DOI: 10.19181/4m.2023.32.1.2
EDN: SJPPOZ

В.А. Ващенко
(*Москва*)

ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ДЛЯ КОРОТКИХ ТЕКСТОВ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ[1]

Устойчивый рост популярности социальных сетей в качестве средства коммуникации актуализирует методологические вопросы, связанные с особенностями обработки коротких текстов, обладающих меньшим семантическим контекстом, чем крупные тексты, широко используемые для обучения и тестирования моделей машинного обучения для работы с текстовыми данными. Тематическое моделирование – метод машинного обучения «без учителя», нацеленный на агрегацию текстов в тематические кластеры, – имеет множество академических и практических приложений в случаях отсутствия подробной разметки текстовых данных. Однако качество работы алгоритмов тематического моделирования может ограничиваться полнотой семантического контекста, необходимого для качественного числового представления единицы текста. В этой статье рассматриваются шесть разных подходов к тематическому моделированию, основанных на различающихся принципах концептуализации текста и тем. Сравнивается качество работы указанных алгоритмов на наборе русскоязычных комментариев в сети TikTok и проводится формальная оценка скорости и когерентности результирующих тем.

Василиса Андреевна Ващенко – стажер-исследователь Международной лаборатории прикладного сетевого анализа Национального исследовательского университета «Высшая школа экономики», Москва, Россия. Email: vvashchenko@hse.ru.

[1] Статья подготовлена в ходе проведения исследования в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ).

Ключевые слова: тематическое моделирование, анализ текстовых данных, блокмоделинг, прикладной сетевой анализ, анализ социальных медиа, трансформерные модели

Введение

Тематическое моделирование – метод вычленения тематических кластеров в корпусе текстов – является важным направлением в методологических исследованиях в области социологии, поскольку предоставляет возможность количественного изучения тематической композиции текстовых материалов: структуры дискуссии, художественных или научных текстов.

Хотя алгоритмы тематического моделирования склонны обобщать выделяемые темы и хуже справляются с выделением узких тематических групп, чем ручное кодирование, что иногда приводит к расхождениям с результатами ручной разметки [1], они успешно применялись в рекомендательных системах [2], наукометрических задачах [3], анализе дискурса [2; 4; 5], автоматической идентификации событий в новостях и социальных медиа [6; 7; 8] и во многих иных практических приложениях, включая анализ нетекстовых источников (например, изображений [9] и аудио [10]).

Вместе с популяризацией социальных сетей в качестве ключевых инструментов массовой коммуникации в современном мире растет и интерес исследователей к социальным медиа как к источнику знания об общественном мнении. В свою очередь крупные объемы информации, производимые пользователями социальных сетей на ежедневной основе, диктуют новые требования к масштабам эмпирической работы [5], необходимой для полноценного описания исследуемого общественного феномена, что стимулирует использование количественных техник анализа текстовых данных исследователями. Как отмечает А. Бызов, алгоритмы тематического моделирования способны оказать исследователю поддержку в разметке или выделении признаков из текстовых данных, особенно

в ситуации, когда объем массива текста затрудняет ручное кодирование [11]. Так, методы тематического моделирования использовались для исследования общественного мнения о COVID-19 на базе больших данных в Twitter [6; 12]. Однако многие алгоритмы, используемые для задач тематического моделирования, расходятся в своих базовых предпосылках с форматом коммуникации, характерным для социальных медиа: короткие тексты сообщений в социальных сетях, подобных Twitter, не позволяют достигать планок качества, сопоставимых с работой тех же инструментов на длинных текстах ввиду недостаточности контекста для слов-токенов [13; 14; 15]. Учитывая, что многие тексты, производимые пользователями социальных сетей, ограничены платформенными лимитами[1], длинные тексты зачастую недоступны исследователям онлайн-дискурса. Следовательно, необходимо формализованное сравнение качества разных алгоритмов тематического моделирования для коротких текстов с целью обнаружения наиболее эффективного с вычислительной и интерпретационной точек зрения подхода. Несмотря на наличие подобных усилий в современном академическом поле, методы, анализируемые исследователями, как правило, ограничены вариациями и адаптациями латентного размещения Дирихле (Latent Dirichlet Allocation, LDA) и словарными эмбеддингами – векторными представлениями, репрезентирующими семантический контекст слова. В свою очередь мы предлагаем проанализировать эффективность методов, альтернативных конвенционально используемому LDA. Мы включаем в анализ методы тематического моделирования, основанные как на вероятностном моделировании соприсутствия слов, так и на методах выделения сообществ в бимодальных сетях и кластеризации предобученных словарных эмбеддингов.

[1] Все платформы массовой коммуникации используют лимиты символов на разные виды текстов, производимых на платформе: публикации, комментарии, описания и т.д. Как правило, эти лимиты достаточно низкие, что заметно на примере двух из трех крупнейших социальных сетей в России на 2023 г., согласно данным исследовательской компании Mediascope [16]: для TikTok это 150 символов на комментарий, для «ВКонтакте» – 280 символов на комментарий.

Разнообразие сравниваемых методов позволяет более конкретно изучить связь между устройством подхода к тематическому моделированию и особенностями результирующих тематических кластеров, а также отметить значимые направления потенциального развития наличествующих инструментов для задач анализа коротких текстов. В настоящей статье проводится формальное сравнение качества работы новых алгоритмов тематического моделирования на базах данных новостных сводок и комментариев в социальных медиа, а также предлагаются пути выбора модели и улучшения ее качества для задач тематического моделирования в социальных исследованиях.

Обзор исследований по теме тематического моделирования для коротких текстов

Методы тематического моделирования получили широкое признание вместе с появлением в конце 1990-х гг. ныне конвенциональных подходов вероятностного латентного семантического анализа (Probabilistic Latent Semantic Analysis, PLSA) [17] и его последующей адаптации в байесовской парадигме – LDA, моделирующей, в отличие от PLSA, глобальную структуру тем, предполагая, что слова в разных документах одного корпуса производны из одного и того же набора тем [18]. Важным нововведением LDA стало использование распределения Дирихле в качестве априорного для распределения тем в документах и слов в темах[1]. Этот шаг выполняет регуляризирующую функцию для тематической модели, выравнивая распределения тем за счет ограничения экстремальных значений и снижая чувствительность модели к отдельным документам.

Вероятностные модели представления парных связей между документами и словами-токенами хорошо подошли к анализу

[1] За более подробным описанием технической части алгоритма LDA читатель может обратиться к оригинальной статье [18] или существующим обзорам на русском языке [19; 20].

художественных и научных текстов, поскольку качество предсказания для таких моделей улучшается пропорционально количеству наблюдений и объему контекста: чем больше информации о паттернах соприсутствия слов представлено в обучающем корпусе, тем более связные латентные семантические кластеры производятся LDA. Однако для коротких текстов, доминирующих в современных медиа и представляющих непосредственный интерес для исследований общественного мнения [21], традиционные алгоритмы могут не подходить ввиду двух ограничений.

1. Вероятностные модели опираются на локальные паттерны соприсутствия в рамках документа – семантический контекст слова, который в коротких документах зачастую оказывается недостаточным.

2. Предпосылка наличия распределения тем внутри документа спорна для коротких текстов, поскольку многие из них, в отличие от более длинных документов, содержат только одну тему [15; 22].

Для разрешения этих проблем существует несколько способов: так, короткие текстовые данные предлагается обогащать метаданными [23], тегами авторов [24] и хештегами [22], использовать длинные тексты для обучения модели тематического моделирования для предсказания тем на коротких текстах [25], агрегации коротких текстов в «псевдодлинные» по заданному признаку, чтобы применять к ним традиционные методы тематического моделирования [14]. Отмечается, что использование метаданных или иных форм дополнительной информации методологически нежелательно, поскольку во многих случаях подобные данные могут быть недоступны [15]. Более того, эти адаптации не решают ряд иных значимых методологических проблем – отсутствия формального критерия выбора количества тем[1] и обоснования предпосылки о соответствии распределения тем и слов внутри

[1] Стоит отметить, что формальные критерии выбора количества тем развиваются и обсуждаются в литературе: например, примечательно применение энтропийного подхода к задаче настройки этого гиперпараметра [19].

темы распределению Дирихле [26]. Как отмечают авторы сетевого подхода к тематическому моделированию М. Герлах, Т.П. Пейшото и Э.Г. Алтманн, несмотря на высококачественные результаты работы алгоритма LDA, выбор распределения Дирихле для моделирования распределения тем в текстовом корпусе является его ограничением, поскольку репрезентация текста в LDA противоречит иным известным в лингвистике паттернам в языке (к примеру, закону Ципфа), указывающим на неравномерность в частотности появления и соприсутствия слов в текстах [26].

В качестве альтернативы семантическому обогащению модели за счет метаданных или расширения контекста Герлах, Пейшото и Алтманн предлагают подойти к задаче тематического моделирования как к задаче выделения сообществ в сетях. Доказывая формальную эквивалентность между PLSA и стохастической блокмоделью (Stochastic Blockmodel, SBM[1]), авторы рассматривают LDA как частный кейс своего непараметрического алгоритма, предлагающего возможность отказа от предпосылок об униформном распределении тематических кластеров в корпусе за счет использования иерархии априорных распределений для моделирования распределения тем [26]. Иерархическая модель допускает гетерогенность в данных и неоднородность выделяемых тем: темы могут быть более или менее «плотными», а также объединяться в структуры более высокого порядка при подъеме по «ступеням» иерархии. Более того, непараметрические модели позволяют избавиться от количества тем как предзаданного гиперпараметра и моделировать количество тем в ходе обучения [28]

Идея иерархичности в формировании тем применяется также в новых разработках среди моделей тематического моделирования,

[1] Стохастические блокмодели (SBM) являются генеративными вероятностными моделями, используемыми для группировки вершин в сетях в блоки или сообщества. Главной предпосылкой модели является положение, указывающее, что вершины, относящиеся к одному блоку, имеют более высокую вероятность связи между ними, чем вершины, относящиеся к разным блокам [27].

