Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Центр дополнительного, дополнительного профессионального и онлайн-образования "Пуск"
Кафедра дискретной математики

Направление подготовки / специальность: 01.04.02 Прикладная математика и информатика
Направленность (профиль) подготовки: Современная комбинаторика

Обучение тематических моделей внимания на тематически несбалансированных текстовых коллекциях

(магистерская диссертация)

Студент:
Нагаев Константин Владимирович

_____
(подпись студента)

Научный руководитель:
Воронцов Константин Вячеславович,
д.ф.-м.н., профессор РАН

_____
(подпись научного руководителя)

Консультант *(при наличии)***:**
_____
_____
(подпись консультанта)

Москва 2026

Оглавление

Введение

Актуальность

Тематическое моделирование широко применяется для разведочного анализа текстовых коллекций, информационного поиска и классификации. Реальные коллекции почти всегда тематически несбалансированы: тематический состав описывается дискретным распределением с «длинным хвостом», и несколько доминирующих тем покрывают большую долю документов, тогда как множество редких тем представлено единичными текстами. Классические модели (LDA, NMF) и современные нейросетевые подходы систематически теряют редкие, но содержательно важные темы, поскольку правдоподобие смещено к частотным образцам, а априорные допущения навязывают равномерность тематического состава.

Модель AARTM (Attentive Additively Regularized Topic Model), предложенная К. В. Воронцовым [1], относится к новому семейству тематических моделей внимания. Она отказывается от гипотезы «мешка слов» и от тематизации документа целиком, приписывая тему слову по его локальному контексту, и в то же время наследует аппарат аддитивной регуляризации ARTM. Теоретическое изложение модели носит аналитический характер; автор прямо формулирует пригодность модели для тематически несбалансированных коллекций как открытый вопрос, требующий эмпирической проверки. Это определяет актуальность настоящей работы – систематическое экспериментальное исследование эффективности AARTM в условиях тематической несбалансированности.

Объект и предмет исследования

Объект исследования – вероятностные тематические модели внимания семейства AARTM и алгоритмы их обучения.

Предмет исследования – качество модели AARTM при обучении на тематически несбалансированных коллекциях, а также влияние гиперпараметров и архитектурных компонентов модели на качество выделения редких (малопредставленных) тем.

Цель и задачи

Цель работы – проверить гипотезу, что новая модель AARTM эффективна при обучении на тематически несбалансированных текстовых коллекциях, выявить влияние ее гиперпараметров на качество тем, включая редкие, и сформулировать обоснованные рекомендации по обучению модели в условиях тематической несбалансированности.

Для достижения цели поставлены следующие **задачи**:

1. Провести обзор методов тематического моделирования и подходов к обработке тематически несбалансированных коллекций; формализовать понятие тематической несба-

лансированности и метрики ее оценки.

2. Описать модель AARTM, ее EM-алгоритм обучения и теоретические предпосылки пригодности для несбалансированных коллекций.
3. Освоить имеющуюся программную реализацию AARTM; провести верификацию ее корректности.
4. Подготовить экспериментальную базу на коллекции 20 Newsgroups с искусственно вводимым тематическим дисбалансом.
5. Сформировать метрический протокол оценки качества, ориентированный на раздельный анализ частых и редких тем.
6. Провести сравнительные и абляционные эксперименты, оценить влияние гиперпараметров модели и степени дисбаланса на качество редких тем.
7. Проанализировать результаты, выявить условия эффективности AARTM на несбалансированных данных и сформулировать практические рекомендации.

Гипотезы исследования

- **Г1.** При переходе от сбалансированной коллекции к несбалансированной качество AARTM на редкой теме деградирует слабее, чем у базовых моделей.
- **Г2.** Контекстный член N_{tw} повышает когерентность тем, в первую очередь редких.
- **Г3.** Существуют значения гиперпараметров AARTM (число тем, коэффициент сглаживания γ, длина контекста), при которых качество тем максимально; зависимость качества от этих параметров носит немонотонный характер.

Научная новизна

1. Впервые проведено систематическое эмпирическое исследование эффективности модели AARTM, ранее описанной лишь теоретически, на тематически несбалансированных коллекциях.
2. Получены результаты сравнения качества выделения редких тем модели AARTM с базовыми моделями тематического моделирования на сбалансированной и несбалансированных коллекциях.
3. Установлен вклад архитектурных компонентов и гиперпараметров AARTM в качество редких тем в условиях несбалансированности.
4. Предложен метрический протокол оценки тематических моделей на несбалансированных коллекциях с раздельным анализом частых и редких тем.

Практическая значимость

Результаты работы дают обоснованные рекомендации по выбору гиперпараметров и условий применения модели AARTM при обработке реальных несбалансированных текстовых коллекций. Метрический протокол и сценарии искусственного дисбаланса применимы для воспроизводимого сравнения тематических моделей в задачах, где важно выделение редких тем. Выявленные ограничения модели определяют направления ее дальнейшего развития.

Положения, выносимые на защиту

1. Предложена методика оценивания тематических моделей в условиях тематической несбалансированности данных, позволяющая дезагрегировать оценки по отдельным темам.
2. Подтверждена гипотеза о слабой деградации тематической модели внимания AARTM при усилении несбалансированности данных; показано, что качество выделения редких тем у AARTM снижается незначительно.
3. Показано, что в модели AARTM важно учитывать не только размер локального контекста, но и попарную сочетаемость слов в локальных контекстах, что повышает когерентность и различность тем.

Структура работы

Работа состоит из введения, четырех глав, заключения и списка литературы. Первая глава содержит обзор методов тематического моделирования и подходов к обработке несбалансированных коллекций. Вторая глава описывает модель AARTM и ее EM-алгоритм обучения. Третья глава определяет методику экспериментального исследования. Четвертая глава содержит результаты экспериментов и проверку гипотез. В заключении сформулированы выводы и направления дальнейшей работы.

1 Обзор методов тематического моделирования и подходов к несбалансированным коллекциям

1.1 Постановка задачи тематического моделирования

Тематическое моделирование [1] – класс методов статистического анализа текстов, выявляющих в режиме обучения без учителя латентную тематическую структуру коллекции документов. Каждая *тема* описывается распределением вероятностей над словарем, а документ – смесью тем.

Наблюдаемыми данными служат счетчики n_{dw} – частоты вхождения терма $w \in W$ в документ $d \in D$; собственно тематическая структура коллекции остается скрытой. Ее описывают две стохастические матрицы, подлежащие оценке: $\Phi = (\varphi_{wt})$, $\varphi_{wt} = p(w \mid t)$ – распределения слов в темах, и $\Theta = (\theta_{td})$, $\theta_{td} = p(t \mid d)$ – распределения тем в документах. В предположении условной независимости слова и документа при известной теме вероятность встретить терм в документе раскладывается по темам:

$$p(w \mid d) = \sum_{t \in T} \varphi_{wt}\, \theta_{td}. \tag{1.1}$$

Восстановление Φ и Θ по наблюдаемым счетчикам n_{dw} есть задача стохастического матричного разложения матрицы «документ–слово» [1]. Это разложение неединственно, что служит фундаментальным источником неустойчивости тематических моделей [1].

Параметры Φ и Θ оценивают по принципу максимума правдоподобия: максимизируют логарифм правдоподобия коллекции при ограничениях неотрицательности и нормировки на столбцы матриц:

$$\mathcal{L}(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt}\, \theta_{td} \;\to\; \max_{\Phi, \Theta}; \tag{1.2}$$

$$\sum_{w \in W} \varphi_{wt} = 1,\ \varphi_{wt} \geq 0; \qquad \sum_{t \in T} \theta_{td} = 1,\ \theta_{td} \geq 0. \tag{1.3}$$

Задача является некорректно поставленной: она имеет в общем случае бесконечно много решений, что и составляет упомянутую неустойчивость и мотивирует введение регуляризации (раздел 1.3).

1.2 Несбалансированные коллекции: типология и трудности

Реальные текстовые коллекции почти никогда не бывают тематически однородными по объему: одни темы представлены тысячами документов, другие – единицами. Это свойство является не дефектом конкретной выборки, а структурной характеристикой естественного текстового потока, и тематическая модель, претендующая на практическую применимость,

должна оцениваться именно в условиях несбалансированности. В настоящем разделе систематизируются виды несбалансированности и разбираются механизмы, по которым она ухудшает качество тематических моделей.

Виды несбалансированности. Термин «несбалансированность» объединяет несколько различных по природе явлений, которые целесообразно различать.

Несбалансированность распространенности тем (тематический «длинный хвост») – неравномерное распределение документов по латентным темам: небольшое число доминирующих тем покрывает основную массу коллекции, а длинный хвост редких тем – лишь малую ее долю. Именно этот вид несбалансированности является предметом настоящей работы, а его крайний случай – редкая (малопредставленная) тема, для выделения которой модели доступно мало данных.

Несбалансированность объема категорий проявляется, когда коллекция снабжена внешней разметкой: размеры размеченных категорий различаются на порядки. При соответствии категорий темам этот вид несбалансированности служит управляемой моделью тематического распределения с «длинным хвостом» (heavy-tailed / long-tailed distribution) – занижение объема отдельной категории воспроизводит редкую тему в контролируемых условиях.

Несбалансированность длин документов – разброс числа слов в документах: короткие документы (заголовки, сообщения, аннотации) дают разреженные счетчики и ненадежные оценки распределения тем в документе. Для коротких текстов разработаны специальные модели, агрегирующие статистику совместной встречаемости токенов [24, 23]. Этот вид несбалансированности в настоящей работе не рассматривается.

Отдельно следует различать несбалансированность *обучающей* и *контрольной* выборок. Модель может обучаться на сбалансированной коллекции, но применяться к потоку с иным, несбалансированным тематическим составом; качество на редкой теме в этом случае определяется уже не условиями обучения, а переносимостью обученной модели. Разделение этих двух ситуаций существенно для корректной постановки эксперимента.

Количественные характеристики несбалансированности. Степень несбалансированности тематического состава измеряется несколькими взаимодополняющими величинами. Простейшая из них – *коэффициент несбалансированности*, отношение объема самой частой темы к объему самой редкой. Более устойчивы интегральные меры: нормированная энтропия распределения документов по темам (близость к нулю означает доминирование одной темы, к единице – равномерный состав) и коэффициент Джини. Такие характеристики позволяют задавать несбалансированность как управляемый параметр эксперимента, а не как качественную данность коллекции.

Механизмы потери редких тем. Стандартные тематические модели систематически недооценивают редкие темы; ниже разобраны основные причины этого.

Доминирование правдоподобия частыми темами. Логарифм правдоподобия складывается из вкладов отдельных токенов, и вклад редкой темы пропорционален доле описываемых ею документов. При максимизации функционал почти полностью определяется частыми темами, и редкая тема может быть принесена в жертву ради малого прироста общего правдоподобия. Тем же свойством обладают агрегированные метрики качества (раздел 1.4).

Симметричные априорные распределения. Латентное размещение Дирихле с симметричным априорным распределением над распределением тем неявно постулирует примерно равную распространенность тем. На несбалансированных данных это предположение нарушается, и модель оказывается смещена в сторону равномерного состава; частичным средством служит переход к асимметричным априорным распределениям [11].

Самоусиливающаяся динамика обучения. Итеративные алгоритмы обучения (EM и его аналоги) подвержены эффекту «богатый богатеет»: тема, захватившая на очередной итерации больше токенов, получает более четкое распределение слов (неопределенность снижается) и на следующей итерации притягивает еще больше токенов. Редкая тема, изначально получившая мало статистики, в этой динамике вытесняется.

Поглощение тематически близкой темой. Редкая тема, лексически пересекающаяся с крупной (например, узкая подтема внутри широкой предметной области), особенно уязвима: модели «выгоднее» описать ее документы доминирующей темой-соседом, чем поддерживать отдельную редкую тему. Поэтому деградация редкой темы зависит не только от ее объема, но и от наличия близких конкурирующих тем.

Коллапс тем. Нейросетевые тематические модели на основе вариационного автокодировщика склонны к вырождению части тем в повторяющиеся неинформативные распределения; в несбалансированном составе в число вырождающихся в первую очередь попадают редкие темы. Для противодействия коллапсу вводятся специальные регуляризаторы [14].

Доминирующий кластер. Модели, основанные на жесткой кластеризации документов (в частности, на плотных эмбеддингах), на несбалансированных данных порождают один-два крупных кластера и обширный класс шумовых документов с фоновым распределением, в который вытесняются документы редких тем.

Артефактность и неустойчивость редких тем. Отдельная трудность несбалансированных коллекций – размывание границы между настоящей редкой темой и артефактом разложения. Стохастическое матричное разложение неединственно (раздел 1.1), и при малом объеме статистики редкая тема может оказаться неустойчивой – воспроизводиться лишь при части инициализаций модели. Из этого следует, что вывод о выделении редкой темы требует проверки на воспроизводимость, а не однократного запуска модели.

Постановка в рамках настоящей работы. Из перечисленных видов несбалансированности предметом исследования является несбалансированность распространенности тем, а конкретной изучаемой ситуацией – судьба редкой темы при обучении модели внимания на коллекции с тематическим длинным хвостом. Несбалансированность длин документов и распределений частот слов с длинным хвостом остаются за рамками работы. Несбаланси-

рованность задается управляемо – искусственным занижением объема отдельной категории размеченной коллекции, что позволяет сравнивать поведение модели в сбалансированном и несбалансированном режимах при прочих равных условиях.

1.3 Обзор моделей тематического моделирования

В настоящем разделе дается обзор основных семейств тематических моделей в порядке их исторического появления; для каждого семейства разбирается, насколько его устройство приспособлено к обработке несбалансированных коллекций. Завершает раздел сводное сравнение семейств по этому признаку.

1.3.1 Матричные и вероятностные модели

Исторически первым подходом стал *латентно-семантический анализ* (LSA) [7]: усеченное сингулярное разложение матрицы «документ–слово» выделяет ортогональные латентные факторы. LSA не имеет вероятностной интерпретации, а ортогональность факторов и допустимость отрицательных значений затрудняют их трактовку как тем. В контексте несбалансированности существен другой недостаток: сингулярное разложение упорядочивает факторы по убыванию дисперсии, поэтому редкая тема, дающая малый вклад в дисперсию, либо вытесняется в младшие компоненты, либо вовсе не выделяется.

Неотрицательное матричное разложение (NMF) [10] устраняет проблему интерпретируемости, требуя неотрицательности обеих матриц-сомножителей, благодаря чему компоненты допускают аддитивную трактовку как темы. NMF остается востребованной моделью, в особенности на коротких текстах. Однако обучение NMF также управляется минимизацией ошибки реконструкции, в которой доминируют частые слова и крупные темы; специальных механизмов поддержки редких тем модель не содержит.

Вероятностный латентно-семантический анализ (PLSA) [8] – первая полноценно вероятностная тематическая модель, в которой документ описывается как смесь тем, а тема – как распределение над словарем; обучение ведется EM-алгоритмом. PLSA задала стандартную постановку задачи (раздел 1.1), но не имеет порождающей модели на уровне коллекции и подвержена переобучению, число параметров Θ растет с числом документов.

