. и др. // Информатика и ее Применения. – 2022. – С. 80-86.
8. Мера подобия текстов как инструмент оценки интертекстуальности при анализе больших коллекций документов / Золотарев О.В. и др. // Вестник российского нового университета. Серия: сложные системы: модели, анализ и управление. –2016. – С. 62-71.
9. Программа выделения терминов из корпуса текстов / Золотарев О.В. и др. // Автономная

некоммерческая организация высшего образования «Российский новый университет». – 2023. – С. 1-2.

10.Программа построения структурированного корпуса текстов на основе электронных баз публикаций / Золотарев О.В. и др. // Автономная некоммерческая организация высшего образования «Российский новый университет». – 2023. – С. 1-2.

11.Farea A., Tripathi Sh., Glazko G., Emmert-Streib F. Investigating the optimal number of topics by advanced text-mining techniques: Sustainable energy research // Engineering Applications of Artificial Intelligence. V. 136, part A. – 2024. Available from: https://doi.org/10.1016/j.engappai.2024.108877.

12.Li Y., Wang W., Yan X., Gao M., Xiao M. Research on the Application of Semantic Network in Disease Diagnosis Prompts Based on Medical Corpus / International Journal of Innovative Research in Computer Science and Technology (IJIRCST). – 2024. – 1-9 p. Available from: https://doi.org/10.55524/ijircst.2024.12.2.1

13.Бручес Е.П. Методы и алгоритмы распознавания и связывания сущностей для построения систем автоматического извлечения информации из научных текстов: дис. на соискание учёной степени кандидата технических наук. – Новосибирск: Федеральное государственное бюджетное учреждение науки Институт систем информатики им. Ершова, 2021. – 112 с.

14.Дударин П.В. Исследование и разработка моделей и методов нечеткой кластеризации коротких текстов: дис. на соискание учёной степени кандидата технических наук. – Ульяновск: «Ульяновский государственный технический университет», 2021. – 141 с.

15.Тутубалина Е.В. Модели и методы автоматической обработки неструктурированных данных в биомедицинской области: дис. доктора компьютерных наук. – Казань: Казанский (Приволжский) федеральный университет, 2023. – 225 с.

16.Корней А.О. Методы и алгоритмы аспектного анализа тональности на основе гибридной семантико-статистической модели естественного языка: дис. на соискание ученой степени кандидата технических наук. – Барнаул: Федеральное государственное бюджетное образовательное учреждение высшего образования «Алтайский государственный технический университет им. И.И. Ползунова», 2021. – 134 с.

References

1. Yurchak V.A. Tools for solving problems of recognition and clustering of data from documents using machine learning methods / Zolotarev O.V., Yurchak V.A. // IVD. – 2023. – No. 2(98). – 156-164 P.

2. Kornei A.O. Semantic-statistical algorithm for determining the categories of aspects in sentiment analysis problems / Kornei A.O., Kryuchkova E.N. // Bulletin of the Southern Federal University. Technical sciences. – 2020. – No. 6 (216). – 66-74 p.

3. Klimenko S.V. Using the ontological approach to analyze natural language texts / Klimenko S.V., Zolotarev O.V., Sharin M.M. // Bulletin of the Russian New University. Series: Complex Systems: Models, Analysis and Management. – 2017. – P. 67-71.

4. Khakimova A.Kh. Approaches to Creating a Multilingual Lexical Resource for Semantometric Assessment of Interlingual Semantic Similarity of Texts / Khakimova A.Kh., Zolotarev O.V., Sharnin M.M. / Nizhny Novgorod State University of Architecture and Civil Engineering, Research Center for Physics and Engineering Informatics. – 2019. – P.319-324.

5. Zolotarev O.V., Khakimova A.Kh., Sharnin M.M. Development of Methods for Intelligent Analysis of Scientific Publications to Monitor Priority Directions for the Development of Preventive and Personalized Medicine / O.V. Zolotarev, A.Kh. Khakimova, M.M. Sharnin // Bulletin of the Russian New University. Series "Complex Systems". – 2019. – P. 110-117.

6. Methodology for constructing an associative-hierarchical portrait of a subject area: hierarchy of categories / Klimenko S.V. et al. // Autonomous Non-Commercial Organization "Institute of Physical and Technical Informatics". – 2017. – P. 251-260.

7. Model and technology for extracting new terms from medical texts / Zolotarev O.V. et al. // Informatics and its Applications. – 2022. – 80-86 p.

8. The measure of similarity of texts as a tool for assessing intertextuality in the analysis of large collections of documents / Zolotarev O.V. et al. // Bulletin of the Russian New University Series: complex systems: models, analysis, and management. – 2016. – 62-71 p.

9. Program for the allocation of terms from the corpus of texts / Zolotarev O.V. et al. // Autonomous non–profit organization of Higher Education "Russian New University" – 2023. – 1-2 p.

10.A program for building a structured corpus of texts based on electronic databases of publications / Zolotarev O.V. et al. // Autonomous non–profit organization of Higher Education "Russian New University" – 2023. – 1-2 p.

11.Farea A., Tripathi Sh., Glazko G., Emmert-Streib F. Investigating the optimal number of topics by advanced

text-mining techniques: Sustainable energy research // Engineering Applications of Artificial Intelligence. V. 136, part A. – 2024. URL: https://doi.org/10.1016/j.engappai.2024.108877.

12. Li Y., Wang W., Yan X., Gao M., Xiao M. Research on the Application of Semantic Network in Disease Diagnosis Prompts Based on Medical Corpus / International Journal of Innovative Research in Computer Science and Technology (IJIRCST). – 2024. – 1-9 p. Available from: https://doi.org/10.55524/ijircst.2024.12.2.1

13. Bruches E.P. Methods and algorithms for recognizing and linking entities for building systems for automatic extraction of information from scientific texts: dis. for the degree of candidate of technical sciences. – Novosibirsk: Federal State Budgetary Institution of Science Institute of Informatics Systems named after Ershov, 2021. – 112 p.

14. Dudarin P.V. Research and development of models and methods for fuzzy clustering of short texts: dis. for the degree of candidate of technical sciences. – Ulyanovsk: "Ulyanovsk State Technical University", 2021. – 141 p.

15. Tutubalina E.V. Models and methods of automatic processing of unstructured data in the biomedical field: dis Doctor of Computer Science – Kazan: Kazan (Volga Region) Federal University, 2023 – 225 p.

16. Korney A.O. Medical Methods and scientific algorithms of the aspect analysis table for determining tonality based on the corpus of a hybrid presented semantic-statistical model of prompts of a natural language: dis. for the abso academic UMLS degree of Candidate of International Technical Semantic Sciences. – Barnaul: allocation of the Federal semantic state budgetary foreign educational institution of higher international education "International Altai State Medical Technical University named after I.I. Polzunov", the consequence of 2021. – 134 p.

Юрчак Владимир Александрович, аспирант 3-го года обучения кафедры ИСЭУ, АНО ВО «Российский новый университет (РосНОУ)», г. Москва, Россия
Yurchak Vladimir Alexandrovich, 3rd year postgraduate student of the ISEU Department, Russian New University (RosNOU), Moscow, Russia