ориентированных на расширение семантического контекста анализируемых документов [29]. Так, активно развивается приложение трансформерных моделей к задаче тематического моделирования, поскольку эмбеддинги, используемые трансформерами, позволяют репрезентировать сложные семантические структуры в текстах на естественном языке, сохраняя больше локальной информации, чем статические предобученные эмбеддинги. «Трансформером» называется разновидность нейросетевой архитектуры, широко применяемой в обработке естественного языка; наиболее значимым отличием трансформеров является использование механизма внимания (attention) при тренировке для оценки взаимной важности слов в контексте друг для друга, а также внедрение позиционных меток в эмбеддинги слов для сохранения последовательности в текстовых данных. Создание представлений текста при помощи трансформеров предоставляет возможность не только различать омонимичные выражения, но и учитывать структурную позицию токена[1] в документе при кодировке [30]. В случае использования представлений трансформера для выделения тем, тематическое моделирование заключается в кластеризации эмбеддингов с сокращенной размерностью, где использование иерархических методов кластеризации позволяет выделять кластеры разного размера и плотности, а значит – достигается ранее упомянутая репрезентация гетерогенности в данных [29; 31].

Таким образом, область тематического моделирования активно развивается в сторону расширения типов данных, для которых выделение латентных семантических структур работает эффективно и корректно, включая короткие тексты. Наиболее популярным на-

[1] В лингвистических моделях оптимальной единицей анализа не всегда является слово. В зависимости от задачи, единицей анализа могут выступать слово, словосочетание или часть слова (слог или пары/триады букв). Такая единица анализа называется токеном (от *англ*. ‘token’ – знак, символ). При обучении моделей типа «трансформер» в качестве токенов, как правило, выступают высокочастотные сочетания букв, необязательно составляющие полный слог.

правлением анализа является обогащение контекста коротких текстов при помощи предобученных эмбеддингов или иных видов дополнительных данных для модели, однако присутствуют также и шаги в сторону смены концептуальной рамки подхода к задаче тематического моделирования, одним из которых является использование методов сетевого анализа для выделения тематических кластеров.

Методология

Выбор архитектур для тематического моделирования

Поскольку ключевой целью настоящей статьи является сравнение разных подходов к тематическому моделированию, мы прибегаем к сравнению алгоритмов, относящихся к разным группам методов. Опираясь на категоризацию, предложенную в обзоре А. Абдельразика с соавторами [31], мы используем модели из трех разных выделенных в обзоре групп: алгебраических, вероятностных и нейросетевых моделей тематического моделирования (табл. 1).

К алгебраическим моделям в нашей подборке алгоритмов относится NMF. Этот алгоритм позволяет выявить темы путем разложения матрицы «документ-токен» на неотрицательные матрицы (матрицы, все элементы которых больше или равны нулю), отражающие связи «документ-тема» и «тема-токен», что обеспечивает более интерпретируемое и контекстуально значимое представление текстовых данных, поскольку в случае репрезентации связей токенов, тем и документов негативные значения неинтерпретируемы.

Вероятностные модели представлены LDA и SBMTM. LDA моделирует тематическую структуру, вероятностно распределяя слова по темам, а темы по документам на основании априорного распределения Дирихле, что позволяет обнаруживать скрытые тематические паттерны. В SBMTM темы выделяются как сооб-

Таблица 1

КАТЕГОРИЗАЦИЯ СРАВНИВАЕМЫХ АЛГОРИТМОВ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Модель	Документация	Категория
Non-Negative Matrix Factorization (NMF) [32]	Реализация модели в библиотеке Gensim для Python: https://radimrehurek.com/gensim/models/nmf.html	Алгебраические
Latent Dirichlet Allocation (LDA) [18]	Реализация модели в библиотеке Gensim для Python: https://radimrehurek.com/gensim/models/ldamodel.html	Вероятностные
Hierarchical Stochastic Block Model for Topic Modeling (SBMTM) [26]	Реализация модели при помощи библиотеки graph-tool для Python: https://github.com/martingerlach/hSBM_Topicmodel	
Embedded Topic Model (ETM) [33]	Исходный код для Python и PyTorch: https://github.com/adjidieng/ETM	Нейросетевые
Product-of-Experts LDA (ProdLDA) [34]	Исходный код для PyTorch реализации: https://github.com/estebandito22/PyTorchAVITM	
Contextualized Topic Model (CTM) [35]	Исходный код для Python и подробные примеры применения: https://github.com/MilaNLProc/contextualized-topic-models	
BERTopic [29]	Описание алгоритма и инструкции по установке: https://maartengr.github.io/BERTopic/index.html Исходный код для Python: https://github.com/MaartenGr/BERTopic	

щества в бимодальной сети соприсутствия слов, где в качестве двух типов вершин выступают документы и слова, при помощи иерархических стохастических блокмоделей. Иерархичность подхода позволяет генерализовать предпосылки о распределениях, описывающих связь между темами и словами. Герлах, Пейшото и Альтманн утверждают, что такой подход позволяет формировать более качественные темы за счет учета как плотных, так и разреженных групп разного размера [26].

Ввиду сравнительной новизны нейросетевых моделей и перспективности их приложения к анализу коротких текстов за счет более нюансированного задействования семантического контекста, они наиболее широко представлены в нашей подборке алгоритмов для анализа. Мы рассмотрим четыре модели: ETM, ProdLDA, CTM и BERTopic. ETM объединяет в себе векторные представления слов и вероятностное моделирование тем. В отличие от традиционных методов, ETM представляет документы и темы в непрерывном семантическом пространстве, что позволяет словам вносить переменный вклад в темы. Это позволяет ETM улавливать тонкие семантические связи между словами и темами, предоставляя более гибкий и интерпретируемый подход к моделированию тем.

ProdLDA является адаптацией LDA, задействующей в своей архитектуре вариационные автокодировщики[1] (Variational

[1] Вариационные автокодировщики (VAE) – это генеративная модель, предназначенная для обучения вероятностному представлению входных данных и генерации новых образцов на основе изученного распределения. VAE состоят из кодировщика, который отображает входные данные в вероятностное скрытое пространство, и декодировщика, который восстанавливает данные из образцов, взятых из скрытого пространства. Обучение VAE побуждает выучиваемое латентное пространство следовать заданному распределению вероятностей и облегчает генерацию новых, значимых образцов данных. В целом VAE используют вероятностные принципы для обучения структурированному и непрерывному представлению входных данных, что делает их крайне полезными для таких задач, как генеративное моделирование, синтез данных и обучение латентных векторных представлений (эмбеддингов).

Autoencoders, VAE) для улучшения репрезентации тем [34]. Еще одной значимой чертой ProdLDA является использование product of experts («произведение экспертов») подхода: такие модели используют несколько «экспертных» моделей, специализирующихся на разных аспектах/частях данных, вместо одной, для вероятностного моделирования тренировочных данных. В традиционной модели LDA предполагается, что темы генерируются независимо для каждого документа. Вместо предположения о независимости ProdLDA моделирует совместное распределение тем и документов как произведение распределений моделей-«экспертов». Каждый «эксперт» связан с отдельным документом, и каждый эксперт вносит свой вклад в общее распределение. Итоговое распределение для документа получается путем взятия произведения распределений экспертов. Это позволяет модели отражать более сложные зависимости между темами и документами, что может быть особенно полезно при наличии нетривиальных зависимостей, которые немогут быть адекватно отражены при допущении независимости.

Модель CTM учитывает контекст каждого слова внутри документа, что позволяет генерировать темы, учитывающие нюансы содержания документа. Модель использует эмбеддинги предобученной модели-трансформера типа BERT для изучения контекстуализированных представлений слов, обеспечивая более точное и динамичное моделирование тем.

Наконец, в BERTopic темы выделяются при помощи иерархической кластеризации внутренних представлений текста предобученной модели трансформера. В рамках использования обеих моделей для запуска необходимо указать предобученную модель-трансформер, векторные представления слов которой ложатся в основу моделирования близости между ними. Мы используем paraphrase-multilingual-mpnet-base-v2 – наилучшую по качеству на открытых бенчмарках предобученную мультиязычную модель

из библиотеки sentence-transformers[1]. Важно отметить различие между CTM и BERTopic в том, что последняя модель выделяет темы как кластеры в пространстве словарных эмбеддингов сокращенной размерности. Мы используем Uniform Manifold Approximation and Projection (UMAP)[2] для сокращения размерности BERT-эмбеддингов и Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)[3] для кластеризации. UMAP выбирается за счет высокого качества сохранения локальной структуры данных, в то время как HDBSCAN способен выделять кластеры разного размера и плотности, что позволяет обнаруживать разнородные темы в документах. Значимой компонентой качества BERTopic является алгоритм c-TF-IDF (class Total Frequency – Inverse Document Frequency) взвешивания – адаптация классического TF-IDF взвешивания, используемая для отражения уникальности слов в документе по отношению к другим словам той же темы. Она направлена на выделение слов, которые являются хорошо различающимися для конкретной темы, в то время как слова, которые являются общими для всего корпуса, отводятся на второй план [29].

Для обучения ETM, – обогащенного предобученными статическими словарными эмбеддингами метода LDA, – в рамках наших экспериментов сравниваются две предобученные модели статических словарных эмбеддингов (табл. 2).

1. GloVe (Global Vectors) для русского языка Navec [36]. Выбор обоснован малым объемом памяти, необходимой для их использования и дообучения при большом объеме словаря: используемые эмбеддинги, обученные на массиве русскоязычной художествен-

[1] Таблица сравнения доступных моделей представлена по ссылке: https://sbert.net/docs/pretrained_models.html (дата обращения: 05.01.2024).

[2] Подробная документация Python-имплементации UMAP доступна по ссылке: https://umap-learn.readthedocs.io (дата обращения: 05.01.2024).

[3] Подробная документация Python-имплементации HDBSCAN доступна по ссылке: https://hdbscan.readthedocs.io (дата обращения: 05.01.2024).

ной литературы[1], покрывают 98% слов в художественных текстах, занимая 50,6 Мб памяти.