Латентное размещение Дирихле (LDA) [9] – наиболее распространенная тематическая модель – помещает обе матрицы PLSA – распределения тем в документах и слов в темах – под априорные распределения Дирихле. В отличие от PLSA, где Θ оценивается напрямую и число ее параметров растет с размером коллекции, априорное распределение Дирихле фиксирует число параметров, снижает переобучение и доопределяет PLSA до порождающей модели уровня всей коллекции. Вместе с тем именно априорное распределение является источником трудностей на несбалансированных данных: симметричное априорное распределение над распределением тем неявно постулирует примерно равную распространенность тем (раздел 1.2) и смещает модель к равномерному тематическому составу.

1.3.2 Адаптации к неравномерному тематическому составу

Осознание этого ограничения породило семейство моделей, изначально допускающих неравномерный тематический состав. Базовый прием – переход к *асимметричным априорным распределениям* [11]: асимметричное априорное распределение над распределением тем в документах позволяет модели усвоить, что разные темы имеют разную распространенность; при этом было показано, что асимметрия априорного распределения именно над темами документов полезна, тогда как асимметрия априорного распределения над словами тем скорее вредна. Асимметричное априорное распределение смягчает смещение LDA, но требует оценивания параметров априорного распределения и не устраняет доминирование правдоподобия частыми темами.

Более радикальное решение дают *непараметрические* модели. Иерархический процесс Дирихле (HDP) [12] и модели на основе процесса Питмана–Йора (Pitman-Yor Process, PYP) не фиксируют число тем заранее, а выводят его из данных. Порождающий процесс таких моделей напрямую задает распределение с длинным хвостом: он естественно дает небольшое число частых тем и длинный хвост редких. Тем самым непараметрические модели концептуально лучше всего согласованы со структурой несбалансированных коллекций. Их практические недостатки – вычислительная сложность вывода и чувствительность к параметрам процесса. Ещё один недостаток непараметрических моделей, в частности HDP, в том, что они содержат гиперпараметр, от которого итоговое число тем зависит критически, а при его варьировании фактически может принимать произвольное значение [3].

Для *коротких и несбалансированных текстов* разработаны модели, преодолевающие разреженность пословных счетчиков за счет перехода к статистике совместной встречаемости. Модель битермов (biterm TM, BTM) [24] моделирует не документы, а пары совместно встречающихся слов на уровне всей коллекции. Сетевая модель тем (WNTM) [23] строит граф совместной встречаемости слов и применяет тематическое моделирование к нему; авторы прямо позиционируют ее как решение для коротких и несбалансированных текстов. Эти модели смягчают несбалансированность длин документов, но не нацелены специально на тематический длинный хвост.

1.3.3 Аддитивная регуляризация (ARTM)

Подход *аддитивной регуляризации тематических моделей* (ARTM) [2, 4] отказывается от байесовского вывода и трактует тематическое моделирование как многокритериальную оптимизацию: к логарифму правдоподобия добавляется взвешенная сумма критериев-регуляризаторов, а решение строится EM-подобным алгоритмом. Тем самым ARTM разрешает некорректность задачи стохастического матричного разложения (раздел 1.1) не априорным распределением, а явными критериями.

Для несбалансированных коллекций существенна *комбинируемость* регуляризаторов: в одной модели сочетаются сглаживание фоновых тем, разреживание предметных тем, декоррелирование тем и частичное обучение по словам-затравкам. Промышленная реализация подхода – библиотека BigARTM [4]. Ускоренная векторизация текста для обработ-

ки больших массивов данных описана в [5]. Принципиально важно, что в рамках ARTM возможна *целенаправленная балансировка* тематического состава: регуляризатор балансировки тем [6] позволяет управлять распространенностью тем и поддерживать редкие темы. Это единственное из рассматриваемых здесь семейств, где работа с несбалансированностью вынесена на уровень явно конструируемого критерия, а не является побочным следствием устройства модели. Модель AARTM, исследуемая в настоящей работе, принадлежит линии развития ARTM.

1.3.4 Нейросетевые модели

Нейросетевые тематические модели [16] обучаются градиентными методами; доминирующая архитектура – *вариационный автокодировщик*, кодирующий документ в распределение над темами (ProdLDA, ETM [13]). ETM дополнительно представляет и слова, и темы векторами в общем эмбеддинг-пространстве, что улучшает работу с редкими словами. Главная трудность этого семейства на несбалансированных данных – *коллапс тем*: часть тем вырождается в повторяющиеся неинформативные распределения, и в число вырождающихся в первую очередь попадают редкие темы. Модель ECRTM [14] вводит специальный регуляризатор, заставляющий каждую тему занимать обособленную область эмбеддинг-пространства и тем самым прямо противодействующий коллапсу.

1.3.5 Контекстуализированные модели и подходы на основе языковых моделей

Отдельную линию образуют модели на *контекстных эмбеддингах*, использующие предобученные языковые представления. BERTopic [17] кластеризует документы в пространстве плотных эмбеддингов и извлекает темы из кластеров. Контекстуализированные тематические модели [18] объединяют эмбеддинги предложений с вариационным автокодировщиком. FASTopic [15] строит темы через оптимальный транспорт между эмбеддингами документов, тем и слов. Слабое место кластеризующих моделей на несбалансированных данных – зависимость от алгоритма кластеризации: на неравномерных данных кластеризация порождает один-два доминирующих кластера и обширный класс шумовых документов, в который вытесняются документы редких тем. Подходы на основе больших языковых моделей (TopicGPT [19]) переводят задачу в режим инструктирования модели; их поведение на редких темах пока слабо изучено и зависит от формулировки запроса.

1.3.6 Сводное сравнение по отношению к несбалансированности

Сопоставление семейств моделей по способу работы с несбалансированными коллекциями приведено в таблице 1.1.

Сравнение выявляет несколько закономерностей. Большинство семейств не содержит механизма поддержки редкой темы вовсе: в LSA, NMF и контекстуализированных моделях редкая тема теряется как побочное следствие оптимизируемого функционала. Часть семейств приспособлена к неравномерному составу *структурно* – непараметрические модели благодаря порождающему процессу распределений с длинным хвостом, нейросетевая

Таблица 1.1 – Семейства тематических моделей и их отношение к несбалансированности тематического состава

Семейство	Механизм относительно редких тем	Ограничение на несбалансированных данных
LSA	Отсутствует; факторы упорядочены по дисперсии	Редкая тема вытесняется в младшие компоненты
NMF	Отсутствует	Реконструкция управляется частыми словами
PLSA, LDA	Симметричное априорное распределение (LDA)	Смещение к равномерному составу; переобучение PLSA
LDA с асимметричным априорным распределением	Асимметричное априорное распределение над темами документов	Доминирование правдоподобия частыми темами сохраняется
Непараметрические (HDP, Питман–Йор)	Порождающий процесс распределений с длинным хвостом	Вычислительная сложность вывода
Модели коротких текстов (BTM, WNTM)	Агрегирование совместной встречаемости	Нацелены на длину документов, не на тематический хвост
ARTM	Явный регуляризатор балансировки тем	Требует подбора весов регуляризаторов
Нейросетевые (ETM, ECRTM)	Регуляризатор против коллапса тем (ECRTM)	Склонность к коллапсу; чувствительность к обучению
Контекстуализированные (BERTopic, FASTopic)	Специальный механизм отсутствует	Доминирующий кластер поглощает редкие темы

ECRTM благодаря антиколлапсному регуляризатору, – но эта приспособленность фиксирована устройством модели и не управляема. Лишь ARTM выносит работу с несбалансированностью на уровень явно конструируемого критерия, допускающего целенаправленную балансировку. Этим обусловлен выбор линии ARTM в качестве основы настоящей работы: модель AARTM наследует от ARTM управляемость через регуляризаторы и дополняет ее механизмом локального контекстного внимания, рассматриваемым в главе 2.

1.4 Метрики оценки на несбалансированных коллекциях

Оценка качества тематической модели сама по себе является нетривиальной исследовательской задачей: тематическая структура латентна, эталонная разметка обычно недоступна, а интерпретируемость тем – свойство, плохо поддающееся формализации. На несбалансированных коллекциях задача дополнительно осложняется тем, что стандартная практика сообщает по каждой метрике единственное агрегированное число, тогда как качество редкой темы в этом числе теряется. Ниже подходы к оценке рассматриваются в исторической перспективе, что позволяет проследить, как формировалось и осознавалось это ограничение.

Перплексия и правдоподобие. Исторически первой парадигмой оценки была вероятностная: качество модели измерялось правдоподобием на отложенной выборке или эквивалентной ему *перплексией*; именно так оценивалось латентное размещение Дирихле [9]. Перплексия остается необходимым инструментом сравнения предсказательной способности моделей, однако обладает двумя принципиальными недостатками. Во-первых, она доминируется частыми словами и темами: вклад редкой темы в суммарное правдоподобие пропорционален доле описываемых ею слов, поэтому провал качества на редкой теме почти не отражается в агрегированной перплексии. Во-вторых, перплексия слабо и даже отрицательно коррелирует с воспринимаемой человеком интерпретируемостью тем [22]: модель с меньшей перплексией может давать темы худшего качества. Причина расхождения в том, что перплексия и человеческая оценка измеряют разное. Перплексия вознаграждает точное предсказание любого терма, тогда как человек судит о теме по семантической связности и четкости ее границ. Поэтому модели выгодно уверенно предсказывать лексику, не несущую тематической нагрузки, – стоп-слова, предлоги, союзы, высокочастотные служебные термы (например, http, jpg, com, RT), технические токены разметки: они дают весомый вклад в правдоподобие, но не входят в семантическое ядро темы. Тот же разрыв порождает две стратегии, понижающие перплексию ценой интерпретируемости:

- *фрагментация* – дробление темы на узкие подтемы: локальная подгонка снижает перплексию, но темы дублируются и теряют обобщенность;
- *слияние* – объединение разных тем в одну: перплексия падает за счет сглаживания распределений, а темы становятся размытыми или «мусорными».

Расхождение между перплексией и интерпретируемостью стало причиной перехода к метрикам интерпретируемости.

Когерентность тем. Вторая парадигма – оценка *когерентности*, то есть смысловой связности топ-слов темы, вычисляемой по статистике их совместной встречаемости. Меры этого семейства развивались от UMass-когерентности, оптимизирующей семантическую связность [21], к мерам на основе поточечной взаимной информации; систематическое сравнение мер когерентности и обоснование меры NPMI и обобщенной меры C_v дано в [20]. Когерентность устранила разрыв между перплексией и интерпретируемостью, однако не сняла другого ограничения: она по-прежнему сообщается как единственное усредненное по темам значение. Высокая средняя когерентность совместима с полной потерей нескольких редких тем, поскольку вклад одной темы в среднее составляет лишь $1/|T|$.

Метрики структуры тематического пространства. Параллельно развивались метрики, характеризующие тематическое пространство в целом: разнообразие тем (доля уникальных слов среди топ-слов всех тем), уникальность отдельной темы, попарные расстояния между темами (в частности, расстояние Хеллингера между распределениями термов). Эти метрики чувствительны к вырождению и слиянию тем – явлениям, которым редкие темы подвержены в

первую очередь, – и потому особенно уместны на несбалансированных данных. Тем не менее в стандартной форме большинство из них также сводится к одному агрегированному числу.

Критика автоматической оценки. Современный этап характеризуется критическим пересмотром автоматических метрик. Систематические исследования показали, что автоматические меры когерентности не всегда надежно воспроизводят человеческие суждения о качестве тем, в особенности для нейросетевых моделей [22]. Это привело к возврату интереса к оценке с участием экспертов, к более осторожной интерпретации автоматических метрик и к осознанию того, что единственное агрегированное значение не описывает модель адекватно.

Несбалансированность как слепое пятно стандартной оценки. Сквозное наблюдение ретроспективы состоит в том, что на всех перечисленных этапах единицей отчетности оставалось одно агрегированное число на метрику. Ни перплексия, ни когерентность, ни метрики разнообразия в их стандартной форме не были спроектированы так, чтобы делать наблюдаемым качество *отдельной* редкой темы. На сбалансированной коллекции это несущественно: «типичная» тема репрезентативна для всех. На несбалансированной коллекции агрегат описывает доминирующие темы, а редкая тема, ради выделения которой и применяется модель, в нем растворяется. Таким образом, оценка редких тем оказывается слепым пятном устоявшейся методологии.

Подходы к оценке редких тем. Ограничение агрегированной оценки породило несколько методических направлений, прицельно ориентированных на редкие темы. Первое – отказ от единственного значения в пользу *распределения* метрики по темам с акцентом на нижние квантили, описывающие наименее качественные темы. Второе – противопоставление микро- и макро-усреднения: макро-усреднение, вычисляющее метрику отдельно для каждой темы и усредняющее с равным весом, дает редкой теме голос наравне с частой. Третье – стратифицированная (раздельная по темам) оценка, при которой предсказательное качество измеряется для каждой темы по отдельности. Четвертое – оценка устойчивости тем между запусками с разными зернами инициализации: редкая тема, не воспроизводящаяся в части запусков, скорее является артефактом разложения, чем устойчивой находкой. Пятое – целевой качественный анализ конкретной редкой темы. Перечисленные приемы не образуют общепринятого стандарта; они представляют собой направления, логически следующие из принципа дезагрегации.

Постановка в рамках настоящей работы. Опираясь на этот принцип, в настоящей работе формируется метрический протокол, в котором каждая допускающая это метрика рассматривается как распределение по темам, а для редкой темы вводятся прицельные метрики – стратифицированная перплексия по темам и уникальность темы. Протокол подробно описан в главе 3; там же стратифицированная перплексия и уникальность темы позиционируются как собственное методическое решение работы.

1.5 Выводы по главе

Несбалансированность – типичное структурное свойство реальных текстовых данных, затрагивающее все семейства тематических моделей. Систематизированы виды несбалансированности и механизмы потери редких тем; показано, что большинство семейств тематических моделей не содержит управляемого механизма поддержки редкой темы, а линия ARTM выделяется возможностью целенаправленной балансировки тематического состава через явный регуляризатор. Ретроспективный разбор подходов к оценке показал, что стандартные метрики в агрегированной форме маскируют провал качества на редких темах, и корректная оценка требует дезагрегации метрик по темам. Модель AARTM, исследуемая в настоящей работе, относится к семейству аддитивно регуляризованных моделей локального контекста и описывается в следующей главе.

2 Модель тематического моделирования внимания AARTM

В настоящей главе описывается модель AARTM (Attentive Additively Regularized Topic Model), исследуемая в работе. Изложение опирается на работы К. В. Воронцова [1] и теорию аддитивной регуляризации тематических моделей.