2. Word2Vec (Continuous Skipgram) [37; 38] – эмбеддинги для русского языка, обученные на Национальном корпусе русского языка (НКРЯ). В отличие от Navec, эта модель обучена на словаре с тегами частей речи, что позволяет в некоторых случаях решить проблему омонимии. Предобученная модель представлена в открытом доступе как часть библиотеки Gensim для Python[2].

Таблица 2

ОПИСАНИЕ ИСПОЛЬЗУЕМЫХ ЭМБЕДДИНГОВ

Модель	Размерность	Тип эмбеддинга	Размер словаря (103)	Размер модели (*Mb*)
Navec	300	GloVe	500	50,6
ruscorpora_300	300	Continuous Skipgram	180	198,8

Так, мы сравниваем две предобученные модели статических словарных эмбеддингов, основанные на разных методах формирования векторных представлений слов: GloVe [39] и Continuous Skip-Gram. Continuous Skip-Gram является моделью локального предсказания контекста. Ее цель – предсказать контекстные слова заданного целевого слова [37]. Модель обучается таким образом, чтобы максимизировать вероятность предсказания контекстных слов по заданному слову. GloVe, в свою очередь, основан на глобальном статистическом подходе: целью обучения является минимизация разницы между точечным произведением векторов слов

[1] В настоящей работе используется дефолтная модель Navec, документация которой доступна по ссылке: https://github.com/natasha/navec (дата обращения: 05.01.2024).

[2] Документация соответствующих методов и список доступных моделей доступны по ссылке: https://radimrehurek.com/gensim/models/word2vec.html (дата обращения: 15.12.2023).

и логарифмом вероятностей соприсутствия в корпусе на основе матрицы соприсутствия. В отличие от Word2Vec, эмбеддинги GloVe используют не только слова в непосредственной близости друг друга (контекст), но и глобальные статистики совстречаемости слов друг с другом, что позволяет также определить сравнительную значимость слов в тексте [39]. Так, выбранные эмбеддинги представляют альтернативные подходы к векторному представлению слов.

При работе с хештегами используется классический LDA, для остальных массивов данных добавляются словарные эмбеддинги и модель идентифицируется как ETM [33]. Поскольку многие хештеги содержат те или иные разновидности авторского написания (нарочитые ошибки, совмещение фразы в одно слово и т.д.), применение к ним предобученных векторных представлений слов неэффективно из-за чувствительности последних к корректности написания, роду и числу слов: многие хештеги не войдут в словарь моделей предобученных представлений слов.

Сопоставление вышеописанных алгоритмов (табл. 1), а также разных моделей предобученных эмбеддингов (табл. 2) позволяет проанализировать различия в качестве моделирования тем в зависимости от подхода к представлению текстовых данных, а также агрегации документов, включая внедрение иерархически организованных групп.

Критерии сравнения

Для сравнения избранных архитектур используется несколько критериев: время, затрачиваемое на вычисления, когерентность (связность) выделяемых тем, разнообразие/схожесть тем, а также качество классификации текстов как принадлежащих теме на основании смоделированных тематических кластеров. Следует отметить, что в случае с тематическим моделированием существует проблема малого количества «внешних» критериев качества – метрик качества, опирающихся на внешнюю по отношению к продукту моделирования информацию для верификации (как, например, метка

класса в задаче классификации), а не только внутренние свойства результирующих тематических групп [5]. Большинство метрик, используемых для проверки и сравнения алгоритмов тематического моделирования, являются «внутренними»: рассчитываются исходя из собственных свойств тематических групп, смоделированных алгоритмом, и не имеют внешнего референта. Однако на размеченных датасетах задача тематического моделирования может быть трансформирована в задачу классификации, что позволяет использовать «внешние» меры качества.

Метрики когерентности выделяемых алгоритмами тематического моделирования групп измеряют «расстояние» между словами внутри кластера: в случае, если слова, вошедшие в один и тот же кластер, «близки» друг к другу, когерентность кластера будет принимать высокое значение, в обратном случае – низкое. Как правило, для оценки когерентности кластера используются не все входящие в него слова, а N наиболее важных. Мы используем две меры когерентности тем: нормализованную поточечную взаимную информацию (Normalized Pointwise Mutual Information, NPMI) и UMass.

Кратко рассмотрим каждую из них.

1. NPMI количественно определяет родственность слов в теме, учитывая закономерности их совместного появления. Это нормализованная версия поточечной взаимной информации (Pointwise Mutual Information, PMI) между парами слов. PMI измеряет отличие вероятности одновременного появления двух слов в документе относительно ожидаемой при их независимости. Учитывая как силу связи, так и редкость пар слов, что обеспечивает сбалансированную оценку связности, NPMI показывает хорошую корреляцию с человеческой оценкой [28; 40], однако на качество оценки может значимо влиять выбранный размер контекстного окна[1].

[1] «Контекстным окном» называется диапазон слов относительно «центрального» слова, включаемый в анализ при расчете метрик, опирающихся на закономерности и частоты соприсутствия слов. Для примера: контекстное окно размера 3

2. UMass вычисляет PMI между главными словами в теме и их совпадением в референтном корпусе: насколько чаще эти слова встречаются вместе, чем в случайном сценарии [41]. Как и NPMI, эта метрика сенситивна к качеству базового корпуса, однако отличается тем, что измеряет вероятность асимметрично (значение зависит от того, какое слово является «центром», а какое – «контекстом») для топ-слов темы. Эта мера считается наиболее нечувствительной к шуму и является одной из самых быстрых метрик вычисления когерентности.

Разнообразие тем оценивается через метрику, обратную Rank-Biased Overlap (RBO), – меру сходства тем, рекурсивно оценивающую долю пересечений между ранжированными списками слов в темах на разных уровнях «глубины» погружения в список, и меру разнообразия топиков TopicDiversity (далее TD) [33], рассчитываемую как отношение числа уникальных слов к произведению количества тем на k, где k – это количество топ-слов в теме, учитываемых в расчете метрики. Предыдущие исследования обнаруживают, что увеличение различности тем повышает интерпретируемость результирующих тем [42], а также принцип различия часто используется во внешней ручной оценке качества тематических моделей с привлечением экспертов, которым предлагается выделить лишнее в теме слово [43].

Для оценки схожести тем используется мера попарного сходства Жаккара (Pairwise Jaccard) – усредненное по количеству сравнений отношение количества общих слов в темах к их совокупному набору слов.

Качество предсказания оценивается при помощи усредненной по классам F1-меры: $2 * \frac{precision * recall}{precision + recall}$. Такой подход к вычислению F1-меры также называется micro F1-мерой. Micro F1 раскла-

относительно слова А предполагает, что мы считаем слово Б соприсутствующим со словом А в рассматриваемом тексте, если оно находится в пределах трех слов слева и трех слов справа от слова А. Контекстные окна используются для ограничения длины контекста, который мы считаем релевантным для каждого слова.

дывает задачу мультиклассовой классификации на несколько задач бинарной классификации, где каждый класс противопоставляется всем прочим сразу, а не по отдельности. Подобная адаптация классической F1-меры менее чувствительна к дисбалансу классов, что актуализирует ее использование с учетом сильного дисбаланса классов в используемых нами данных (см. Приложение 1).

Эмпирическая база исследования

Сравнение вышеописанных алгоритмов тематического моделирования проводится на четырех базах данных, преследующих разные задачи: двух размеченных (массивы новостных публикаций MLSUM и Corus) и двух неразмеченных (массивы комментариев и хештегов урбанистической тематики из русскоязычного сегмента TikTok, собранные автором самостоятельно в ходе подготовки настоящей статьи). Для восполнения ограничений, связанных с неизвестностью истинного количества тематических кластеров в базе данных комментариев в TikTok, мы начинаем с анализа размеченных баз данных. Это позволяет оценить, насколько хорошо работает каждый из избранных алгоритмов в условиях возможности проверки соответствия между созданными кластерами и темами, размеченными в массивах данных, что делает дальнейшие выводы на неразмеченной базе данных более надежными.

Размеченные массивы данных

Для первичного сравнения качества сравниваемых моделей мы используем два размеченных массива данных для задачи классификации текстов.

1. Размеченная темами выборка сжатых пересказов новостных публикаций из базы данных Multi Lingual Summarization (MLSUM)[1] [44]. Изначальная база данных была очищена от строк

[1] База данных MLSUM с разбиением на тренировочную, валидационную и тестовую выборки доступна на платформе Hugging Face по ссылке: https://huggingface.co/datasets/mlsum (дата обращения: 09.02.2024).

длиннее 150 символов, чтобы приблизить схожесть по длине с комментариями в социальных сетях. Были удалены низкозаполненные и несодержательные темы.

2. Размеченная темами база новостных публикаций Lenta.ru v1.1+ из корпуса Corus[1] (далее мы будем отсылаться к этому датасету как Lenta). В выборку настоящего исследования вошли топ-20 тем по количеству появлений. Для приближения объема текста к референтному – характерному для социальных сетей – в текстах сохранялись первые одно-два предложения. Итоговые тексты не превышают по длине 200 символов.

Оба массива содержат краткие изложения новостных статей и используют назначенные им опубликовавшими их изданиями категории в качестве разметки. Все данные, содержащиеся в массивах MLSUM и Corus, были получены в результате веб-скрейпинга онлайн-страниц избранных новостных изданий, подготовка данных не включала разметки экспертами со стороны авторов массива. Так, соответствие назначенной темы тексту новостной публикации зависит от качества соответствующей работы по категоризации новостей в каждом из использованных в массивах данных изданий.

Неразмеченные массивы данных

Для задачи оценки качества на неразмеченном датасете мы используем массив русскоязычных комментариев к урбанистическим видео в TikTok, а также хештеги, присуждаемые этим видео их создателями. Выбор TikTok в качестве платформы для анализа обусловлен устойчивой популярностью сети до блокировки в России, а также адаптацией формата в прочих социальных сетях, до сих пор доступных на территории РФ (YouTube, «ВКонтакте»). Фокус на урбанистической тематике основан на потребности

[1] Описание баз данных и кода для доступа к ним доступны на платформе GitHub по ссылке: https://github.com/natasha/corus (дата обращения: 09.02.2024).

в общей тематической когерентности анализируемых видео, которую можно использовать в качестве базового бенчмарка тематики. Сама по себе тема урбанистики удобна своей популярностью, предоставляющей достаточно объемную базу для исследования, а также способностью вовлекать аудиторию в дискуссию за счет близости темы зрителям, что обеспечивает наличие содержательных обсуждений в комментариях.

Отбор видео осуществлялся по хештегам, список которых расширялся при помощи рекомендательного алгоритма платформы. Финальный набор содержит 17 хештегов. Из массива видео, относящихся к этим хештегам, доступными оказались 2625 видео и содержащими комментарии – 1271. Выгрузка комментариев и хештегов производилась с помощью инструмента Selenium WebDriver для Chrome и языков программирования JavaScript и Python[1].

Предобработка

Перед анализом полученные массивы подвергались базовой предобработке.

Шаг 1. Лемматизация – слова были приведены к нормальной форме (инфинитиву для глаголов и именительному падежу единственного числа мужского рода для существительных и прилагательных).

Шаг 2. Удаление стоп-слов – из нормализованных текстов удалялись слова, входящие в наборы русскоязычных стоп-слов в библиотеках NLTK и spaCy для Python. В этот список входят слова с низкой семантической значимостью (предлоги, союзы, местоимения). Далее были удалены слова и хештеги, встречающиеся редко (менее 10 раз) или часто (более чем в 80% документов), а также

[1] Полный код для скрейпинга комментариев в TikTok, использованный для сбора данных в настоящей работе, представлен на платформе GitHub по ссылке: https://github.com/vavaschenko/TikTokScraper (дата обращения: 05.01.2024).

«отметки» – ники других пользователей, возникающие в тексте комментария, когда его автор обращается к другому пользователю. Для всех корпусов, кроме массива хештегов, в анализ включались не только отдельные слова, но и биграммы, сформированные при помощи метода Phraser библиотеки Gensim для Python.

Таблица 3

ОПИСАНИЕ ИСПОЛЬЗУЕМЫХ ДАННЫХ

Датасет	Количество тем	Количество наблюдений		Размер словаря после обработки
TikTok	-	Комментарии	152 461	7021
		Хештеги	2625	822
MLSUM	11	24 032		6375
Lenta	20	239 151		7617

Сравнение результатов тематического моделирования на коротких текстах

Оценка качества на размеченной базе данных

Эксперименты по тренировке алгоритмов проводились на виртуальной машине Linux 5.15.109 с Intel(R) Xeon(R) Platinum 8259CL CPU @ 2,50GHz и GPU Tesla V100-SXM2-16GB. Для каждого датасета гиперпараметры (необучаемые параметры, задаются перед началом обучения) рассматриваемых моделей оптимизировались при помощи случайного поиска по сетке значений (см. Приложение 2).

Метрики качества (UMass, NPMI, IRBO, TD, Pairwise Jaccard и F1-мера) в таблицах 4 и 5 приводятся для наилучшей модели по NPMI и рассчитываются на топ-10 словах каждой темы. Для модели SBMTM оцениваются темы, результирующие на двух уровнях иерархии тем: L0 и L1 в таблицах 4 и 5 относятся к нулевому и первому уровню иерархии тем для SBMTM соответственно.

Таблица 4

СРАВНЕНИЕ КАЧЕСТВА ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ НА РАЗМЕЧЕННЫХ ДАННЫХ

Модель	Датасет		Время (*мин.*)	Кол-во тем	Когерентность		Разнообразие		Схожесть	F1
					UMass	NPMI	IRBO	TD	Pairwise Jaccard	
BERTopic	MLSUM		3	265	**-0,28**	-0,13	**0,99**	**0,982**	**0,000**	0,285
	Lenta		7	719	*-0,635*	0,005	*0,99*	*0,912*	*0,000*	0,516
ETM (Word2Vec)	MLSUM		1	14	-3,311	-0,016	0,974	0,279	0,029	0,331
	Lenta		5	49	-3,146	0,065	0,845	0,453	0,117	0,539
ETM (Navec)	MLSUM		0,5	38	-8,338	-0,119	0,938	0,532	0,044	0,438
	Lenta		6	35	-3,986	0,059	0,966	0,737	0,021	0,506
CTM	MLSUM		1,5	38	-10,07	-0,179	0,935	0,579	0,052	0,313
	Lenta		9	49	-3,596	*0,164*	0,989	0,745	0,009	0,582
ProdLDA	MLSUM		1,2	66	-8,636	-0,095	0,956	0,468	0,029	0,429
	Lenta		22,5	61	-4,069	0,138	0,995	0,815	0,004	0,54
SBMTM	MLSUM	L0	11	33	-8,293	-0,006	1,0*	1,0*	0,000*	**0,486**
		L1		6	-4,367	0,023	1,0*	1,0*	0,000*	0,376
	Lenta	L0	21	674	-10,35	-0,019	1,0*	1,0*	0,000*	*0,653*
		L1		150	-8,15	-0,008	1,0*	1,0*	0,000*	0,588
NMF	MLSUM		**0,5**	11	-4,723	**0,01**	0,934	0,67	0,058	0,344
	Lenta		26	24	-3,575	0,075	0,96	0,604	0,032	0,265

Примечание. Жирным шрифтом отмечены лучшие результаты для MLSUM, курсивом с подчеркиванием – для Lenta.

* В текущей версии алгоритма SBMTM не реализована возможность пересечения между выделяемыми сообществами. Как следствие, каждое слово может попадать только в один кластер, что приводит к строго разнообразным кластерам.

Для оценки различительной способности сформированных тематических кластеров при помощи сравниваемых алгоритмов мы используем случайный лес для задачи классификации: целевыми значениями становятся значения уникальных классов, а предикторами – предсказанные вероятности тематических кластеров для документов в массиве.

Нейросетевые модели превосходят альтернативы по когерентности и разнообразности тем, однако SBMTM L0 превосходит их по качеству классификации. BERTopic является абсолютным лидером по метрикам качества тем, однако значительно уступает и более простым алгоритмам в качестве классификации вследствие того, что модель выделяет очень локальные семантические группы, принадлежность к которым хуже генерализуется для предсказания класса. На высокую специфичность выделяемых тем указывает, в частности, высокое значение разнообразия тем, несмотря на техническую возможность альтернативного результата (в отличие от SBMTM). В свою очередь SBMTM хорошо приближает истинные классы, однако формирует темы с более низкой когерентностью. Между альтернативными эмбеддингами для ETM различия невелики: модель, обученная с эмбеддингами Navec, несколько превосходит эмбеддинги Word2Vec по разнообразию тем и качеству классификации на MLSUM, однако уступает для Lenta. Некоторой золотой серединой выступают CTM и ProdLDA, следующие за SBMTM и BERTopic по когерентности и F1.

Так, анализ алгоритмов тематического моделирования на размеченных данных позволил обнаружить содержательные различия в результирующих для каждого из инструментов тематических группах: несмотря на то, что BERTopic лучше агрегирует слова-токены, что позволяет достигать более когерентных тем, SBMTM превосходит BERTopic и ETM в задаче группировки документов в соответствии с выделенными темами.

Оценка качества на неразмеченных данных

Вторым этапом в сравнительной оценке методов тематического моделирования становится их приложение к данным с неизвестным количеством тем. Здесь мы оцениваем качество тематического моделирования на двух типах данных: хештегах, сопутствующих видеопубликациям, и комментариях, оставленных пользователями.

Анализ хештегов представляется перспективным направлением проверки качества инструментов тематического моделирования, поскольку хештеги как тип текстовых данных отличаются большей семантической когерентностью, чем свободные высказывания. За счет того, что при помощи использования хештегов пользователи стремятся привязать свое собственное высказывание к более крупному дискурсу и/или тренду, хештеги зачастую близко друг с другом связаны.

Аналогично BERTopic демонстрирует наилучшую связность и разнообразие кластеров, однако в этом случае мы наблюдаем расхождение в том, насколько хорошо модели справляются с разными типами данных.

Примечательна динамика качества тематических кластеров в зависимости от количества слов, на которых производится оценка когерентности: по мере увеличения количества анализируемых слов в кластере для расчета метрики связности значение связности падает. Этот эффект связан с качеством ранжирования слов в темах внутри самих алгоритмов. Вместе с убыванием значимости слова в теме оно должно терять свою релевантность, а значит – способствовать сокращению метрики связности темы. Устойчиво убывающие графики качества (рис. 1) указывают на хорошее качество ранжирования слов внутри тем, достигаемое за счет c-TF-IDF взвешивания внутри кластеров для BERTopic и методов выделения сообществ в бимодальной сети для SBMTM. Волатильность оценок для ETM (Navec), в свою очередь,

Таблица 5

СРАВНЕНИЕ КАЧЕСТВА ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ НА НЕРАЗМЕЧЕННЫХ ДАННЫХ

Модель	Датасет		Время (*мин*)	Кол-во тем	Когерентность		Разнообразие		Схожесть
					UMass	NPMI	IRBO	TD	Pairwise Jaccard
BERTopic	TikTok (комм.)		5	1421	**-1,308**	**0,0**	**0,999**	**0,78**	**0,000**
	TikTok (хешт.)		7	719	*-0,635*	0,005	*0,999*	*0,912*	*0,000*
ETM (Word2Vec)	TikTok (комм.)		3	73	-3,11	0,061	0,036	0,014	1,0
ETM (Navec)			3	38	-6,487	-0,036	0,903	0,421	0,067
CTM	TikTok (комм.)		10	146	-7,803	-0,016	0,871	0,125	0,114
	TikTok (хешт.)		0,1	12	-9,124	*0,092*	0,959	0,733	0,037
ProdLDA	TikTok (комм.)		10	55	-6,237	-0,021	0,856	0,193	0,106
	TikTok (хешт.)		0,08	18	-7,117	-0,004	0,77	0,333	0,179
LDA	TikTok (хешт.)		0,05	70	-2,992	0,054	0,031	0,016	0,914
SBMTM	TikTok (комм.)	L0	17	11	-7,66	-0,05	1,0*	1,0*	0,000*
		L1		2	-8,086	0,006	1,0*	1,0*	0,000*
	TikTok (хешт.)	L0	1	44	-11,24	0,042	1,0*	1,0*	0,000*
		L1		10	-10,81	-0,029	1,0*	1,0*	0,000*
NMF	TikTok (комм.)		14	47	-4,799	0,043	0,93	0,364	0,048
	TikTok (хешт.)		0,25	46	-7,949	0,052	0,881	0,315	0,091

Примечание. Жирным шрифтом отмечены лучшие результаты для MLSUM, курсивом с подчеркиванием – для Lenta.