2.1 Базовая тематическая модель и аддитивная регуляризация

Пусть D – коллекция текстовых документов, W – словарь термов. При гипотезе «мешка слов» документ d представляется мультимножеством с частотами n_{dw}. Тематическая модель описывает вероятность появления терма через формулу полной вероятности и гипотезу условной независимости $p(w \mid d, t) = p(w \mid t)$:

$$p(w \mid d) = \sum_{t \in T} \varphi_{wt}\, \theta_{td}. \tag{2.1}$$

Параметры оцениваются максимизацией логарифма правдоподобия

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt}\, \theta_{td} \;\to\; \max_{\Phi, \Theta} \tag{2.2}$$

при ограничениях неотрицательности и нормировки. Задача является некорректно поставленной задачей стохастического матричного разложения. Подход *аддитивной регуляризации* (ARTM) разрешает некорректность введением критериев-регуляризаторов $R_k(\Phi, \Theta)$ и решением регуляризованной задачи

$$L(\Phi, \Theta) + \sum_{k=1}^{K} \tau_k R_k(\Phi, \Theta) \;\to\; \max_{\Phi, \Theta}, \tag{2.3}$$

где $\tau_k \geq 0$ – коэффициенты регуляризации. Существенным элементом теории является операция нормировки, проецирующая числовой вектор на единичный симплекс:

$$\operatorname*{norm}_{i \in I}(x_i) = \frac{\max\{0, x_i\}}{\sum_{k \in I} \max\{0, x_k\}}. \tag{2.4}$$

2.2 Симметризованная модель на тематических эмбеддингах

Определение 1. *Тематический эмбеддинг* – это представление произвольного объекта x (терма, документа, контекста, коллекции) вектором в пространстве тем T; координатами вектора служат вероятности $p(t \mid x)$, образующие дискретное распределение на T.

Базовая модель несимметрична: связь документов с темами описывается эмбеддингами $p(t \mid d)$, а связь термов с темами – частотными словарями $p(w \mid t)$. Используя тождество $p(t \mid w)\, p(w) = p(w \mid t)\, p(t)$, где $p(w) = n_w/n$ определяется данными, получаем симметризованную модель

$$p(w \mid d) = \sum_{t \in T} \varphi_{tw}\, \theta_{td}\, \frac{p(w)}{p(t)}, \tag{2.5}$$

в которой матрица $\Phi = (\varphi_{tw})$ содержит тематические эмбеддинги термов $p(t \mid w)$, а $p(t)$ – тематический эмбеддинг всей коллекции. Распределение $p(t)$ возникает в EM-алгоритме как несмещенная оценка тематической массы: $p(t) = n_t/n$.

2.3 Модель локального контекста

Базовые модели опираются на гипотезу тематической однородности документа. Модель локального контекста отказывается от нее, определяя тематику терма по его локальной окрестности. Документы конкатенируются в единую последовательность $\{w_1, \ldots, w_n\}$; для терма w_i вводится *контекст* C_i – множество позиций термов локальной окрестности.

Вероятностная модель локального контекста:

$$p(w \mid C_i) = \sum_{t \in T} \varphi_{tw}\, \frac{p(w)}{p(t)}\, \theta_{ti}, \tag{2.6}$$

где *тематический эмбеддинг контекста* есть средневзвешенное эмбеддингов термов:

$$\theta_{ti} = \sum_{c \in C_i} \alpha_{ci}\, \varphi_{tw_c}, \qquad \sum_{c \in C_i} \alpha_{ci} = 1, \quad \alpha_{ci} \geq 0, \tag{2.7}$$

а α_{ci} – *коэффициенты внимания*. Параметры Φ оцениваются максимизацией регуляризованного правдоподобия $L(\Phi) = \sum_{i=1}^{n} \ln p(w_i \mid C_i) + R(\Phi)$. Применение основной леммы ARTM дает E-шаг

$$p_{ti} = \underset{t \in T}{\operatorname{norm}}\big(\varphi_{tw_i}\, \theta_{ti}/p(t)\big) \tag{2.8}$$

и M-шаг

$$\varphi_{tw} = \underset{t \in T}{\operatorname{norm}}\left(n_{tw} + \varphi_{tw} N_{tw} - \varphi_{tw}\, n_t\, \frac{p(w)}{p(t)} + \varphi_{tw} \frac{\partial R}{\partial \varphi_{tw}}\right), \tag{2.9}$$

где счетчики

$$n_{tw} = \sum_{i=1}^{n} p_{ti}\, [w_i = w], \qquad N_{tw} = \sum_{i=1}^{n} q_{wi}\, \frac{p_{ti}}{\theta_{ti}}, \qquad q_{wi} = \sum_{c \in C_i} \alpha_{ci}\, [w_c = w]. \tag{2.10}$$

Содержательный смысл членов M-шага раскрывается следующими следствиями: при контексте, равном документу, и равномерном внимании справедливо $N_{tw} = n_w$, а после подстановки несмещенных оценок $\varphi^*_{tw} = n_{tw}/n_w$ третий член отбрасывается. Член N_{tw} накапливает совстречаемость терма w с контекстами, отнесенными к теме t: чем чаще w попадает в окружение слов темы t, тем сильнее добавка $\varphi_{tw} N_{tw}$ тянет w в эту тему. В ре-

зультате тема стягивает лексику, собирающуюся в сходных контекстах; ожидаемое следствие – рост когерентности, наиболее заметный на коллекциях коротких документов. В частном случае контекста-документа модель переходит в Theta-less ARTM, что используется для верификации реализации.

2.4 Механизм внимания через скользящие средние

Внимание реализуется вычислительно эффективным способом – двумя экспоненциальными скользящими средними, проходящими по тексту в двух направлениях:

$$\overrightarrow{p}(t \mid i) = \gamma_i \, \varphi_{tw_i} + (1 - \gamma_i) \, \overrightarrow{p}(t \mid i-1), \tag{2.11}$$

$$\overleftarrow{p}(t \mid i) = \gamma_i \, \varphi_{tw_i} + (1 - \gamma_i) \, \overleftarrow{p}(t \mid i+1), \tag{2.12}$$

где $\gamma_i \in [0, 1]$ – коэффициент сглаживания. При постоянном γ веса убывают с расстоянием по геометрической прогрессии: $\alpha_{ci} = \gamma(1-\gamma)^{|i-c|}$; эффективное число усредняемых термов составляет приблизительно $1/\gamma$. Эмбеддинг двустороннего контекста – взвешенное среднее левого и правого. Модель локального контекста с этим механизмом внимания представляет экземпляр в семействе моделей AARTM. Веса внимания здесь не оптимизируются, а полностью определяются позицией терма и коэффициентом γ; обучение AARTM сводится к оценке одной матрицы Φ и остается, по существу, быстрой мягкой кластеризацией коллекции.

2.5 Предпосылки эффективности на несбалансированных коллекциях

Из устройства модели вытекают четыре теоретические предпосылки, по которым AARTM может оказаться более устойчивой к тематическому дисбалансу, чем модели «мешка слов». Все они рассматриваются как обоснованные ожидания, а не доказанные свойства.

Отказ от тематизации документа целиком. В классических моделях ARTM и LDA тематический эмбеддинг $p(t \mid d)$ оценивается для всего документа сразу, и при этом любое слово документа вносит вклад в каждую из тем документа пропорционально весу темы в нем. Документы редкой категории при дисбалансе малочисленны, и их вклад в матрицу Θ оказывается мал; при оценке матрицы Φ редкая тема «обделена» вкладами и склонна растворяться в соседних темах или сливаться с ними в неоднородную тему-агломерат. Модель AARTM не оценивает Θ непосредственно: тема приписывается слову по локальному контексту фиксированного масштаба, и каждое вхождение редкой темы дает вклад в Φ независимо от длины и числа документов, ее содержащих. Это снимает зависимость качества редкой темы от объема выборки соответствующей категории на уровне документов; редкая тема, имея даже малое число документов, может проявиться через множество локальных контекстов внутри этих документов.

Контекстная регуляризация когерентности. Член N_{tw} в М-шаге накапливает совстречаемость терма w с контекстами, отнесенными к теме t (раздел 2.3; подробный вывод в [1]). При нормировке по теме это означает, что для каждой темы предпочитаются термы,

собирающиеся в одинаковых контекстах. Для редкой темы, представленной малым числом документов, такая концентрация особенно важна: даже скромного числа контекстов, в которых редкая лексика встречается совместно, оказывается достаточно для формирования различимого ядра темы. На моделях «мешка слов» такого механизма нет, и редкая тема либо сливается с частыми (если лексика частично разделяется с ними), либо вырождается в шумовое распределение (если частота недостаточна для образования устойчивого паттерна).

Снижение переобучения за счет сокращения размерности параметров. В классической PLSA число параметров матрицы Θ растет с размером коллекции и составляет $|T| \times |D|$. LDA устраняет этот рост введением априорного распределения, но за счет регуляризации, которая работает усреднением и склонна «смазывать» редкие явления. AARTM устраняет рост размерности структурно, исключая Θ из числа параметров: тематический эмбеддинг любого фрагмента вычисляется по матрице Φ за один проход. Это снижает риск переобучения и, что важнее для несбалансированных данных, делает оценку Φ более устойчивой к малому числу документов отдельной категории — параметры Φ оцениваются по всем вхождениям термов, а не по документам как единицам.

Управляющий потенциал распределения $p(t)$. В формуле E-шага распределение $p(t)$ входит в знаменатель: чем больше $p(t)$ темы, тем меньше вероятность отнесения вхождения к этой теме при прочих равных. Если темы оказываются неравномерно представленными в коллекции (тема, доминирующая по объему лексики, имеет большее $p(t)$), знаменатель частично компенсирует это смещение, ослабляя самоусиление частых тем. В [1] отмечено, что $p(t)$ может вводиться в модель как внешний управляющий параметр, что закладывает теоретическую основу для целенаправленной балансировки тем при обучении. Управление $p(t)$ как метод противодействия дисбалансу составляет отдельную исследовательскую программу и в рамках настоящей работы не реализуется; рассматривается только модель AARTM в стандартной конфигурации, в которой $p(t)$ оценивается по данным и роль ее ограничивается естественным эффектом, описанным выше.

Исходная работа [1] носит теоретический характер: модель AARTM описана и обоснована аналитически, но ее эмпирические свойства, в том числе пригодность для несбалансированных коллекций, оставлены как открытый вопрос. Перечисленные четыре предпосылки являются обоснованными ожиданиями и составляют теоретическое основание гипотез исследования, сформулированных в следующей главе.

2.6 Выводы по главе

Описана модель AARTM – тематическая модель внимания семейства аддитивно регуляризованных моделей локального контекста. Модель отказывается от гипотезы «мешка слов» и тематизации документа целиком, приписывая тему слову по локальному контексту с механизмом внимания на основе экспоненциальных скользящих средних. Обучение сводится к EM-алгоритму, наследующему аппарат регуляризации ARTM. Выделены четыре теоретические предпосылки эффективности модели на несбалансированных коллекциях, проверяемые в экспериментальной части работы.

3 Методика экспериментального исследования

3.1 Гипотезы исследования

Экспериментальное исследование организовано вокруг трех гипотез.

- **Г1** (деградация на редкой теме). При переходе от сбалансированной коллекции к несбалансированной качество AARTM на редкой теме деградирует слабее, чем у базовых моделей (LDA, NMF).
- **Г2** (вклад члена N_{tw}). Контекстный член N_{tw} в М-шаге повышает когерентность тем, в первую очередь редких; проверяется абляционным сравнением полной модели с вариантом без N_{tw}.
- **Г3** (немонотонность по гиперпараметрам). Существуют значения гиперпараметров AARTM (число тем, коэффициент γ, длина контекста), при которых качество полученных тем максимально, то есть, зависимость носит немонотонный характер.

3.2 Базовая коллекция и предобработка

В качестве базовой коллекции используется размеченный корпус **20 Newsgroups**, содержащий новостные сообщения 20 тематических категорий. Наличие относительно достоверной разметки позволяет проводить целевой анализ редкой темы; известная категориальная структура делает возможным контролируемое внесение дисбаланса.

Предобработка единообразна для всех моделей и коллекций и выполняется программным кодом `corpus_preparation.py`: удаление служебных полей сообщений (`headers`, `footers`, `quotes`), приведение к нижнему регистру, фильтрация термов по длине (от 3 до 20 символов), удаление стандартных стоп-слов английского языка в пакете nltk, фильтрация словаря (удаление термов, встречающихся менее чем в 3 документах и более чем в 50 % документов), удаление пустых документов. Словарь строится на полном корпусе и используется для всех коллекций.

3.3 Коллекции с искусственным дисбалансом

Тематический дисбаланс вносится искусственно скриптом `dataset_preparation.py` – заданием структуры представленности категорий. Определены четыре коллекции (таблица 3.1).

Коллекция `full` – сбалансированный эталон. Коллекция `minus_rel` уменьшает количество документов в категории soc.religion.christian, для которой присутствуют тематически пересекающиеся соседи (`alt.atheism`, `talk.religion.misc`), `minus_med` – уменьшает количество документов в категории sci.med, тематически изолированной от остальных категорий.

Таблица 3.1 – Структура коллекций

Коллекция	Редкая категория	Док. редкой кат. (обуч.)	Док. редкой кат. (тест)	Прочие кат. (обуч.)
`full`	–	1000	50	1000
`minus_rel`	soc.religion.christian	50	10	1000
`minus_med`	sci.med	50	10	1000
`imbalanced_test`	soc.religion.christian	1000	10	1000

Коллекция `imbalanced_test` сбалансирована по обучающей выборке и несбалансирована по тестовой – используется в эксперименте на перенос. Сопоставление `minus_rel` и `minus_med` разделяет деградацию из-за малого объема данных и деградацию из-за конкуренции с близкими темами.

Для краткости в дальнейшем изложении используются короткие идентификаторы коллекций, совпадающие с именами в программной реализации; их описательные названия приведены в таблице 3.2.

Таблица 3.2 – Описательные названия коллекций

Идентификатор	Описательное название
`full`	Полный (сбалансированный) набор
`minus_rel`	Дисбаланс обучающей и тестовой выборок по категории `soc.religion.christian`
`minus_med`	Дисбаланс обучающей и тестовой выборок по категории `sci.med`
`imbalanced_test`	Дисбаланс только в тестовой выборке по категории `soc.religion.christian`

3.4 Метрический протокол оценки эффективности модели

Оценка тематических моделей на несбалансированных коллекциях методологически сложнее, чем на сбалансированных: стандартные метрики, агрегированные по всей коллекции, систематически маскируют провал качества на редких темах. Настоящий раздел сначала рассматривает стандартные подходы к оценке и их ограничения в условиях несбалансированности (3.4.1), затем описывает предлагаемый протокол с раздельным анализом тем (3.4.2), центральным элементом которого является стратифицированная перплексия по темам.

3.4.1 Стандартные подходы и их ограничения при несбалансированности

Перплексия. Перплексия – классическая мера предсказательного качества тематической модели, определяемая через логарифм правдоподобия на выборке документов:

$$\mathrm{PPL} = \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\ln p(w_i \mid C_i)\right), \tag{3.1}$$

где n – суммарное число термов выборки, а $p(w_i \mid C_i)$ – вероятность терма, предсказываемая моделью. Чем ниже перплексия, тем точнее модель аппроксимирует данные.