* В текущей версии алгоритма SBMTM не реализована возможность пересечения между выделяемыми сообществами. Как следствие, каждое слово может попадать только в один кластер, что приводит к строго разнообразным кластерам.

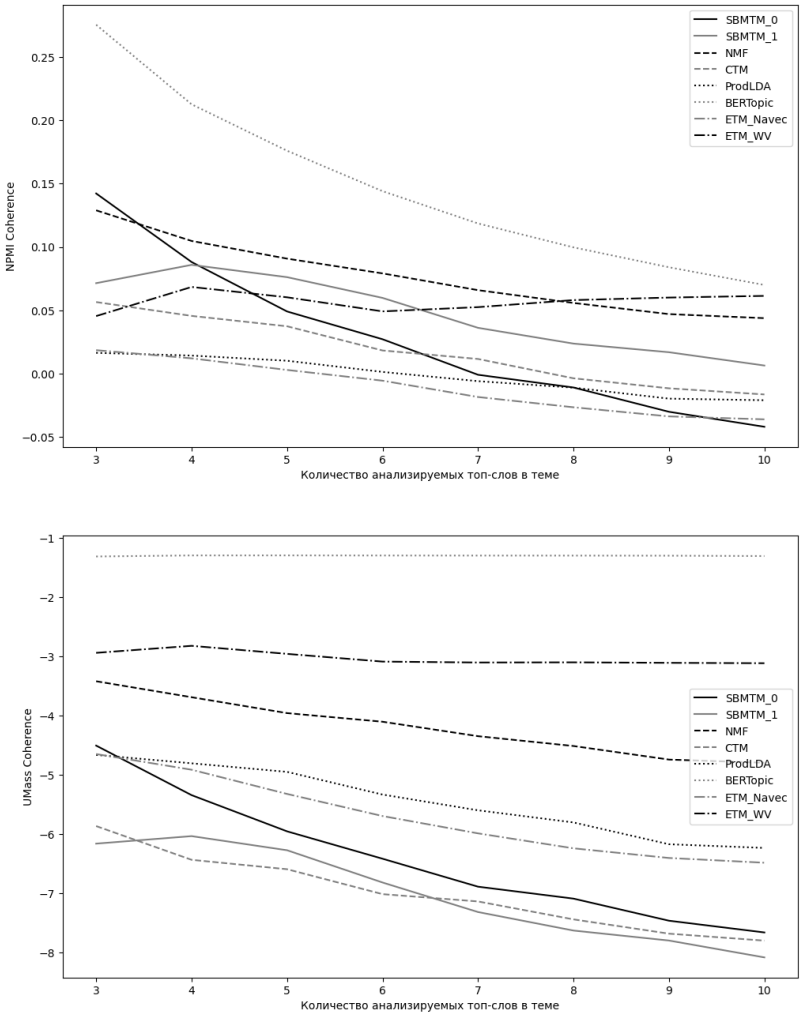

Рис. 1. Динамика метрик качества тем для комментариев в зависимости от количества включаемых в анализ слов в теме

демонстрирует худшее качество ранжирования важности слов в тематических кластерах. Тем не менее наблюдаемая динамика может частично объяснять превосходство прочих моделей перед BERTopic в качестве классификации на размеченных данных: модель отлично выбирает наиболее значимые слова в теме, однако далее по списку значимости качество тем резко падает, в то время как CTM и ProdLDA сохраняют близкие значения когерентности при увеличении числа анализируемых топ-слов.

Следует обратить внимание на то, что между моделями различается и разброс значений когерентности (рис. 2): если BERTopic склонен создавать темы примерно одного и того же среднего уровня качества, то SBMTM производит больше как очень низко-, так и очень высококогерентных тем.

ETM (Word2Vec) демонстрирует очень высокую волатильность качества в зависимости от числа анализируемых слов. В меньшей степени это характерно для ETM (Navec) и NMF. Наибольшую стабильность демонстрируют CTM и ProdLDA – анализ большего числа слов незначительно меняет распределение когерентности тем, указывая на хорошее и стабильное качество ранжирования слов между темами.

Ограничения

Среди ограничений настоящего исследования стоит обратить внимание на несколько деталей, связанных с ограничениями вычислительных мощностей, доступных автору при проведении вышеописанных экспериментов. Во-первых, ввиду ограничений доступных объемов оперативной памяти CPU и GPU (50 и 16 Гб соответственно) для экспериментальной оценки качества алгоритмов тематического моделирования были выбраны сравнительно небольшие массивы данных. Во-вторых, используемый массив неразмеченных коротких текстов из TikTok по той же причине ограничен тематически, однако дополнительным ограничением

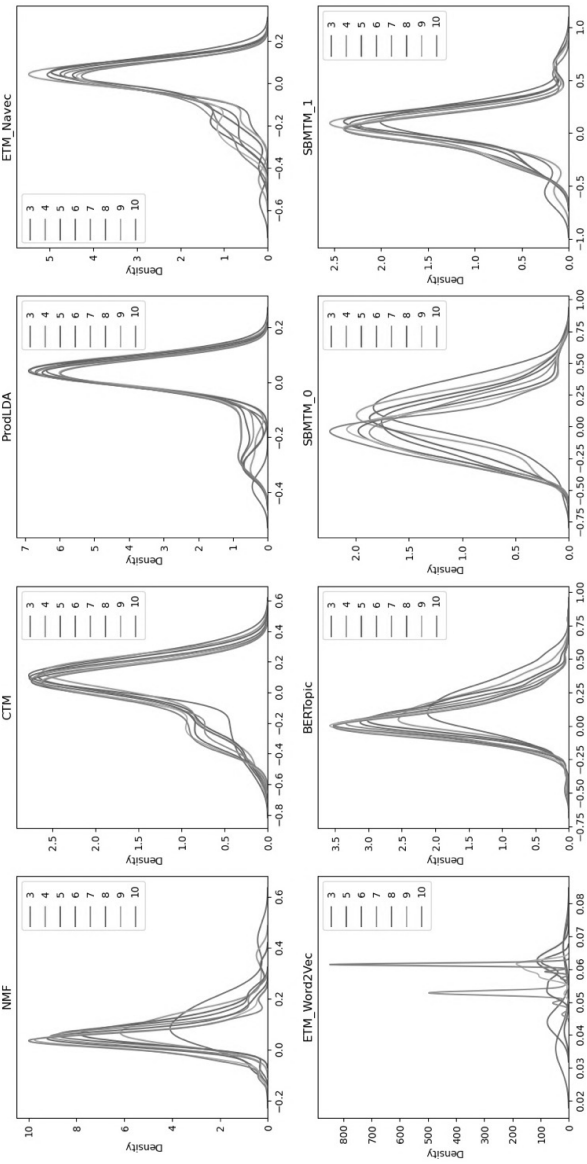

Рис. 2. Распределение метрик качества тем для комментариев в зависимости от количества включаемых в анализ слов в теме

доступа к данным комментариев в TikTok являются внутренние лимиты на автоматизированный сбор данных в социальной сети: в целях противодействия веб-скрейпингу TikTok ограничивает объем выдачи как видео, так и комментариев к ним по заданному запросу. Другими словами, в собранном массиве данных присутствуют не все видео урбанистической тематики, а также хоть и абсолютное большинство, но не все комментарии к полученным видео. Наконец, ограничения доступной памяти и недоступность параллельных вычислений в нашем случае не позволяют расширить сетку значений гиперпараметров, рассматриваемую при оптимизации выдачи каждого из анализируемых алгоритмов, что позволяет предположить, что избранное решение с наилучшим результатом для каждой из моделей не является абсолютным максимумом ее возможного качества.

В качестве методологического ограничения проделанной работы можно выделить разницу в источнике сравниваемых размеченных и неразмеченных массивов данных. Оба размеченных массива содержат тексты новостных публикаций, в то время как неразмеченные данные представляют комментарии пользователей в социальной сети TikTok. Поскольку новостные публикации зачастую имеют тематическую категоризацию на сайтах изданий, они представляют собой крупный и доступный объем данных для тематического моделирования, и многие базы данных для тематического моделирования опираются именно на них. Выбор урбанистической тематики в качестве фильтра позволяет ожидать пересечения в темах обсуждения в комментариях в TikTok с новостными публикациями, освещающими проблемы общества, вопросы транспорта, и территориальными категориями «Москва» и «Московская область» в массиве MLSUM. Тем не менее тексты комментариев могут отличаться лексически, содержать больше сленговых выражений, ненормативной лексики, именованных сущностей, что может затруднять работу моделей, основанных на предобученных словарных эмбеддингах. Так, сравниваемые

массивы различаются не только наличием разметки, но и структурой данных.

Наконец, в рамках настоящего исследования оптимальное число тем подбиралось по сетке значений вместе с прочими гиперпараметрами для моделей, требующих указания числа тем в качестве гиперпараметра, на основании увеличения метрики когерентности. Использование автоматических методов оптимизации тематических моделей, таких как энтропийные тематические модели [19], и регуляризаторов для итеративного сокращения числа тем [20] может помочь быстрее находить оптимальное количество тем и улучшить результаты тематического моделирования.

Заключение

В рамках настоящей работы представлен сравнительный анализ качества неконвенциональных алгоритмов тематического моделирования на четырех разных корпусах коротких текстов. В сравнение вошли методы, дополняющие традиционно используемый LDA семантической информацией при помощи предобученных статических (ETM) и контекстуальных (CTM) эмбеддингов, расширяющие формальное определение LDA за счет внедрения иерархического подхода к моделированию тем в сетевом представлении документов и слов (SBMTM), или product-of-experts структуры (ProdLDA), а также реформирующие задачу тематического моделирования как задачу кластеризации на векторных представлениях слов и документов, произведенных на выходе кодировщика BERT-модели (BERTopic) или матричного разложения (NMF). Избранные алгоритмы сравнивались в задачах классификации и тематического моделирования, для этого были использованы наборы размеченных по темам новостных публикаций и неразмеченных хештегов и комментариев к урбанистическим видео в TikTok.

Анализ качества тематического моделирования по метрикам когерентности и разнообразия тем для всех четырех корпусов

текстов указывает на превосходство BERTopic в задаче создания когерентных тем. Это достигается за счет c-TF-IDF – алгоритма взвешивания слов по их значимости внутри темы, прямые аналоги которого отсутствуют у альтернатив. Однако мы обнаруживаем, что, несмотря на то что темы, выделяемые BERTopic, имеют более высокие показатели когерентности в среднем по модели, SBMTM на нулевом уровне иерархии формирует больше тем с высокими значениями когерентности, которые при усреднении уравновешиваются темами с низкими метриками качества. Вероятно, эта характеристика позволяет SBMTM значительно превосходить BERTopic в задаче классификации документов. Совместное моделирование сообществ для документов и слов в бимодальной сети позволяет SBMTM лучше кластеризировать документы, однако качество предсказания также указывает на то, что темы, производимые SBMTM, лучше подлежат генерализации. Иерархическая структура алгоритма позволяет эффективно привнести гетерогенность в набор производимых тем, в то время как BERTopic склонен фокусироваться исключительно на локальных семантических паттернах.

Проведенные эксперименты подтверждают перспективность использования альтернативных LDA-методов тематического моделирования на коротких текстах и, в частности, сетевого подхода к представлению текстовых данных в задаче тематического моделирования. Мы обнаруживаем, что метод иерархического блокмоделирования превосходит методы, основанные на словарных эмбеддингах в задаче классификации, ранее продемонстрировавших наиболее высокое среди сродных альтернатив качество для классификации документов на англоязычных датасетах [15]PLSA and LDA. Тем не менее следует отметить, что SBMTM уступает альтернативам по производительности, а также качеству ранжирования слов в темах по значимости, на что указывает высокий разброс значений и заметные различия распределения когерентности по темам – в зависимости от числа анализируемых слов.

На трех задействованных в анализе базах текстов: MLSUM, Lenta v1.1+ и TikTok наиболее стабильное качество демонстрируют CTM и ProdLDA. Несмотря на то, что во всех экспериментах по анализу качества тем лидирует BERTopic, высокое значение когерентности в этой модели нестабильно и резко падает с увеличением числа анализируемых топ-слов в теме; BERTopic также хуже справляется с классификацией документов на основании произведенных моделью тем. Лидер классификации, SBMTM, в свою очередь отличается нестабильностью качества тем. Мы отмечаем потенциал применения методов сетевого анализа для тематического моделирования на коротких текстах, однако заключаем, что методы, основанные на создании более сложных внутренних представлений текста в модели, демонстрируют более высокое качество за счет лучшей репрезентации близости между текстами и результирующими темами.

Наконец, в то время как разнообразия сравниваемых размеченных баз данных недостаточно для формулировки формализованных рекомендаций к выбору методов тематического моделирования на основе характеристик доступных данных, представляется возможным обобщение наблюдений о предпочтительности той или иной модели в зависимости от исследовательской задачи. Так, если тематическое моделирование в рамках исследования предусмотрено с целью различения документов, то следует рассмотреть модели, группирующие документы по тематической близости (в рамках настоящего исследования это SBMTM). Для данных с предположительно сложным тематическим составом (особенно если предполагается, что в данных могут присутствовать малые по частоте встречаемости, но значимые для задачи исследования темы) могут быть очень полезны инструменты тематического моделирования с использованием контекстуальных эмбеддингов (здесь CTM, BERTopic). Однако мы подчеркиваем преимущество CTM с точки зрения интерпретации за счет выделения меньшего количества тем, что делает их доступными для ручной проверки

на связность. Тем не менее многие альтернативы LDA могут быть затратны с точки зрения вычислительных ресурсов, поскольку имеют меньше альтернативных эффективных реализаций и не всегда доступны для параллельных вычислений: в случае, если предполагается сравнительная тематическая однородность, не ожидается значимых различий в значении слов в зависимости от контекста и порядок слов в тексте не является принципиальным, LDA может оказаться предпочтительнее моделей, обращающихся к контексту.

ЛИТЕРАТУРА

1. *Brookes G., McEnery T.* The utility of topic modelling for discourse studies: A critical evaluation // Discourse Studies. 2019. Vol. 21, № 1. С. 3–21. DOI: 10.1177/1461445618814032.

2. Using topic models for Twitter hashtag recommendation / F. Godin, V. Slavkovikj, W. De Neve [et al.] // Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil: ACM, 2013. P. 593–596. DOI: 10.1145/2487788.2488002.

3. *Asmussen C.B., Møller C.* Smart literature review: a practical topic modelling approach to exploratory literature review // Journal of Big Data. 2019. Vol. 6, № 1. P. 93. DOI: 10.1186/s40537-019-0255-7. EDN: XBRIWK.

4. On the Globalization of the QAnon Conspiracy Theory Through Telegram / M. Hoseini, P. Melo, F. Benevenuto [et al.] // Proceedings of the 15th ACM Web Science Conference 2023. Austin TX, USA: ACM, 2023. P. 75–85. DOI: 10.1145/3578503.3583603.

5. *Кольцова О.Ю., Маслинский К.А.* Выявление тематической структуры российской блогосферы: автоматические методы анализа текстов // Социология: методология, методы, математическое моделирование (Социология: 4М). 2013. № 36. С. 113–139. EDN: RCFOWJ.

6. *Lyu J.C., Han E.L., Luli G.K.* COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis // Journal of Medical Internet Research. 2021. Vol. 23, № 6. P. e24435. DOI: 10.2196/24435.

7. ET-LDA: Joint topic modeling for aligning, analyzing and sensemaking of public events and their Twitter feeds / Y. Hu, A. John, F. Wang [et al.] // Cornwall University [site]. 08.10.2012. URL: https://arxiv.org/abs/1210.2164 (дата обращения: 01.09.2023).

8. Multi-modal event topic model for social event analysis / S. Qian, T. Zhang, C. Xu, J. Shao // IEEE Transactions on Multimedia. 2016. Vol. 18, № 2. P. 233–246. DOI: 10.1109/TMM.2015.2510329.

9. *Zheng Y., Zhang Y.-J., Larochelle H.* Topic Modeling of Multimodal Data: An Autoregressive Approach // 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. P. 1370–1377. DOI: 10.1109/CVPR.2014.178.

10. *Gong Y., Poellabauer C.* Topic Modeling Based Multi-modal Depression Detection // Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. Mountain View, California, USA: ACM, 2017. P. 69–76. DOI: 10.1145/3133944.3133945.

11. *Бызов А.А.* Интеллектуальный анализ текстов в социальных науках // Социология: методология, методы, математическое моделирование (Социология: 4М).2019. № 49. С. 131–160. EDN: GCIIVL.

12. *Boon-Itt S., Skunkan Y.* Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study // JMIR Public Health and Surveillance. 2020. Vol. 6, № 4. P. e21978. DOI: 10.2196/21978.

13. *Albalawi R., Yeap T.H., Benyoucef M.* Using topic modeling methods for short-text data: A comparative analysis // Frontiers in artificial intelligence. 2020. Vol. 3. P. 42. DOI: 10.3389/frai.2020.00042.

14. *Hong L., Davison B.D.* Empirical study of topic modeling in Twitter // Proceedings of the First Workshop on Social Media Analytics. Washington, D.C.: ACM, 2010. P. 80–88. DOI: 10.3390/ijerph18126487.

15. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey / Q. Jipeng, Q. Zhenyu, L. Yun [et al.] // IEEE Trans. Knowl. Data Eng. 2022. Vol. 34, № 3. P. 1427–1445. DOI: 10.1109/TKDE.2020.2992485. EDN: ACFCRC.

16. Медиапотребление 2023 // Mediascope [сайт]. [2023]. URL: https://mediascope.net/upload/iblock/226/e7lwh96qizxpwhf1rj2ttfzkwlie8vr8/медиапотребление%202023.pdf (дата обращения: 09.02.2024).

17. *Hofmann T.* Probabilistic latent semantic analysis // Cornwall University [site]. 22.01.2013. URL: https://arxiv.org/abs/1301.6705 (дата обращения: 01.09.2023).

18. *Blei D.M., Ng A.Y., Jordan M.I.* Latent dirichlet allocation // Journal of machine learning research. 2003. Vol. 3. P. 993–1022.

19. *Кольцов С.Н.* Применение энтропийного подхода к проблеме выбора числа тем в тематических моделях // Социофизика и социоинженерия'2018: труды второй Всероссийской междисциплинарной конференции. Москва, 23–25 мая 2018 г. М.: Ин-т проблем управления им. В.А. Трапезникова РАН, 2018. С. 235–236. DOI: 10.21883/PJTF.2017.12.44713.16725. EDN: XYERBR.

20. *Потапенко А.А.* Семантические векторные представления текста на основе вероятностного тематического моделирования: дис. … канд. физ.-мат. наук / НИУ ВШЭ. М., 2017. 147 с. EDN: DNXEFS.

21. Digital Trace Data in the Study of Public Opinion: An Indicator of Attention Toward Politics Rather Than Political Support / A. Jungherr, H. Schoen, O. Posegga, P. Jürgens // Social Science Computer Review. 2017. Vol. 35, № 3. P. 336–356. DOI: 10.1177/0894439316631043.

22. *Ahuja A., Wei W., Carley K.M.* Topic modeling in large scale social network data // SSRN electronic journal. January 2015. DOI: 10.2139/ssrn.2720333.

23. Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs / Y. Wang, J. Liu, Y. Huang, X. Feng // IEEE Transactions on Knowledge and Data Engineering. 2016. Vol. 28, № 7. P. 1919–1933. DOI: 10.1109/TKDE.2016.2531661.