В условиях несбалансированности перплексия обладает двумя существенными ограничениями, разобранными в разделе 3.4.1 обзора. Первое следует прямо из вида (3.1): сумма берется по всем термам коллекции, и редкая тема с малой долей термов почти не сдвигает итог – ее провал остается за пределами агрегированной оценки. Второе – установленная эмпирически слабая (и часто отрицательная) корреляция перплексии с человеческой оценкой интерпретируемости. Поэтому перплексия используется в работе как вспомогательная метрика общего качества модели, но не как основной критерий оценки качества редких тем.

Когерентность. Когерентность оценивает семантическую связность темы по ее топ-словам. Распространенная мера – нормированная поточечная взаимная информация (NPMI) [20]. Для пары термов w_i, w_j

$$\mathrm{NPMI}(w_i, w_j) = \frac{1}{-\ln p(w_i, w_j)} \ln \frac{p(w_i, w_j)}{p(w_i)\, p(w_j)}, \tag{3.2}$$

где $p(w_i), p(w_j)$ – вероятности встречаемости термов в документах опорного корпуса, $p(w_i, w_j)$ – вероятность их совместной встречаемости. Когерентность темы t с топ-k словами есть среднее NPMI по всем парам:

$$C_{\mathrm{NPMI}}(t) = \frac{2}{k(k-1)} \sum_{1 \le i < j \le k} \mathrm{NPMI}(w_i^{(t)}, w_j^{(t)}). \tag{3.3}$$

Стандартная практика – сообщать единственное значение, усредненное по всем темам: $\overline{C}_{\mathrm{NPMI}} = \frac{1}{|T|} \sum_t C_{\mathrm{NPMI}}(t)$. Именно усреднение является источником проблемы при несбалансированности: высокое среднее значение может сосуществовать с полной потерей одной или нескольких редких тем, поскольку вклад редкой темы в среднее составляет лишь $1/|T|$. Среднее по темам не отвечает на вопрос, выделена ли редкая тема качественно, – оно характеризует «типичную» тему, которой в несбалансированной коллекции почти всегда оказывается частая тема.

Способ усреднения. Более общая методологическая проблема – противопоставление микро- и макро-усреднения. *Микро-усреднение* агрегирует величину по всем элементарным наблюдениям (термам, документам) и потому доминируется крупными темами. *Макро-усреднение* вычисляет метрику в данном случае отдельно для каждой темы и усредняет результаты с равным весом, благодаря чему редкая тема влияет на итог наравне с частой. Для несбалансированных коллекций информативно как макро-усредненное значение, так и само расхождение микро- и макро-оценок: значительный разрыв между ними служит индикатором того, что общее качество модели держится на доминирующих темах.

Вывод. Стандартные метрики в их обычной, агрегированной форме недостаточно пригодны для прямой оценки качества редких тем. Корректная оценка на несбалансированной

коллекции требует рассмотрения метрик не как скаляров, а как распределений по темам, и введения метрик, прицельно характеризующих качество отдельной редкой темы.

3.4.2 Предлагаемый протокол с раздельным анализом тем

Предлагаемый метрический протокол основан на принципе *дезагрегации*: каждая метрика, допускающая это, рассматривается как распределение по темам, а качество редкой темы оценивается отдельно от качества коллекции в целом. Протокол включает пять групп метрик.

Стратифицированная перплексия по темам. Центральный элемент протокола – стратифицированная (per-topic) перплексия, дающая оценку предсказательного качества отдельно для каждой темы. Для ее вычисления каждый терм w_i выборки относится к своей *доминирующей теме* – теме с наибольшей вероятностью в тематическом эмбеддинге его контекста:

$$t^*(i) = \arg\max_{t \in T} \theta_{ti}, \qquad \theta_{ti} = p(t \mid C_i). \tag{3.4}$$

Перплексия темы t вычисляется по подмножеству термов, для которых она доминирующая:

$$\mathrm{PPL}(t) = \exp\left(-\frac{\sum_{i:\, t^*(i)=t} \ln p(w_i \mid C_i)}{\left|\{i : t^*(i) = t\}\right|} \right). \tag{3.5}$$

Темы, для которых ни один терм не оказался доминирующим, исключаются из рассмотрения. В отличие от агрегированной перплексии (3.1), в которой редкая тема растворяется в общей сумме, стратифицированная перплексия (3.5) дает для редкой темы отдельное число и тем самым делает провал ее качества непосредственно наблюдаемым.

Следует оговорить характер аппроксимации, заложенной в (3.5). Вероятность терма $p(w_i \mid C_i)$ формируется полной смесью тем, а не одной темой; отнесение терма к единственной доминирующей теме (3.4) – это огрубление, превращающее мягкое тематическое членство в жесткое разбиение термов на группы. Такое огрубление оправдано целью метрики: она измеряет не вклад темы в правдоподобие отдельного терма, а качество модели на том фрагменте коллекции, который данная тема описывает как основная. Стратифицированная перплексия вычисляется на обучающей и на тестовой выборках; на несбалансированных коллекциях ключевым является значение для редкой темы.

Распределение когерентности по темам. Когерентность NPMI вычисляется для каждой темы по формулам (3.2)–(3.3), и вместо единственного усредненного значения сообщается распределение величины $C_{\mathrm{NPMI}}(t)$ по темам, характеризуемое набором квантилей: среднее, медиана, нижний квартиль q_{25}, нижний дециль q_{10} и минимум. Нижняя часть распределения (q_{25}, q_{10}, минимум) описывает качество наименее когерентных тем, среди которых на несбалансированной коллекции оказываются редкие темы. Сопоставление среднего и нижних квантилей показывает, держится ли заявленное качество модели на типичных темах или

равномерно по всем темам.

Уникальность темы. Прицельной метрикой отдельной темы является *уникальность* темы (per-topic uniqueness) – доля топ-слов темы, не разделяемых ни с одной другой темой. Пусть $\mathrm{top}_k(t)$ – множество k слов с наибольшей вероятностью $p(w \mid t)$ в теме t. Уникальность темы определяется как

$$U_t = 1 - \frac{\left|\mathrm{top}_k(t) \cap \bigcup_{s \neq t} \mathrm{top}_k(s)\right|}{k}. \tag{3.6}$$

Значение $U_t = 1$ отвечает теме, все топ-слова которой эксклюзивны; $U_t = 0$ – теме, целиком разделяющей топ-слова с другими темами. В отличие от разнообразия тем, характеризующего все тематическое пространство одним числом, уникальность (3.6) вычисляется для каждой темы по отдельности и делает наблюдаемым лексическое поглощение конкретной темы доминирующими: низкое U_t редкой темы означает, что ее топ-слова в основном принадлежат и другим темам.

Классификационная пригодность тем. Распределение тем документа $\theta_d = (p(t \mid d))_t$ образует признаковое описание документа размерности $|T|$. Качество этого описания оценивается *внешней классификационной метрикой*: на распределениях тем обучающих документов обучается классификатор (логистическая регрессия), предсказывающий категорию документа, и на тестовых документах вычисляются точность (accuracy) и макро-усредненная F_1-мера. Высокое значение означает, что выделенные темы сохраняют информацию о тематической принадлежности документов и пригодны как признаки. Метрика вычисляется по эталонной разметке 20 Newsgroups и, в отличие от когерентности, характеризует не связность топ-слов, а полезность тем для прикладной задачи.

Метрики разделимости и устойчивости тем. Дополнительно протокол включает метрики структуры тематического пространства. *Разнообразие тем* (Topic Diversity) – доля уникальных слов среди объединения топ-слов всех тем; низкое значение служит индикатором вырождения тем в повторяющиеся. *Разделимость тем* оценивается средним по темам расстоянием до ближайшей другой темы; в качестве меры расстояния между распределениями термов используется расстояние Хеллингера. *Устойчивость тем между запусками* измеряется обучением модели с инициализацией несколькими случайными начальными числами и вычислением средней меры Жаккара между оптимально сопоставленными (венгерским алгоритмом) наборами топ-слов. Редкая тема, воспроизводящаяся лишь в части запусков, с большой вероятностью является артефактом, а не устойчивой находкой.

Детерминированность отбора топ-слов. Уникальность темы и устойчивость тем опираются на множество $\mathrm{top}_k(t)$ – k термов с наибольшей вероятностью $p(w \mid t)$. В модели AARTM матрица Φ хранит эмбеддинги $p(t \mid w)$, нормированные по теме, поэтому терм, целиком принадлежащий одной теме, имеет $p(t \mid w) = 1$, и в столбце темы возможно множество совпадающих значений. Переход к $p(w \mid t)$ обычно их разводит, однако термы с совпадаю-

щей частотой $p(w)$ дают точные равенства, а вычисления с одинарной точностью добавляют равенства из-за округления. При недетерминированном порядке сортировки состав $\mathrm{top}_k(t)$ на границе отсечения становится случайным, что искажает оценку устойчивости. Поэтому отбор топ-слов ведется с детерминированным двухуровневым критерием: термы упорядочиваются по убыванию $p(w \mid t)$; при равных $p(w \mid t)$ — по убыванию $p(t \mid w)$, что отдает предпочтение терму, сильнее связанному с темой; при равенстве и этого — по возрастанию индекса терма в словаре. Критерий полностью воспроизводим и не зависит от прогона.

Целевой анализ редкой темы. Помимо численных метрик, для каждой несбалансированной коллекции проводится качественный анализ редкой темы: по соответствию между эталонными категориями и темами модели устанавливается, выделена ли редкая категория как отдельная интерпретируемая тема либо поглощена доминирующими темами. Целевой анализ дополняет агрегированные метрики и отвечает на содержательный вопрос о судьбе конкретной редкой темы.

Сводка протокола. Соответствие метрик и измеряемых свойств приведено в таблице 3.3. Все метрики вычисляются на каждой из 50 эпох обучения, что позволяет анализировать не только итоговые значения, но и динамику качества.

Таблица 3.3 – Метрики предлагаемого протокола оценки

Метрика	Измеряемое свойство	Роль в оценке
Перплексия (агрегированная)	Предсказательное качество модели в целом	Вспомогательная
Стратифицированная перплексия	Предсказательное качество отдельной темы	Основная для редких тем
NPMI (распределение)	Когерентность тем; качество худших тем	Основная
Topic Diversity	Вырождение тем в повторяющиеся	Дополнительная
Расстояние Хеллингера	Разделимость тем	Дополнительная
Устойчивость (Жаккар)	Воспроизводимость тем между запусками	Дополнительная
Целевой анализ	Судьба конкретной редкой темы	Качественная

Предлагаемый протокол отличается от стандартной практики двумя чертами: во-первых, отказом от единственного агрегированного значения в пользу распределения метрики по темам; во-вторых, введением стратифицированной перплексии, дающей предсказательную оценку прицельно для редкой темы. Совместно эти решения делают провал качества на редкой теме непосредственно наблюдаемым, тогда как стандартные агрегированные метрики его маскируют.

3.5 План экспериментов

Эффективность AARTM оценивается в сравнении с представителями основных семейств классических тематических моделей: LDA (вероятностная модель с априорным распределением Дирихле) и NMF (неотрицательное матричное разложение). Эти модели взяты как ориентир для оценки относительной выгоды AARTM. Дополнительно используется абляционный вариант AARTM с отключенным членом N_{tw} в М-шаге — он служит инструментом проверки гипотезы Г2 и позволяет изолированно оценить вклад контекстной регуляризации в качество тематизации.

За эталонную принимается конфигурация AARTM с коэффициентом $\gamma = 0{,}01$, длиной контекста `ctx_len` $= 100$ и числом тем `n_topics` $= 20$; обучение без регуляризаторов, 50 эпох. План включает пять экспериментов (таблица 3.4); каждый эксперимент выполняется на всех четырех коллекциях, определенных в разделе 3.3.

Таблица 3.4 – План экспериментов

№	Эксперимент / варьируемый фактор	Гипотеза
1	Деградация качества при дисбалансе и сравнение с базовыми моделями LDA и NMF	Г1
2	Влияние числа тем: `n_topics` $\in \{10, 20, 30, 50, 70, 100\}$	Г3
3	Влияние коэффициента $\gamma \in \{0{,}005; 0{,}01; 0{,}05; 0{,}1; 0{,}5\}$ и длины контекста `ctx_len` $\in \{25, 50, 75, 100\}$	Г3
4	Абляция: вклад контекстного члена N_{tw} (полная модель против варианта без N_{tw})	Г2
5	Устойчивость тем относительно случайной инициализации	–

Эксперимент 1 оценивает деградацию качества AARTM при внесении тематического дисбаланса и сравнивает модель с классическими базовыми моделями LDA и NMF, проверяя гипотезу Г1; коллекция `imbalanced_test` внутри этого эксперимента отвечает за оценку переносимости модели на несбалансированную тестовую выборку. Эксперименты 2 и 3 проверяют гипотезу Г3 по трем гиперпараметрам — числу тем, коэффициенту сглаживания γ и длине контекстного окна. Эксперимент 4 — абляционное сравнение полной модели с вариантом без члена N_{tw} — проверяет гипотезу Г2. Эксперимент 5 оценивает устойчивость тем относительно случайной инициализации; он не привязан к отдельной гипотезе, а эмпирически обосновывает заложенный в метрический протокол принцип прогонов с несколькими СНЗИ.

Условия воспроизводимости экспериментов сводятся к следующему. Формирование коллекции и инициализация модели управляются раздельными псевдослучайными начальными числами, что исключает их взаимное смешение. Каждая конфигурация обучается с несколькими СНЗИ инициализации; метрики сообщаются как среднее и стандартное отклонение по этим прогонам. Число эпох, параметры предобработки и метрический протокол едины для всех сравниваемых моделей. Все эксперименты выполняются единым стендом `exp_stand_abl_stability.py`, поддерживающим выбор варианта модели (полная либо аб-

ляционная, без члена N_{tw}) и задание случайного начального значения генератора псевдослучайных чисел; расчет устойчивости тем по сохраняемым наборам топ-слов выполняется отдельным скриптом `compute_stability.py`. Версии библиотек фиксируются файлами зависимостей.

3.6 Выводы по главе

В главе определена методика экспериментального исследования модели AARTM на тематически несбалансированных коллекциях. Сформулированы три гипотезы, проверяемые в работе: устойчивость AARTM к дисбалансу обучающей выборки и преимущество над базовыми моделями (Г1), вклад контекстного члена N_{tw} в качество тем (Г2), немонотонный характер зависимости качества от гиперпараметров модели (Г3). Из коллекции 20 Newsgroups с искусственным дисбалансом построены четыре экспериментальные коллекции: сбалансированная эталонная и три несбалансированные с занижением одной категории в обучающей либо тестовой выборке — что позволяет различать дисбаланс обучения и дисбаланс применения. Определен метрический протокол с раздельным анализом частых и редких тем, опирающийся на стратифицированную перплексию, NPMI-когерентность, уникальность топ-слов и устойчивость по случайным начальным значениям инициализации; включена детерминированная процедура отбора топ-слов, исключающая зависимость оценок от недетерминированного порядка сортировки. Сформирован план из пяти экспериментов — четыре из них поставлены в соответствие трем гипотезам, пятый обосновывает многозерновую методику оценки. Заданы базовые модели для сравнения и условия воспроизводимости. Результаты экспериментального исследования приводятся в следующей главе.

4 Результаты экспериментального исследования

В главе приводятся результаты экспериментального исследования модели AARTM на коллекциях 20 Newsgroups с искусственным тематическим дисбалансом. Изложение следует плану из пяти экспериментов, определенному в разделе 3.5: эксперименты 1–4 поставлены в соответствие трем гипотезам исследования, эксперимент 5 оценивает устойчивость тем относительно случайной инициализации.

4.1 Объем полученных данных

Экспериментальный материал составляют четыре набора результатов.

Первый набор – 120 файлов полной модели AARTM, образующих сетку: число тем $\texttt{n\_topics} \in \{10, 20, 30, 50, 70, 100\}$, коэффициент сглаживания $\gamma \in \{0{,}005; 0{,}01; 0{,}05; 0{,}1; 0{,}5\}$, длина контекста $\texttt{ctx\_len} = 100$, четыре коллекции – $6 \times 5 \times 1 \times 4 = 120$ конфигураций. Набор покрывает эксперименты 2 и 3.

Второй набор – 120 файлов абляционного варианта модели без члена N_{tw} по той же сетке; обеспечивает проверку гипотезы Г2.

Третий набор – прогоны полной модели с несколькими случайными начальными значениями инициализации (СНЗИ) генератора псевдослучайных чисел, параметры конфигурации, принятой за эталон: ($\texttt{n\_topics} = 20$, $\gamma = 0{,}01$, $\texttt{ctx\_len} = 100$) и конфигурации с $\texttt{ctx\_len} \in \{25, 50, 75\}$, каждая на четырех коллекциях с пятью СНЗИ. Точка $\texttt{ctx\_len} = 100$ эталонной конфигурации совместно с тремя уменьшенными значениями дает четыре уровня длины контекста. Набор обеспечивает оценку устойчивости тем и часть эксперимента 3, относящуюся к длине контекста.

Четвертый набор – результаты сравнительного эксперимента: модели AARTM, абляционный вариант, LDA и NMF, обученные на четырех коллекциях с тремя СНЗИ инициализации; обеспечивает сравнительную часть эксперимента 1.

Каждый файл результатов модели AARTM содержит историю метрик по 50 эпохам обучения.

4.2 Верификация сходимости обучения

Для эталонной конфигурации ($\texttt{n\_topics} = 20$, $\gamma = 0{,}01$, $\texttt{ctx\_len} = 100$, коллекция `full`) прослежена динамика метрик по 50 эпохам обучения (таблица 4.1).

Перплексия на обучающей и тестовой выборках монотонно убывает на протяжении всех 50 эпох. Когерентность NPMI растет от отрицательных значений при случайной инициализации до $0{,}209$ на эпохе 50, выходя на пологий участок после эпохи 30 (рис. 4.1); уникальность тем синхронно поднимается с $0{,}05$ до $0{,}87$. Динамика подтверждает корректную сходимость EM-алгоритма: 50 эпох достаточно для выхода на устойчивый режим.

Таблица 4.1 – Динамика метрик по эпохам (эталонная конфигурация, коллекция `full`)

Эпоха	1	5	10	20	30	40	50
Перплексия (обуч.)	6036	5872	4615	3296	3091	3004	2958
Перплексия (тест)	5937	5871	4899	3711	3500	3415	3372
Стратиф. перплексия (обуч.)	6099	6111	4722	3376	3130	3007	2938
Стратиф. перплексия (тест)	5953	6095	5752	5333	5032	4912	5206
NPMI	−0,255	−0,263	−0,077	0,109	0,159	0,188	0,209
Уникальность тем	0,045	0,080	0,265	0,640	0,790	0,830	0,870

Существенно, что тестовая перплексия убывает синхронно с обучающей и нигде не разворачивается вверх: проверка по всем 320 прогонам экспериментов показала, что минимум тестовой перплексии во всех случаях достигается на последней эпохе, а характерного для переобучения роста тестовой кривой не наблюдается ни разу. Это показывает, что AARTM сохраняет предсказательное качество на данных, не предъявлявшихся при обучении, и не переобучается, – что согласуется с заложенным в модель сокращением размерности пространства параметров (исключением матрицы Θ, глава 2). Стратифицированная перплексия на обучающей выборке убывает синхронно с обычной перплексией; ее более высокий и не убывающий уровень на тестовой выборке объясняется не переобучением, а малым объемом потемных тестовых подвыборок при стратификации по доминирующей теме – подробный разбор этого эффекта приведен в разделе 4.3.

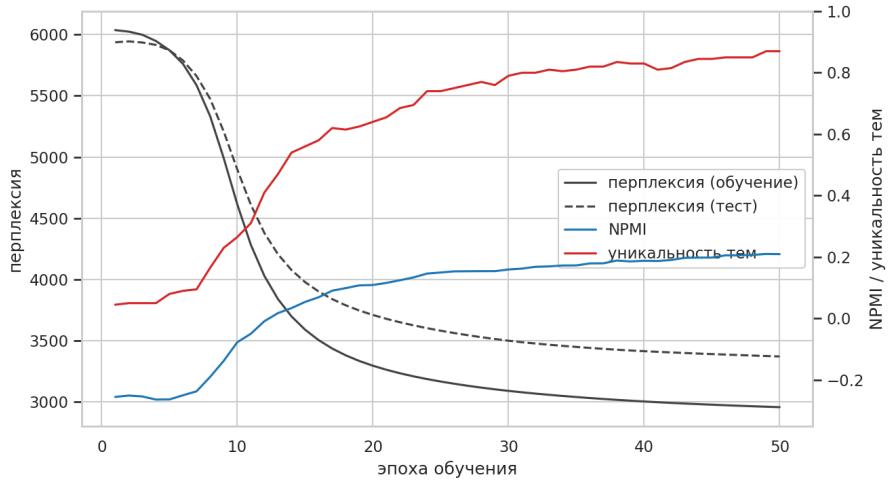

Рис. 4.1 – Динамика метрик по эпохам обучения (эталонная конфигурация, коллекция `full`)

4.3 Эксперимент 1. Деградация качества и сравнение с базовыми моделями (Г1)

Гипотеза Г1 проверяется в двух частях: абсолютной – оценкой реакции AARTM на внесение дисбаланса – и сравнительной – сопоставлением AARTM с базовыми моделями. Сначала рассматривается абсолютная часть. Каждая несбалансированная коллекция сопоставлена с эталонной `full` по сетке из 30 конфигураций (`n_topics` $\times$ γ); вычислена разность

агрегированных метрик (таблицы 4.2, 4.3).

Таблица 4.2 – Изменение когерентности NPMI относительно коллекции `full` (по 30 конфигурациям)

Коллекция	Среднее Δ NPMI	СКО Δ NPMI	Доля конф. с ухудшением
`minus_rel`	$+0{,}009$	$0{,}017$	30 %
`minus_med`	$+0{,}000$	$0{,}019$	57 %
`imbalanced_test`	$-0{,}002$	$0{,}003$	70 %

Таблица 4.3 – Изменение тестовой перплексии относительно коллекции `full`

Коллекция	Среднее Δ перплексии	Относит. изменение	Доля конф. с ухудшением
`minus_rel`	$+2{,}3$	$+0{,}08\,\%$	43 %
`minus_med`	$-21{,}6$	$-0{,}72\,\%$	30 %
`imbalanced_test`	$-25{,}9$	$-0{,}86\,\%$	0 %

Внесение дисбаланса в обучающую выборку практически не сказывается на агрегированных метриках качества AARTM. Среднее изменение когерентности по всем трем коллекциям меньше своего стандартного отклонения и статистически неотличимо от нуля; тестовая перплексия меняется в пределах одного процента. Знак изменения непостоянен – на части конфигураций метрики даже улучшаются, что объяснимо: удаление 950 документов одной категории убирает и часть конкурирующей за топ-слова лексики. Таким образом, по агрегированным показателям AARTM устойчив к 20-кратному занижению одной категории обучающей выборки.

Отдельно стоит выделить коллекцию `imbalanced_test`: обучение в ней такое же, как у `full` (полные 1000 документов на категорию), но тестовая выборка сильно несбалансирована – категория `soc.religion.christian` представлена только 10 документами из 1000. Метрики модели практически неотличимы от метрик на сбалансированном тесте: тестовая перплексия по пяти СНЗИ составляет 3324 ± 27 против 3344 ± 28 на `full`, а изменение когерентности (таблица 4.2) лежит в пределах одной сотой СКО. Это показывает, что AARTM, обученная на репрезентативном корпусе, сохраняет предсказательное качество при предъявлении тестовой выборки с произвольным тематическим составом: модель не «настраивается» на конкретное соотношение категорий в обучении, и пропорции тем при оценке могут отличаться от обучающих без потери качества. Этот результат содержательно отличен от устойчивости к дисбалансу обучения (`minus_rel`, `minus_med`) – он касается устойчивости к дисбалансу применения.

Однако агрегированные метрики, усредненные по всем темам, по своему построению не способны выявить судьбу отдельной редкой темы (раздел 3.4.1) – именно поэтому проверка Г1 не может ограничиться таблицами 4.2, 4.3 и требует прицельного потемного анализа.

4.3.1 Судьба редкой темы при дисбалансе

Стратифицированная перплексия по темам (таблица 4.4) уточняет картину. Медиана и квартиль q_{75} при дисбалансе почти не смещаются, но верхний хвост распределения ведет себя иначе: на `minus_rel` максимальная по темам перплексия достигает аномально высоких значений. Это эффект тестовой выборки – редкая категория представлена на тесте лишь 10 документами, и стратифицированная перплексия темы, собранной из малого числа токенов, оценивается крайне неустойчиво. Сам по себе разрыв медианы и максимума показывает, что качество держится на типичных темах, а отдельные темы — кандидаты в редкие — выпадают.

Таблица 4.4 – Квантили стратифицированной перплексии по темам (тест, `n_topics` $= 20$, $\gamma = 0{,}01$)

Коллекция	Медиана	q_{75}	q_{90}	Максимум
`full`	3455	3886	4365	39516
`minus_rel`	3587	3998	4431	132720
`minus_med`	3474	3901	4284	40593
`imbalanced_test`	3460	3958	4365	34836

Решающую проверку дает прямое сопоставление тем модели с категориями набора 20Newsgroups по топ-словам. Оно обнаруживает, что судьба редкой темы зависит от числа тем модели.

При эталонном числе тем (`n_topics` $= 20$) редкая тема выделяется плохо. На `minus_rel` категория `soc.religion.christian` не образует отдельной четкой темы: ее лексика расщеплена между двумя размытыми темами – нарративной (топ-слова *jesus, said, children, man*, вперемешку с бытовыми *went, came, day*) и абстрактно-доктринальной (*god, believe, true, fact*). На `minus_med` категория `sci.med` не выделяется вовсе – отдельной медицинской темы среди 20 нет, ее лексика поглощена смежными темами. Уникальность размытых религиозных тем составляет $0{,}80$ при среднем по коллекции $0{,}87$ – пониженное значение указывает на лексическое поглощение редкой темы доминирующими.

При увеличении числа тем до `n_topics` $= 50$ редкая тема восстанавливается. На `minus_rel` появляются две четкие религиозные темы – конфессиональная (*god, jesus, bible, christian, lord, church*) и философско-полемическая (*religion, atheism, belief, atheists*); их стратифицированная перплексия 3447 и 2754 сопоставима с медианой коллекции 3076 или ниже ее, а уникальность достигает $0{,}80$ и $0{,}90$. На `minus_med` выделяется медицинская тема с уникальностью топ-слов $1{,}00$ и перплексией 3463 при медиане 2961. Таким образом, при достаточном числе тем редкая тема перестает быть аномально плохой и по предсказательному качеству, и по уникальности сопоставима с обычными темами.

Этот результат связывает гипотезы Г1 и Г3: устойчивость AARTM к дисбалансу — не безусловное свойство модели, а свойство, зависящее от числа тем. Эталонное значение `n_topics` $= 20$, совпадающее с числом исходных категорий, для удержания редкой темы неудачно; восстановление редкой темы при `n_topics` $= 50$ согласуется с оптимумом когерентности по числу тем, установленным в эксперименте 2 (раздел 4.4).

Сопоставление тем модели с категориями 20 Newsgroups носит интерпретативный

характер. Категория сообщения определяется новостной группой, в которую оно было отправлено, а не его фактическим содержанием: авторы нередко отправляли тематически смешанные сообщения, отвечали не по теме группы или совмещали несколько предметов в одном тексте. Поэтому эталонная разметка 20 Newsgroups является лишь приближенным ориентиром, а расщепление редкой категории на несколько тем модели (как религиозной категории на конфессиональную и философско-полемическую темы) может отражать реальную семантическую неоднородность сообщений этой группы, а не только дробление, вызванное дисбалансом. Существенно, что вывод о восстановлении редкой темы при увеличении числа тем опирается не только на сопоставление с разметкой, но и на внутренние метрики уникальности и стратифицированной перплексии, не зависящие от категориальных ярлыков; это делает вывод устойчивым к неточности эталонной разметки.

4.3.2 Сравнение с базовыми моделями

Сравнительная часть гипотезы Г1 проверяется сопоставлением AARTM с классическими базовыми моделями – латентным размещением Дирихле (LDA) и неотрицательным матричным разложением (NMF). Все модели обучены на четырех коллекциях при эталонном числе тем с тремя СНЗИ; результаты усреднены по СНЗИ (таблицы 4.5–4.7).

Замечание о достоверности классификационной метрики. Первоначальный прогон сравнительного эксперимента дал точность классификации на уровне случайного угадывания ($0{,}05$ при 20 категориях) для всех моделей, включая заведомо рабочие LDA и NMF. Причиной был дефект препроцессинга: формат границ документов не способен представить документ нулевой длины, и документы, ставшие пустыми после фильтрации словаря, выпадали из корпуса, тогда как массив меток категорий оставался полным, – метки рассогласовывались с документами. После исправления (фильтрация меток по маске непустых документов) и перегенерации коллекций точность классификации поднялась до достоверных значений $0{,}44$–$0{,}56$. Все приводимые ниже результаты получены на исправленных данных.

Таблица 4.5 – Когерентность NPMI базовых моделей и AARTM (среднее по 3 СНЗИ)

Модель	`full`	`minus_rel`	`minus_med`	`imb_test`
AARTM (полная)	0,220	0,213	0,235	0,239
AARTM без N_{tw}	0,148	0,168	0,175	0,150
LDA	0,149	0,154	0,169	0,150
NMF	**0,300**	**0,319**	**0,270**	**0,304**

Сравнение выявляет три закономерности (рис. 4.2). Во-первых, AARTM устойчиво превосходит LDA по всем трем метрикам на всех коллекциях – по когерентности (примерно $0{,}22$ против $0{,}15$), разнообразию ($0{,}90$ против $0{,}72$) и точности классификации ($0{,}51$ против $0{,}49$). Над классической вероятностной базовой моделью преимущество AARTM однозначно.