24. The author-topic model for authors and documents / M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth // Cornwall University [site]. 11.01.2012. URL: https://arxiv.org/abs/1207.4169 (дата обращения: 01.09.2023).

25. *Phan X.-H., Nguyen L.-M., Horiguchi S.* Learning to classify short and sparse text & web with hidden topics from large-scale data collections // Proceedings of the 17th international conference on World Wide Web. Beijing, China: ACM, 2008. P. 91–100. DOI: 10.1145/1367497.1367510.

26. *Gerlach M., Peixoto T.P., Altmann E.G.* A network approach to topic models // Sci. Adv. 2018. Vol. 4, № 7. P. eaaq1360. DOI: 10.1126/sciadv.aaq1360.

27. Mixed Membership Stochastic Blockmodels / E.M. Airoldi, D. Blei, S. Fienberg, E. Xing // Advances in Neural Information Processing Systems. Vancouver, British Columbia, Canada: Curran Associates Inc., 2008. P. 33–40.

28. *Коршунов А., Гомзин А.* Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН. 2012. Т. 23. С. 215–244. DOI: 10.15514/ISPRAS-2012-23-13. EDN: PLUXDR.

29. *Grootendorst M.* BERTopic: Neural topic modeling with a class-based TF-IDF procedure // Cornwall University [site]. 11.03.2022. URL: https://arxiv.org/abs/2203.05794 (дата обращения: 01.09.2023).

30. Attention is All you Need / A. Vaswani, N. Shazeer, N. Parmar [et al.] // Advances in Neural Information Processing Systems. Long Beach, CA, USA: Curran Associates Inc., 2017. P. 5998–6008.

31. Topic modeling algorithms and applications: A survey / A. Abdelrazek, Y. Eid, E. Gawish [et al.] // Information Systems. 2022. Vol. 112. P. 102131. DOI: 10.1016/j.is.2022.102131. EDN: WLYLKR.

32. *Lee D., Seung H.S.* Algorithms for Non-negative Matrix Factorization // Advances in Neural Information Processing Systems. Denver, CO, USA: MIT Press, 2000. P. 556–562.

33. *Dieng A.B., Ruiz F.J.R., Blei D.M.* Topic Modeling in Embedding Spaces // Transactions of the Association for Computational Linguistics. 2020. Vol. 8. P. 439–453. DOI: 10.1162/tacl_a_00325.

34. *Srivastava A., Sutton C.* Autoencoding Variational Inference for Topic Models // Cornwall University [site]. 04.03.2017. URL: https://arxiv.org/abs/1703.01488 (дата обращения: 01.09.2023).

35. Cross-lingual Contextualized Topic Models with Zero-shot Learning / F. Bianchi, S. Terragni, D. Hovy [et al.] // Proceedings of the 16th Conference of the

European Chapter of the Association for Computational Linguistics. April 19–23, 2021 / Ed. by P. Merlo, J. Tiedemann, R. Tsarfaty. Potsdam, Germany: Association for Computational Linguistics, 2021. P. 1676–1683. DOI: 10.18653/v1/2021.eacl-main.143.

36. *Кукушкин А.* Navec – компактные эмбеддинги для русского языка // Проект Natasha – набор Python-библиотек для обработки текстов на естественном русском языке [сайт]. 2022. URL: https://natasha.github.io/navec/ (дата обращения: 05.01.2024).

37. Efficient estimation of word representations in vector space / T. Mikolov, K. Chen, G. Corrado, J. Dean // Cornwall University [site]. 16.01.2013. URL: https://arxiv.org/abs/1301.3781 (дата обращения: 01.09.2023).

38. Distributed representations of words and phrases and their compositionality / T. Mikolov, I. Sutskever, K. Chen [et al.] // Advances in Neural Information Processing Systems. 2013. Vol. 26. P. 3111–3119.

39. *Pennington J., Socher R., Manning C.D.* Glove: Global vectors for word representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014. P. 1532–1543. DOI: 10.3115/v1/D14-1162.

40. *Aletras N., Stevenson M.* Evaluating topic coherence using distributional semantics // Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers. Potsdam, Germany: Association for Computational Linguistics, 2013. P. 13–22.

41. Optimizing Semantic Coherence in Topic Models / D. Mimno, H.M. Wallach, E. Talley [et al.] // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP). Edinburgh, Scotland, UK: Association for Computational Linguistics, 2011. P. 262–272.

42. *Tan Y., Ou Z.* Topic-weak-correlated Latent Dirichlet allocation // 2010 7th International Symposium on Chinese Spoken Language Processing. Tainan, Taiwan: IEEE, 2010. P. 224–228. DOI: 10.1109/ISCSLP.2010.5684906.

43 *Newman D., Karimi S., Cavedon L.* External Evaluation of Topic Models // ADCS 2009 – Proceedings of the Fourteenth Australasian Document Computing Symposium. Sydney, Australia: University of Sydney, 2011. P. 1–8.

44. MLSUM: The Multilingual Summarization Corpus / T. Scialom, P.-A. Dray, S. Lamprier [et al.] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [s. l.]: Association for Computational Linguistics, 2020. P. 8051–8067. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.647.

Приложение 1

РАСПРЕДЕЛЕНИЕ КЛАССОВ В РАЗМЕЧЕННЫХ МАССИВАХ ТЕКСТОВ

РАСПРЕДЕЛЕНИЕ ТЕМ В ВЫБОРКЕ ИЗ MLSUM

Тема	Количество	Тема	Количество
Общество	7168	Москва	1528
Политика	5310	Экономика	1514
Спорт	3073	Московская область	604
Культура	2623	Наука	506
Происшествия	1634	Авто	72

РАСПРЕДЕЛЕНИЕ ТЕМ В ВЫБОРКЕ ИЗ Lenta.ru v1.1+

Тема	Количество	Тема	Количество
Политика	36 093	Музыка	7054
Общество	31 963	Наука	6693
Украина	19 920	Люди	6250
Происшествия	19 212	Квартира	4656
Футбол	14 301	Преступность	4652
Госэкономика	14 277	ТВ и радио	3945
Кино	10 117	Космос	3686
Интернет	8516	События	3414
Бизнес	8018	Конфликты	3380
Следствие и суд	7784	Соцсети	3314

Приложение 2

ОПИСАНИЕ ГИПЕРПАРАМЕТРОВ, ПОДБИРАЕМЫХ В РАМКАХ ОПТИМИЗАЦИИ АЛГОРИТМОВ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Модель	Гиперпараметр	Влияние на результат	Рассматриваемые значения
NMF	Количество тем	Задает количество тем, которое будет выделять алгоритм	{10, 20, … 200}
	Темп обучения (learning rate) – определяет размер шага оптимизатора при обучении модели	Высокий темп обучения способствует ускорению обучения, однако может мешать определению глобального минимума функции потерь	{0.005, 0.01, ... 0.03}
LDA	Количество тем	См. выше	{10, 20, … 200}
ETM	Количество тем	См. выше	{10, 20, … 200}
	ρ (ро)	Контролирует разреженность эмбеддингов сокращенной размерности в модели. Чем выше это значение, тем более разрежены вектора, отображающие отношения документов, и тем меньше тем выделяется в каждом документе	{0.1, 0.2, 0.5, 1.0}

Продолжение прилож. 2

Модель	Гиперпараметр	Влияние на результат	Рассматриваемые значения
ETM	Dropout rate – доля нейронов, случайным образом удаляемых из заданного слоя модели в ходе обучения	Значения dropout rate отображают силу регуляризации в модели: чем выше dropout rate, тем сильнее противодействие переобучению при тренировке модели	{0, 0.1, … 0.6}
	Темп обучения (learning rate)	См. выше	{0.005, 0.01.. 0.03}
ProdLDA и CTM	Количество тем	См. выше	{10, 20, … 200}
	Функция активации – функция, преобразующая выходы полносвязных слоев в нейросетевой архитектуре	Функции активации добавляют нелинейность в модель. От выбора функции активации зависят качество и эффективность обучения нейросетевой модели	{ReLU, LeakyReLU, Softplus}
	Dropout rate	См. выше	{0, 0.1, … 0.6}
	Количество слоев – отображает глубину обучаемой архитектуры VAE	Более глубокие архитектуры способны улавливать более сложные связи в массиве обучающих данных, однако более склонны к переобучению	{1, 2, 3, 4}

Окончание прилож. 2

Модель	Гиперпараметр	Влияние на результат	Рассматриваемые значения
BERTopic	Минимальный размер темы – минимально допустимый размер кластера при кластеризации эмбеддингов сокращенной размерности	Более низкие значения позволяют лучше улавливать малые темы, однако увеличивают время вычисления	{10, 20, 30, 40, 50}
	Размерность сокращенного пространства эмбеддингов перед кластеризацией	Чем больше измерений используется при сокращении размерности эмбеддингов, тем лучше сохраняется структура оригинальных данных, однако это увеличивает время вычисления	{2, 3, 4, 5}
	Количество «соседей» – количество ближайших наблюдений, задействуемых при вычислении эмбеддингов сокращенной размерности	Более высокие значения количества соседей стимулируют модель сокращения размерности отображать более глобальную структуру данных, более низкие – локальную	{10, 15, 20, 25, 30}

Vashchenko Vasilisa A.,
Research Assistant at the International Laboratory for Applied Network Research, HSE University, Moscow, Russia, vvashchenko@hse.ru

Topic modeling for short texts: comparative analysis of algorithms

The steady increase in the popularity of social media as a means of communication actualizes methodological issues related to processing of short texts with less semantic context than large corpora, which are widely used for training and testing machine learning models for textual data. Topic modeling, an unsupervised machine learning technique aimed at aggregating texts into topic clusters, has many academic and practical applications where information on true groupings of texts is not available. However, the performance of topic modeling algorithms may be limited by requirement of a sufficient semantic context for a high-quality numerical representation of a unit of text, which may not be derived effectively from a short document. This paper is dedicated to discussing 6 different approaches to topic modeling, comparing their performance on a set of Russian-language comments on TikTok and formally evaluating their performance based on speed and coherence of the resulting topics.
Keywords: topic modeling, analysis of textual data, blockmodeling, applied network analysis, social media analysis, transformer models

References

1. Brookes G., McEnery T. The utility of topic modelling for discourse studies: A critical evaluation, *Discourse Studies*, 2019, vol. 21, no. 1, p. 3–21. DOI: 10.1177/1461445618814032.
2. Godin F., Slavkoviki V., De Neve W. et al. Using topic models for Twitter hashtag recommendation, *Proceedings of the 22nd International Conference on World Wide Web*. ACM, Rio de Janeiro, 2013, p. 593-596. DOI: 10.1145/2487788.2488002.
3. Asmussen C.B., Møller C. Smart literature review: a practical topic modelling approach to exploratory literature review, *Journal of Big Data*, 2019, vol. 6, no 1, p. 93. DOI: 10.1186/s40537-019-0255-7.
4. Hoseini M., Melo P., Benevenuto F. et al. On the Globalization of the QAnon Conspiracy Theory Through Telegram, *Proceedings of the 15th ACM Web Science Conference*. ACM: Austin, 2023, p. 75-85. DOI: 10.1145/3578503.3583603.