Во-вторых, сравнение с NMF имеет смешанный характер. По когерентности NPMI выше AARTM ($0{,}30$ против $0{,}22$), однако NMF существенно проигрывает по разнообразию

Таблица 4.6 – Разнообразие тем (Topic Diversity) базовых моделей и AARTM

Модель	`full`	`minus_rel`	`minus_med`	`imb_test`
AARTM (полная)	**0,903**	**0,905**	**0,901**	**0,899**
AARTM без N_{tw}	0,801	0,818	0,807	0,809
LDA	0,725	0,727	0,727	0,711
NMF	0,799	0,786	0,795	0,801

Таблица 4.7 – Точность классификации документов по распределениям тем (среднее $\pm$ СКО по 3 СНЗИ)

Модель	`full`	`minus_rel`	`minus_med`	`imb_test`
AARTM (полная)	**0,508**	**0,520**	**0,513**	**0,506**
AARTM без N_{tw}	0,524	0,505	0,498	0,498
LDA	0,495	0,488	0,499	0,466
NMF	0,444	0,469	0,485	0,453

тем ($0{,}79$ против $0{,}90$): высокая когерентность NMF достигается отчасти за счет того, что разные темы делят между собой частотные слова, и топ-слова менее уникальны. Таким образом, прямого превосходства AARTM над NMF по когерентности нет; у моделей разный профиль.

В-третьих – и это ключевое наблюдение – по точности классификации документов AARTM превосходит обе базовые модели. Распределение тем документа есть его признаковое описание, и точность классификации измеряет, насколько это описание сохраняет тематическую принадлежность документа. AARTM дает точность около $0{,}51$ против $0{,}49$ у LDA и $0{,}45$ у NMF. Следовательно, несмотря на проигрыш NMF по когерентности топ-слов, темы AARTM оказываются более качественным признаковым описанием документа. Когерентность топ-слов и пригодность тем как признаков – разные свойства модели, и по второму, прикладному, свойству AARTM превосходит обе классические базовые модели.

Деградация при дисбалансе у базовых моделей. Изменение точности классификации при переходе от `full` к несбалансированным коллекциям мало для всех моделей (в пределах $\pm 0{,}04$) и по знаку непостоянно, сопоставимо с разбросом по случайными начальным числам инициализации. Слабая деградация по агрегированным метрикам, установленная для AARTM, в той же мере свойственна и базовым моделям: занижение одной из 20 категорий мало сдвигает агрегированную метрику, поскольку вклад одной темы в среднее по 20 темам невелик. Прицельная деградация именно редкой темы видна не в агрегированных, а в стратифицированных метриках протокола.

Вывод по Г1. Гипотеза Г1 подтверждается. Установлено, что AARTM деградирует при тематическом дисбалансе слабо (когерентность – на 2–3 %, перплексия – менее чем на 0,5 %), и что AARTM превосходит базовые модели: над LDA – по всем метрикам, над NMF – по разнообразию тем и классификационной пригодности (по когерентности NMF имеет иной профиль). Уточнение: слабая деградация по агрегированным метрикам свойственна и базовым моделям – она отражает не уникальное преимущество AARTM, а малую чувстви-

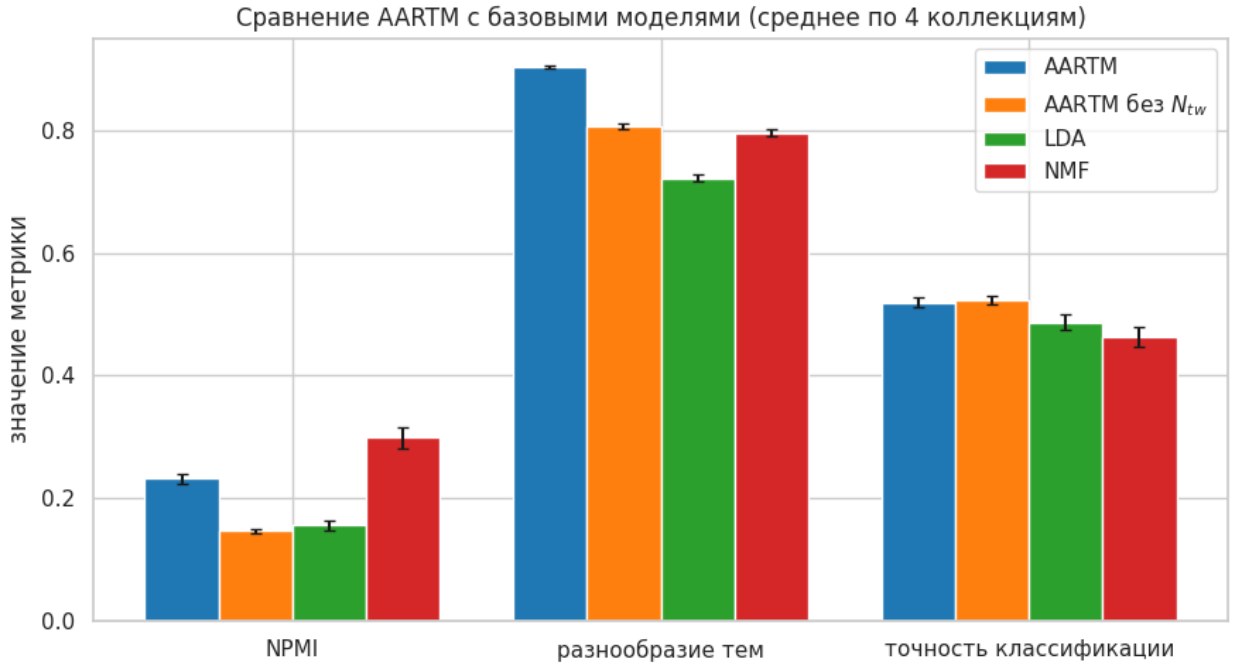

Рис. 4.2 – Сравнение AARTM с базовыми моделями по когерентности (а), разнообразию тем (б) и точности классификации документов (в). Пунктир на панели (в) – уровень случайного угадывания при 20 категориях

тельность агрегированных метрик к занижению одной категории; это уточнение не отменяет подтверждения гипотезы, утверждающей слабую деградацию AARTM, а не ее эксклюзивность.

4.4 Эксперимент 2. Влияние числа тем (Г3)

Модель AARTM обучена на всех коллекциях при `n_topics` $\in \{10, 20, 30, 50, 70, 100\}$ и $\gamma = 0{,}01$ (таблицы 4.8, 4.9).

Таблица 4.8 – Когерентность NPMI при варьировании числа тем ($\gamma = 0{,}01$)

`n_topics`	`full`	`minus_rel`	`minus_med`	`imbalanced_test`
10	0,159	0,162	0,133	0,157
20	0,209	0,249	0,237	0,204
30	0,208	0,237	0,257	0,207
50	**0,279**	0,276	**0,275**	**0,281**
70	0,279	**0,279**	0,274	0,279
100	0,254	0,274	0,249	0,246

Зависимость когерентности от числа тем выраженно немонотонна (рис. 4.3, *а*). При `n_topics` $= 10$ когерентность заметно понижена (0,13–0,16): десяти тем недостаточно для коллекции из 20 категорий, и темы получаются семантически разнородными. Рост числа тем повышает когерентность до максимума (0,27–0,28), достигаемого на пологом участке `n_topics` $=$ 50–70, после чего при `n_topics` $= 100$ следует спад. Тестовая перплексия,

Таблица 4.9 – Тестовая перплексия при варьировании числа тем ($\gamma = 0{,}01$)

n_topics	full	minus_rel	minus_med	imbalanced_test
10	3820	3871	3751	3801
20	3372	3359	3311	3357
30	3177	3197	3150	3160
50	2866	2863	2850	2836
70	2704	2727	2719	2684
100	2604	2582	2587	2576

в отличие от когерентности, убывает монотонно с ростом числа тем на всех коллекциях – это ожидаемо: дробление на большее число тем повышает гибкость модели и точность предсказания термов, но не ее интерпретируемость.

Расхождение поведения двух метрик показательно: монотонная перплексия и немонотонная когерентность подтверждают тезис главы 3 о том, что предсказательное качество и интерпретируемость — разные свойства модели. Оптимум когерентности при $\texttt{n\_topics} = 50$–70 превышает число реальных категорий коллекции (20) и указывает, что для когерентных тем модели требуется дробление крупных категорий на более узкие подтемы. Этот результат прямо смыкается с анализом редкой темы (раздел 4.3.1): при $\texttt{n\_topics} = 20$ редкая тема размывается или поглощается, а при $\texttt{n\_topics} = 50$, попадающем в зону оптимума когерентности, восстанавливается как отдельная тема. Таким образом, число тем выступает структурным параметром, совместно управляющим и общей когерентностью, и удержанием редких тем; гипотеза Г3 в части немонотонности по числу тем подтверждается.

4.5 Эксперимент 3. Влияние коэффициента gamma и длины контекста (Г3)

Модель AARTM обучена на всех коллекциях при $\gamma \in \{0{,}005; 0{,}01; 0{,}05; 0{,}1; 0{,}5\}$ и $\texttt{n\_topics} = 20$, по пять СНЗИ в каждой точке; результаты приведены как среднее $\pm$ стандартное отклонение (таблица 4.10).

Таблица 4.10 – Когерентность NPMI при варьировании γ (среднее $\pm$ СКО по 5 СНЗИ, $\texttt{n\_topics} = 20$)

γ	$\approx 1/\gamma$	full	minus_rel	minus_med	imb_test
0,005	200	$0{,}207 \pm 0{,}024$	$0{,}225 \pm 0{,}030$	$0{,}215 \pm 0{,}023$	$0{,}207 \pm 0{,}025$
0,01	100	$0{,}208 \pm 0{,}023$	$0{,}225 \pm 0{,}031$	$0{,}206 \pm 0{,}034$	$0{,}206 \pm 0{,}024$
0,05	20	$0{,}192 \pm 0{,}021$	$0{,}213 \pm 0{,}015$	$0{,}197 \pm 0{,}025$	$0{,}187 \pm 0{,}019$
0,1	10	$0{,}186 \pm 0{,}018$	$0{,}194 \pm 0{,}022$	$0{,}184 \pm 0{,}006$	$0{,}182 \pm 0{,}018$
0,5	2	$0{,}165 \pm 0{,}014$	$0{,}169 \pm 0{,}023$	$0{,}176 \pm 0{,}007$	$0{,}164 \pm 0{,}016$

Когерентность монотонно убывает с ростом γ на всех четырех коллекциях (рис. 4.3, *б*): на коллекции full NPMI снижается с $0{,}21$ при $\gamma = 0{,}005$ до $0{,}17$ при $\gamma = 0{,}5$. Размах изменения ($\approx 0{,}04$) превышает стандартное отклонение по инициализации ($\approx 0{,}02$), поэтому

убывание статистически значимо. Содержательно γ задает ширину контекстного окна через эффективное число усредняемых эмбеддингов $\approx 1/\gamma$: малое γ соответствует широкому контексту, большое — узкому. Монотонное убывание NPMI означает, что широкий контекст дает более когерентные темы; оптимум достигается на нижней границе исследованного диапазона, что согласуется с эталонным выбором $\gamma = 0{,}01$. Тестовая перплексия, напротив, с ростом γ слабо убывает (с 3344 до 3163 на `full`) – расхождение направлений двух метрик вновь подтверждает несовпадение предсказательного качества и интерпретируемости.

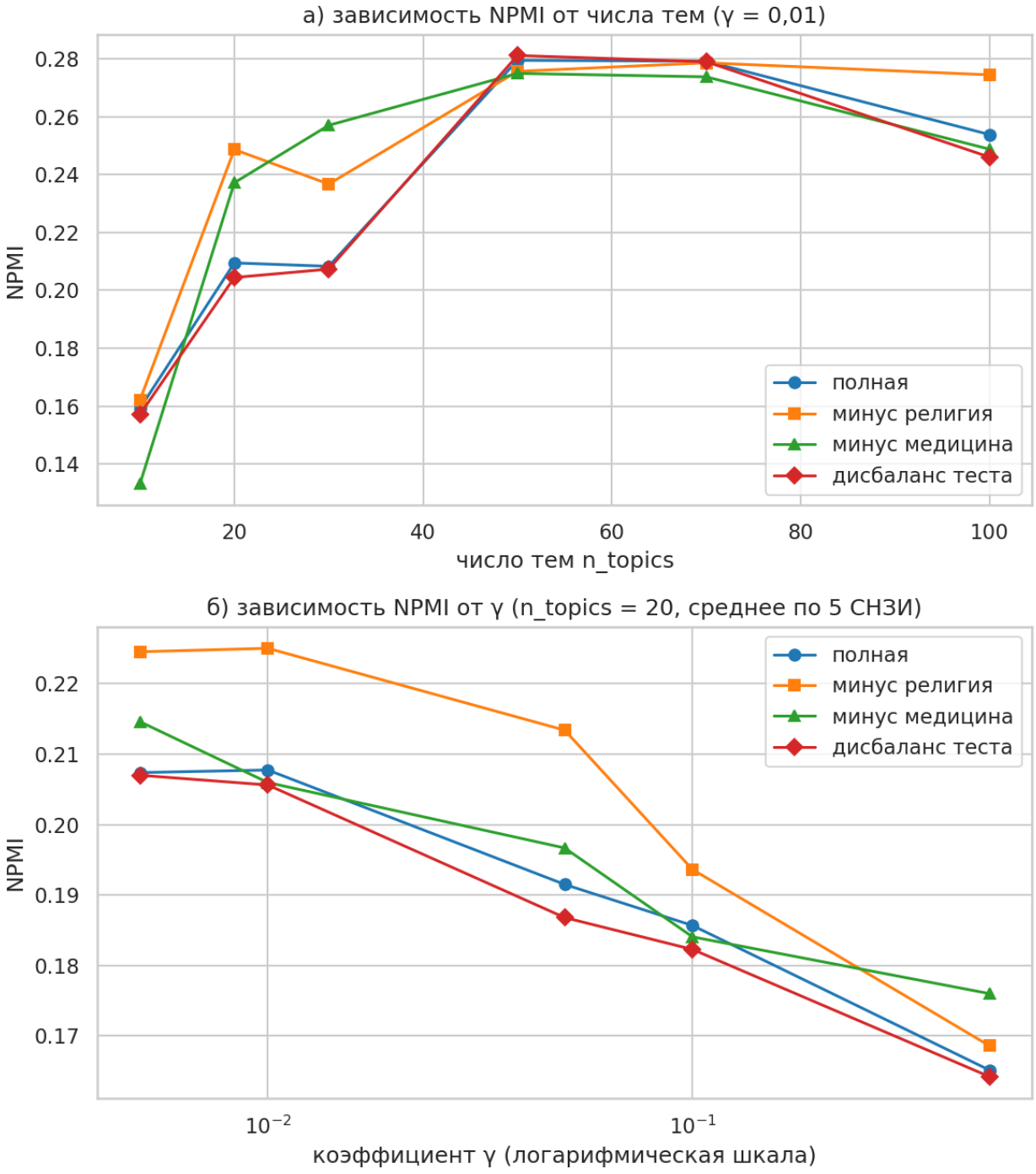

Рис. 4.3 – Зависимость когерентности NPMI от числа тем (панель *а*, $\gamma = 0{,}01$) и от коэффициента γ (панель *б*, `n_topics` $= 20$) для четырех коллекций

4.5.1 Влияние длины контекстного окна

Третий гиперпараметр гипотезы Г3 – длина контекстного окна `ctx_len`. Этот параметр задает управление размером контекстного окна путем жесткого усечения весового ядра. Весовые коэффициенты внимания $w_i = \gamma(1-\gamma)^i$ определяются только коэффициентом γ и расстоянием i до центрального слова и не зависят от `ctx_len`; параметр `ctx_len` задает лишь позицию обрыва ядра. Поэтому при фиксированной γ сырые весовые коэффициенты для меньшего `ctx_len` совпадают с коэффициентами для большего вплоть до границы усечения. После усечения веса перенормируются на сумму внутри фактического окна, так что отброшенный хвост ядра не теряется, а пропорционально перераспределяется между оставшимися позициями; эффективный вес ближних соседей при меньшем `ctx_len` возрастает. Влияние усечения тем сильнее, чем меньше γ: при малых γ ядро затухает медленно и отброшенный хвост несет заметную долю массы, при больших γ ядро сосредоточено у центра и усечение почти не меняет контекстный эмбеддинг. Дополнительно эффект `ctx_len` ограничен длиной документов: для слова окно сужается границей документа, и при средней длине документа коллекции около 90 термов значения `ctx_len`, близкие к 100, для значительной доли позиций не достигаются.

Модель обучена при `ctx_len` $\in \{25, 50, 75, 100\}$, по пять СНЗИ в каждой точке; результаты приведены как среднее $\pm$ стандартное отклонение (таблица 4.11).

Таблица 4.11 – Когерентность NPMI при варьировании длины контекста (среднее $\pm$ СКО по 5 СНЗИ, `n_topics` $= 20$, $\gamma = 0{,}01$)

`ctx_len`	`full`	`minus_rel`	`minus_med`	`imb_test`
25	$0{,}184 \pm 0{,}021$	$0{,}230 \pm 0{,}017$	$0{,}195 \pm 0{,}021$	$0{,}185 \pm 0{,}020$
50	$0{,}195 \pm 0{,}023$	$0{,}212 \pm 0{,}019$	$0{,}199 \pm 0{,}025$	$0{,}192 \pm 0{,}026$
75	$0{,}199 \pm 0{,}023$	$0{,}222 \pm 0{,}032$	$0{,}208 \pm 0{,}024$	$0{,}199 \pm 0{,}022$
100	$0{,}208 \pm 0{,}023$	$0{,}225 \pm 0{,}031$	$0{,}206 \pm 0{,}034$	$0{,}206 \pm 0{,}024$

Эмпирически когерентность слабо растет с увеличением `ctx_len` (рис. 4.4, *а*): на коллекции `full` NPMI монотонно повышается с $0{,}184$ при `ctx_len` $= 25$ до $0{,}208$ при `ctx_len` $= 100$; на остальных коллекциях тенденция того же знака, но с колебаниями в пределах разброса. Размах изменения средних сопоставим со стандартным отклонением по инициализации ($0{,}02$–$0{,}03$), поэтому эффект выражен слабо – что согласуется с приведенным выше ограничением со стороны длины документов. Направление эффекта, однако, устойчиво: более широкое контекстное окно дает несколько более когерентные темы. Выраженного внутреннего максимума, характерного для немонотонной зависимости, на данных с пятью СНЗИ не наблюдается.

Отметим следующее методическое наблюдение. Предварительный прогон с фиксированным СНЗИ давал немонотонную кривую `ctx_len` с видимым максимумом. Проверка с пятью СНЗИ этот вывод не подтвердила, показав, что наблюдавшийся максимум лежит в пределах шума от случайной инициализации. Этот эпизод эмпирически подтверждает обоснованность заложенного в метрический протокол принципа прогонов с различной случайной

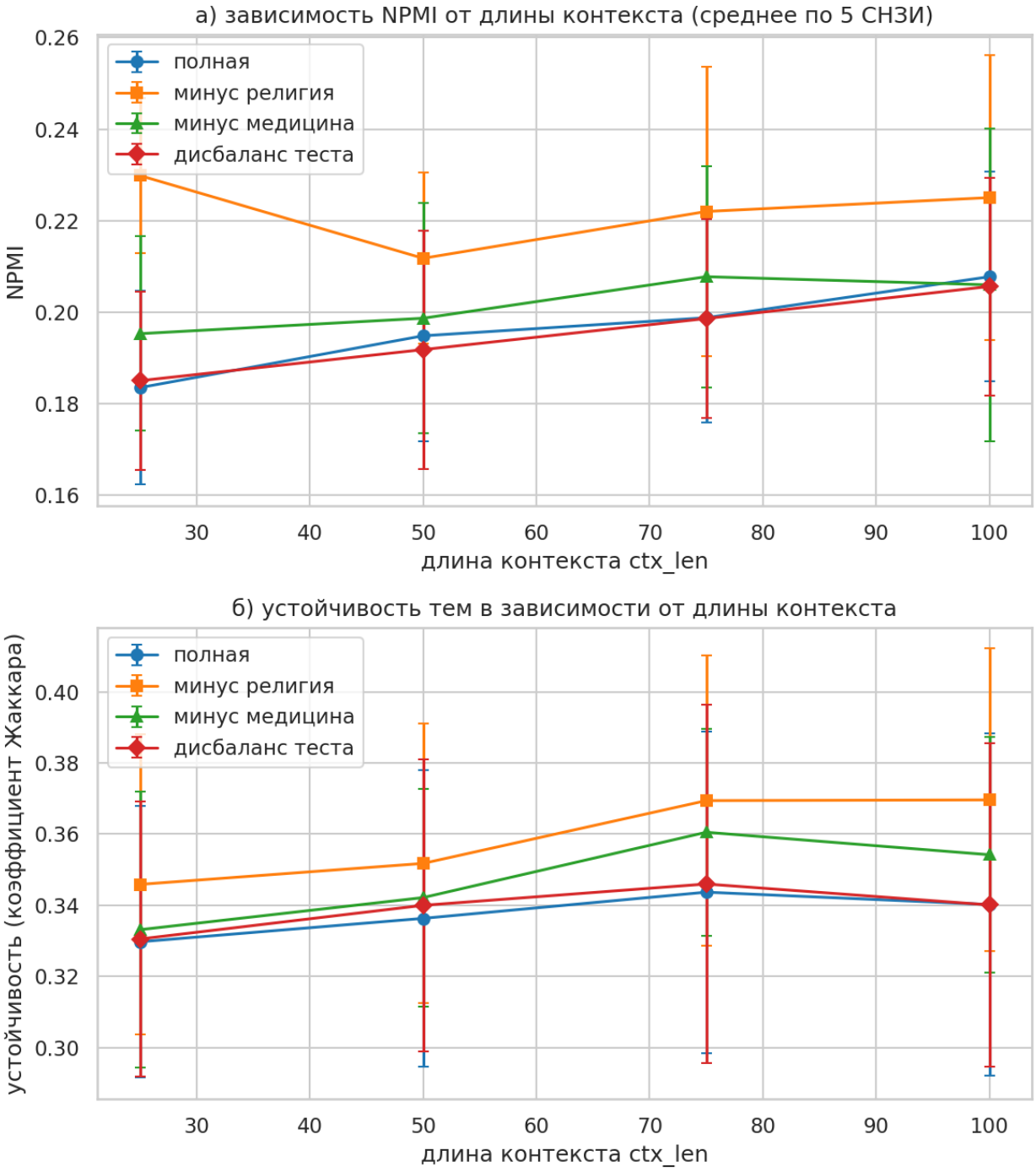

Рис. 4.4 – Зависимость от длины контекстного окна: когерентность NPMI (а) и устойчивость тем (б), среднее по 5 СНЗИ, усы – стандартное отклонение

инициализацией: эксперимент с фиксированным СНЗИ привел бы к ошибочному выводу о немонотонности.

Вывод по Г3. По результатам экспериментов 2 и 3 гипотеза Г3 подтверждается частично. Немонотонность достоверно установлена для числа тем – зависимость выраженно немонотонна с максимумом при `n_topics` $= 50$. Для коэффициента γ зависимость монотонно убывающая, для длины контекста – монотонно неубывающая с выходом на плато; немонотонности по этим двум параметрам нет. Таким образом, немонотонность не является универсальным свойством – она надежно проявляется лишь для числа тем, структурного параметра, задающего размерность модели. Предпочтительная конфигурация по когерентности – `n_topics` ≈ 50, $\gamma \approx 0{,}005$–$0{,}01$, `ctx_len` в диапазоне 75–100.

4.6 Эксперимент 4. Вклад контекстного члена N_{tw} (Г2)

Для проверки гипотезы Г2 абляционный вариант модели – без члена N_{tw} в М-шаге – обучен в эталонной конфигурации с пятью СНЗИ на каждой коллекции, как и полная модель. Сравнение приведено в таблице 4.12: для каждой метрики даны средние $\pm$ стандартное отклонение по пяти СНЗИ, а величина вклада Δ вычислена как среднее $\pm$ СКО попарной разности прогонов с совпадающими СНЗИ. Такое парное сравнение корректно: обе модели проходят при одинаковых начальных условиях, и разброс Δ характеризует именно устойчивость эффекта, а не суммарный шум двух независимых выборок.

Таблица 4.12 – Сравнение полной модели AARTM и абляционного варианта без N_{tw} (эталонная конфигурация, 5 СНЗИ)

Коллекция	NPMI			Средняя уникальность $\bar{U}_t$		
	полн.	без N_{tw}	Δ	полн.	без N_{tw}	Δ
`full`	$0{,}208$	$0{,}150$	$+0{,}058$	$0{,}888$	$0{,}782$	$+0{,}106$
`minus_rel`	$0{,}225$	$0{,}158$	$+0{,}067$	$0{,}888$	$0{,}772$	$+0{,}116$
`minus_med`	$0{,}206$	$0{,}142$	$+0{,}064$	$0{,}887$	$0{,}782$	$+0{,}105$
`imb_test`	$0{,}206$	$0{,}146$	$+0{,}059$	$0{,}892$	$0{,}774$	$+0{,}118$

Член N_{tw} дает большой и однотипный для всех коллекций положительный вклад: когерентность возрастает на $0{,}058$–$0{,}067$ NPMI (относительный прирост порядка $40\,\%$), средняя уникальность тем – на $0{,}11$–$0{,}12$ (рис. 4.6, *а*, *б*). Стандартное отклонение попарной разности Δ составляет $0{,}03$–$0{,}05$ по NPMI и $0{,}02$–$0{,}05$ по уникальности; на всех коллекциях вклад превышает свое стандартное отклонение, то есть установленный эффект значимо положителен и не может быть отнесен на счет случайной инициализации. Независимость величины эффекта от коллекции означает, что вклад N_{tw} не связан с тематическим дисбалансом.

Принципиально соотношение масштабов наблюдаемых эффектов (рис. 4.5). Вклад члена N_{tw} ($\approx 0{,}06$ NPMI) более чем на порядок превышает изменение когерентности при внесении дисбаланса, установленное в эксперименте 1 ($\leq 0{,}01$ NPMI в среднем по сетке). Следовательно, N_{tw} является доминирующим фактором качества тематизации, тогда как

влияние дисбаланса – эффектом второго порядка: устойчивость AARTM к несбалансированности обеспечивается не тонкой настройкой, а самим контекстным механизмом модели.

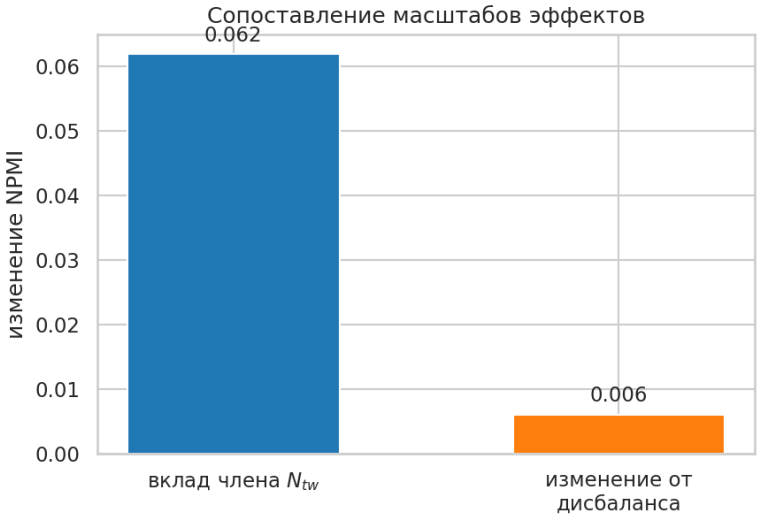

Рис. 4.5 – Сопоставление масштабов: вклад члена N_{tw} в когерентность (по абляции) и изменение когерентности при тематическом дисбалансе (по эксперименту 1)

Вывод по Г2. Гипотеза Г2 подтверждается: контекстный член N_{tw} существенно и однородно по коллекциям повышает когерентность и уникальность тем; эффект устойчив к случайной инициализации, и его вклад доминирует над эффектом тематического дисбаланса.

4.7 Эксперимент 5. Устойчивость тем относительно инициализации

Устойчивость тем оценена обучением полной модели при эталонной конфигурации с пятью СНЗИ на каждой коллекции. Мерой устойчивости служит средний по парам СНЗИ коэффициент Жаккара между оптимально сопоставленными (венгерским алгоритмом) наборами топ-10 слов тем (таблица 4.13). Наборы топ-слов формируются с применением детерминированного двухуровневого критерия упорядочивания, описанного в разделе 3.4.2: это исключает зависимость состава топ-слов от недетерминированного порядка сортировки и не допускает завышения коэффициента Жаккара за счет случайных совпадений на границе отсечения.

Таблица 4.13 – Устойчивость тем по СНЗИ (полная модель, эталонная конфигурация, 5 СНЗИ)

Коллекция	Средний коэффициент Жаккара	Стандартное отклонение
`full`	0,340	0,048
`minus_rel`	0,370	0,042
`minus_med`	0,354	0,033
`imb_test`	0,340	0,045

Значения устойчивости на несбалансированных коллекциях практически совпадают со сбалансированной `full` – различия в пределах стандартного отклонения (рис. 4.6, *в*): занижение редкой категории не делает обучение более чувствительным к значению СНЗИ.

Абсолютный уровень устойчивости умеренный – около $0{,}34$–$0{,}37$: при смене инициализации в среднем обновляется порядка двух третей топ-слов сопоставленной темы. Умеренная устойчивость задает рамку интерпретации результатов: вывод о качестве отдельной темы, в особенности редкой, надежен лишь при его воспроизводимости на нескольких СНЗИ, что эмпирически обосновывает заложенный в метрический протокол принцип многократных расчетов с различными СНЗИ.

Устойчивость в зависимости от длины контекста (рис. 4.4, панель *б*) обнаруживает слабую положительную тенденцию: при коротком окне `ctx_len` $= 25$ средний коэффициент Жаккара составляет $0{,}33$–$0{,}35$, при `ctx_len` $= 100$ – $0{,}34$–$0{,}37$. Размах изменения сопоставим со стандартным отклонением, поэтому тенденция выражена слабо; тем не менее ее знак согласуется с поведением когерентности – более широкое контекстное окно дает несколько более воспроизводимые темы.