5. Koltsova O., Maslinsky K. Revealing the thematic structure of the Russian blogosphere: automatic methods of text analysis (in Russian), *Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling)*, 2013, no. 36, p. 113-139.
6. Lyu J.C., Han E.L., Luli G.K. COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis, *Journal of Medical Internet Research*, 2021, vol. 23, no. 6, p. e24435. DOI: 10.2196/24435.
7. Hu Y., John A., Wang F., et al. ET-LDA: Joint topic modeling for aligning, analyzing and sensemaking of public events and their Twitter feeds, *Cornwall University* [site], 08.10.2012. URL: https://arxiv.org/abs/1210.2164 (date of access: 01.09.2023).
8. Qian S., Zhang T., Xu C., Shao J. Multi-modal event topic model for social event analysis, *IEEE Transactions on Multimedia*, 2016, vol. 18, no. 2, p. 233–246. DOI: 10.1109/TMM.2015.2510329.
9. Zheng Y., Zhang Y.-J., Larochelle H. "Topic Modeling of Multimodal Data: An Autoregressive Approach", in: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA: IEEE, 2014, p. 1370–1377. DOI: 10.1109/CVPR.2014.178.
10. Gong Y., Poellabauer C. "Topic Modeling Based Multi-modal Depression Detection", in: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. Mountain View, California, USA: ACM, 2017, p. 69–76. DOI: 10.1145/3133944.3133945.
11. Byzov A. Text mining in social sciences (in Russian), *Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling)*, 2019, no. 49, p. 131-160.
12. Boon-Itt S., Skunkan Y. Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study, *JMIR Public Health and Surveillance*, 2020, vol. 6, no. 4, p. e21978. DOI: 10.2196/21978.
13. Albalawi R., Yeap T.H., Benyoucef M. Using topic modeling methods for short-text data: A comparative analysis, *Frontiers in artificial intelligence*, 2020, vol. 3, p. 42. DOI: 10.3389/frai.2020.00042.
14. Hong L., Davison B. D. "Empirical study of topic modeling in Twitter", in: *Proceedings of the First Workshop on Social Media Analytics*. Washington, D.C.: ACM, 2010, p. 80–88. DOI: 10.3390/ijerph18126487.
15. Jipeng Q., Zhenyu Q., Yun L., et al. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey, *IEEE Trans. Knowl. Data Eng*. 2022, vol. 34, no. 3, p. 1427–1445. DOI: 10.1109/TKDE.2020.2992485.

16. Mediaconsumption 2023 (in Russian), *Mediascope* [site], 2023. URL: https://mediascope.net/upload/iblock/226/e7lwh96qizxpwhf1rj2ttfzkwlie8vr8/медиапотребление%202023.pdf (date of access: 09.02.2024).
17. Hofmann T. Probabilistic latent semantic analysis, *Cornwall University* [site], 22.01.2013. URL: https://arxiv.org/abs/1301.6705 (date of access: 01.09.2023).
18. Blei D.M., Ng A.Y., Jordan M.I. Latent dirichlet allocation, *Journal of machine learning research*, 2003, vol. 3, p. 993–1022.
19. Koltsov S. "Applying the entropy approach to the problem of choosing the number of topics in topic models", in: *Sociophysics and Socioengineering 2018: Proceedings of the Second All-Russian Cross-disciplinary Conference*. Moscow: V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences, 2018, p. 235–236. DOI: 10.21883/PJTF.2017.12.44713.16725.
20. Potapenko A. *Sematic vector embeddings of text based on probabilistic topic modeling* (in Russian) [Doct. Diss.]. Moscow: Higher School of Economics, 2017, 147 p.
21. Jungherr A., Schoen H., Posegga O., Jürgens P. Digital Trace Data in the Study of Public Opinion: An Indicator of Attention Toward Politics Rather Than Political Support, *Social Science Computer Review*, 2017, vol. 35, no. 3. p. 336–356. DOI: 10.1177/0894439316631043.
22. Ahuja A., Wei W., Carley K.M. Topic modeling in large scale social network data, *SSRN electronic journal*, January 2015. DOI: 10.2139/ssrn.2720333.
23. Wang Y., Liu J., Huang Y., Feng X. Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs, *IEEE Transactions on Knowledge and Data Engineering*, 2016, vol. 28, no. 7, p. 1919–1933. DOI: 10.1109/TKDE.2016.2531661.
24. Rosen-Zvi M., Griffiths T., Steyvers M., Smyth P. The author-topic model for authors and documents, *Cornwall University* [site], 11.01.2012. URL: https://arxiv.org/abs/1207.4169 (date of access: 01.09.2023).
25. Phan X.-H., Nguyen L.-M., Horiguchi S. "Learning to classify short and sparse text & web with hidden topics from large-scale data collections", in: *Proceedings of the 17th international conference on World Wide Web*. Beijing, China: ACM, 2008, p. 91–100. DOI: 10.1145/1367497.1367510.
26. Gerlach M., Peixoto T.P., Altmann E.G. A network approach to topic models, *Sci. Adv*, 2018, vol. 4, no. 7, p. eaaq1360. DOI: 10.1126/sciadv.aaq1360.

27. Airoldi E.M., Blei D., Fienberg S., Xing E. “Mixed Membership Stochastic Blockmodels”, in: *Advances in Neural Information Processing Systems*. Vancouver, British Columbia, Canada: Curran Associates Inc., 2008, p. 33–40.
28. Korshunov A., Gomzin A. Topic modelling for natural language texts (in Russian), *Proceedings of the Institute for System Programming of the Russian Academy of Sciences*, 2012, vol. 34, p. 215-244. DOI: 10.15514/ISPRAS-2012-23-13.
29. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure, *Cornwall University* [site], 11.03.2022. URL: https://arxiv.org/abs/2203.05794 (date of access: 01.09.2023).
30. Vaswani A., Shazeer N., Parmar N., et al. “Attention is All you Need”, in: *Advances in Neural Information Processing Systems*. Long Beach, CA, USA: Curran Associates Inc., 2017, p. 5998–6008.
31. Abdelrazek A., Eid Y., Gawish E., et al. Topic modeling algorithms and applications: A survey, *Information Systems*, 2022, vol. 112. p. 102131. DOI: 10.1016/j.is.2022.102131.
32. Lee D., Seung H.S. “Algorithms for Non-negative Matrix Factorization”, in: *Advances in Neural Information Processing Systems*. Denver, CO, USA: MIT Press, 2000, p. 556–562.
33. Dieng A.B., Ruiz F.J.R., Blei D.M. Topic Modeling in Embedding Spaces, *Transactions of the Association for Computational Linguistics*, 2020, vol. 8, p. 439–453. DOI: 10.1162/tacl_a_00325.
34. Srivastava A., Sutton C. Autoencoding Variational Inference for Topic Models, *Cornwall University* [site], 04.03.2017. URL: https://arxiv.org/abs/1703.01488 (date of access: 01.09.2023).
35. Bianchi F., Terragni S., Hovy D., et al. “Cross-lingual Contextualized Topic Models with Zero-shot Learning”, in: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, ed. by P. Merlo, J. Tiedemann, R. Tsarfaty. Potsdam, Germany: Association for Computational Linguistics, 2021, p. 1676–1683. DOI: 10.18653/v1/2021.eacl-main.143.
36. Kukushkin A. Navec – compact embeddings for the Russian language (in Russian), *Project Natasha – an array of Python libraries for text processing in natural Russian language* (in Russian) [site], 2022. URL: https://natasha.github.io/navec/ (date of access: 05.01.2024).

37. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space, *Cornwall University* [site], 16.01.2013. URL: https://arxiv.org/abs/1301.3781 (date of access: 01.09.2023).
38. Mikolov T., Sutskever I., Chen K, et al. Distributed representations of words and phrases and their compositionality, *Advances in Neural Information Processing Systems*, 2013, vol. 26, p. 3111–3119.
39. Pennington J., Socher R., Manning C.D. "Glove: Global vectors for word representation", in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, p. 1532–1543. DOI: 10.3115/v1/D14-1162.
40. Aletras N., Stevenson M. "Evaluating topic coherence using distributional semantics", in: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*. Potsdam, Germany: Association for Computational Linguistics, 2013, p. 13–22.
41. Mimno D., Wallach H.M., Talley E., et al. "Optimizing Semantic Coherence in Topic Models", in: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Edinburgh, Scotland, UK: Association for Computational Linguistics, 2011, p. 262–272.
42. Tan Y., Ou Z. "Topic-weak-correlated Latent Dirichlet allocation", in: *2010 7th International Symposium on Chinese Spoken Language Processing*. Tainan: IEEE, 2010, p. 224–228. DOI: 10.1109/ISCSLP.2010.5684906.
43. Newman D., Karimi S., Cavedon L. "External Evaluation of Topic Models", in: *ADCS 2009 – Proceedings of the Fourteenth Australasian Document Computing Symposium*. Sydney: University of Sydney, 2011, p. 1–8.
44. Scialom T., Dray P.-A., Lamprier S., et al. "MLSUM: The Multilingual Summarization Corpus", in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, p. 8051–8067. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.647.