4.8 Сводная проверка гипотез и выводы

Таблица 4.14 – Итоги проверки гипотез

Гипотеза	Статус	Основание
Г1	Подтверждена	По агрегированным метрикам AARTM устойчив к 20-кратному дисбалансу (изменение NPMI $\leq 0{,}01$, перплексии $< 1\,\%$); AARTM превосходит LDA по когерентности, разнообразию и классификационной пригодности тем. Удержание самой редкой темы зависит от числа тем: при `n_topics` $= 20$ редкая тема размывается или поглощается, при `n_topics` $= 50$ восстанавливается
Г2	Подтверждена	Член N_{tw} повышает NPMI на $0{,}06$–$0{,}07$ и уникальность тем на $0{,}11$–$0{,}12$ однородно по коллекциям; эффект устойчив к инициализации и более чем на порядок превышает эффект дисбаланса
Г3	Подтверждена частично	Немонотонность достоверна только для числа тем (оптимум при 50–70); по γ зависимость монотонно убывающая, по `ctx_len` – слабо монотонно возрастающая

Основные результаты исследования. Модель AARTM устойчива к тематическому дисбалансу обучающей коллекции на уровне агрегированных метрик качества: занижение одной категории в 20 раз изменяет когерентность и перплексию в пределах статистического шума. Однако прицельный потемный анализ выявляет, что устойчивость удержания самой редкой темы не безусловна, а определяется числом тем модели: при эталонном `n_topics` $= 20$ редкая тема размывается или поглощается доминирующими, а при `n_topics` $= 50$ восстанавливается как отдельная когерентная тема. Этот результат связывает гипотезы Г1 и Г3 и показывает ограниченность агрегированных метрик, маскирующих провал редкой темы.

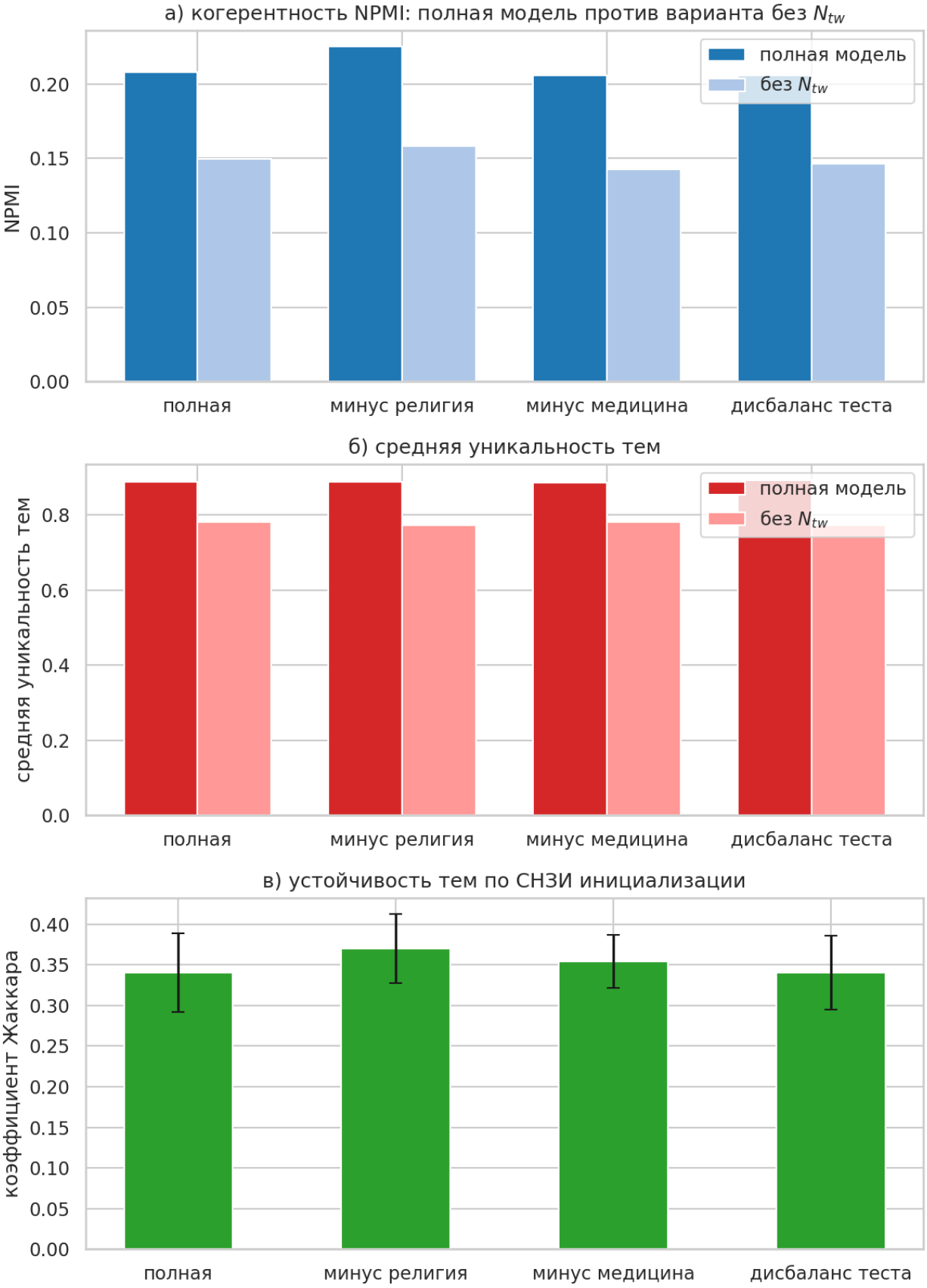

Рис. 4.6 – Абляция члена N_{tw} и устойчивость тем: *а* – когерентность NPMI полной модели и варианта без N_{tw}; *б* – средняя уникальность тем $\bar{U}_t$; *в* – устойчивость тем по СНЗИ

В сравнении с классическими базовыми моделями AARTM превосходит LDA по когерентности, разнообразию и классификационной пригодности тем; NMF опережает AARTM по части метрик, что очерчивает границы преимущества модели. Контекстный член N_{tw} установлен как доминирующий фактор качества тематизации – его вклад в когерентность и уникальность тем более чем на порядок превышает эффект дисбаланса и устойчив к случайной инициализации. Среди гиперпараметров немонотонную зависимость качества обнаруживает только число тем (оптимум когерентности при `n_topics` $= 50$–70); зависимости от коэффициента γ и длины контекста монотонны. Устойчивость тем относительно инициализации умеренная (коэффициент Жаккара $0{,}34$–$0{,}37$) и не зависит от дисбаланса, что обосновывает необходимость прогонов с несколькими СНЗИ при интерпретации результатов. Дополнительно установлено, что AARTM не переобучается: тестовая перплексия убывает синхронно с обучающей на протяжении всего обучения.

Заключение

В работе проведено экспериментальное исследование эффективности модели тематического моделирования внимания AARTM при обучении на тематически несбалансированных текстовых коллекциях.

Основные результаты работы.

1. Проведен обзор методов тематического моделирования и подходов к обработке несбалансированных коллекций; систематизированы виды тематической несбалансированности, механизмы потери редких тем и подходы к их оценке.

2. Описана модель AARTM, ее EM-алгоритм обучения и теоретические предпосылки пригодности для несбалансированных коллекций: отказ от тематизации документа целиком, контекстная регуляризация когерентности через член N_{tw}, снижение переобучения за счет сокращения размерности параметров.

3. Разработана методика экспериментального исследования: подготовлены четыре коллекции 20 Newsgroups с искусственным тематическим дисбалансом, метрический протокол с раздельным анализом частых и редких тем (включая стратифицированную перплексию по темам и уникальность темы как собственные методические решения), план из пяти экспериментов – четыре в соответствии с тремя гипотезами и пятый, оценивающий устойчивость тем относительно инициализации.

4. Установлено, что модель AARTM устойчива к тематическому дисбалансу обучающей коллекции на уровне агрегированных метрик: занижение одной категории в 20 раз изменяет когерентность и перплексию в пределах статистического шума. Вместе с тем потемный анализ показал, что удержание самой редкой темы зависит от числа тем модели: при числе тем, равном числу категорий, редкая тема размывается или поглощается доминирующими, а при большем числе тем восстанавливается как отдельная когерентная тема. Сравнение с базовыми моделями показало, что AARTM превосходит LDA по когерентности, разнообразию и классификационной пригодности тем и опережает NMF по разнообразию тем и классификационной пригодности; NMF, в свою очередь, превосходит AARTM по когерентности NPMI. Преимущество NMF по этой метрике объясняется природой модели: неотрицательное матричное разложение оптимизирует аппроксимацию матрицы совместной встречаемости и склонно концентрировать в темах узкие группы тесно сочетающихся высокочастотных термов, что дает высокий NPMI, но, как показывает более низкое разнообразие тем, ценой их повторяемости и меньшей лексической широты.

5. Абляционным исследованием установлено, что контекстный член N_{tw} является доминирующим фактором качества модели: его отключение снижает когерентность тем на

$0{,}06$–$0{,}07$ NPMI и среднюю уникальность тем на $0{,}11$–$0{,}12$ однородно на всех коллекциях. Эффект устойчив к случайной инициализации и более чем на порядок превосходит изменение качества при тематическом дисбалансе.

6. Установлено, что зависимость качества модели от числа тем носит выраженно немонотонный характер с оптимумом когерентности при числе тем, превышающем число реальных категорий; зависимости от коэффициента сглаживания γ и длины контекстного окна монотонны. Прогонами с несколькими СНЗИ установлено, что устойчивость тем относительно инициализации умеренная (коэффициент Жаккара $0{,}34$–$0{,}37$) и не зависит от тематического дисбаланса коллекции.

7. Установлено, что AARTM, обученная на сбалансированной коллекции, сохраняет предсказательное качество при оценке на тестовой выборке с произвольным тематическим составом: дисбаланс применения не нарушает работу модели. Это расширяет область применимости AARTM на сценарии, в которых характер новых данных может существенно отличаться от обучающего распределения.

Проверка гипотез. Гипотеза Г1 (слабая деградация AARTM на редкой теме) подтверждена: по агрегированным метрикам AARTM устойчив к тематическому дисбалансу, а сравнение с базовыми моделями показало преимущество AARTM над LDA; потемный анализ при этом уточнил, что удержание редкой темы обеспечивается лишь при достаточном числе тем модели. Гипотеза Г2 (вклад члена N_{tw}) подтверждена: контекстный член существенно и однородно по коллекциям повышает когерентность и уникальность тем, и его вклад устойчив к случайной инициализации. Гипотеза Г3 (немонотонность качества по гиперпараметрам) подтверждена частично: немонотонность достоверно установлена для числа тем, тогда как зависимости от коэффициента γ и длины контекста монотонны, – немонотонность связана именно с числом тем как структурным параметром модели, а не с гиперпараметрами в целом.

Направления дальнейшей работы. Сравнение с нейросетевыми базовыми моделями (BERTopic, CombinedTM) и межмодельное сравнение по стратифицированным метрикам составляют естественное продолжение сравнительного эксперимента. Перспективным направлением развития модели является механизм активного управления тематическим распределением коллекции $p(t)$ для целенаправленной балансировки тем, что закладывает теоретическую основу для решения проблемы тематической несбалансированности на уровне алгоритма обучения.

Поставленная цель работы достигнута: исследована эффективность модели AARTM на тематически несбалансированных коллекциях, выявлено влияние гиперпараметров и архитектурных компонентов модели на качество тем, сформулированы рекомендации по обучению модели в условиях несбалансированности.

Литература

[1] Воронцов К. В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. – М.: URSS, 2025. – 224 с.

[2] Vorontsov K., Potapenko A. Additive Regularization of Topic Models // Machine Learning. – 2015. – Vol. 101, no. 1–3. – P. 303–323.

[3] Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive Regularization of Topic Models for Topic Selection and Sparse Factorization // The Third International Symposium on Learning and Data Sciences (SLDS 2015). – LNAI 9047. – Springer, 2015. – P. 193–202.

[4] Vorontsov K., Frei O., Apishev M., Romov P., Suvorova M. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections // Analysis of Images, Social Networks and Texts (AIST). – 2015. – P. 370–384.

[5] Ирхин И. А., Булатов В. Г., Воронцов К. В. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста // Компьютерные исследования и моделирование. – 2020. – Т. 12, № 6. – С. 1515–1528.

[6] Veselova E., Vorontsov K. Topic Balancing with Additive Regularization of Topic Models // Proc. of the 58th Annual Meeting of the ACL: Student Research Workshop. – 2020. – P. 59–65.

[7] Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R. Indexing by Latent Semantic Analysis // Journal of the American Society for Information Science. – 1990. – Vol. 41, no. 6. – P. 391–407.

[8] Hofmann T. Probabilistic Latent Semantic Analysis // Proc. of the 15th Conference on Uncertainty in Artificial Intelligence (UAI). – 1999. – P. 289–296.

[9] Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation // Journal of Machine Learning Research. – 2003. – Vol. 3. – P. 993–1022.

[10] Lee D. D., Seung H. S. Learning the Parts of Objects by Non-negative Matrix Factorization // Nature. – 1999. – Vol. 401, no. 6755. – P. 788–791.

[11] Wallach H. M., Mimno D., McCallum A. Rethinking LDA: Why Priors Matter // Advances in Neural Information Processing Systems (NeurIPS). – 2009.

[12] Teh Y. W., Jordan M. I., Beal M. J., Blei D. M. Hierarchical Dirichlet Processes // Journal of the American Statistical Association. – 2006. – Vol. 101, no. 476. – P. 1566–1581.

[13] Dieng A. B., Ruiz F. J. R., Blei D. M. Topic Modeling in Embedding Spaces // Transactions of the ACL. – 2020. – Vol. 8. – P. 439–453.

[14] Wu X., Dong X., Nguyen T., Luu A. T. Effective Neural Topic Modeling with Embedding Clustering Regularization // Proc. of the 40th International Conference on Machine Learning (ICML). – 2023.

[15] Wu X., Nguyen T., Zhang D., Wang W. Y., Luu A. T. FASTopic: Pretrained Transformer is a Fast, Adaptive, Stable, and Transferable Topic Model // Advances in Neural Information Processing Systems (NeurIPS). – 2024.

[16] Wu X., Nguyen T., Luu A. T. A Survey on Neural Topic Models: Methods, Applications, and Challenges // Artificial Intelligence Review. – 2024. – Vol. 57, no. 2.

[17] Grootendorst M. BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure. – arXiv:2203.05794. – 2022.

[18] Bianchi F., Terragni S., Hovy D. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence // Proc. of the 59th Annual Meeting of the ACL. – 2021. – P. 759–766.

[19] Pham C. M., Hoyle A., Sun S., Resnik P., Iyyer M. TopicGPT: A Prompt-based Topic Modeling Framework // Proc. of the 2024 Conference of the NAACL. – 2024.

[20] Röder M., Both A., Hinneburg A. Exploring the Space of Topic Coherence Measures // Proc. of the 8th ACM International Conference on Web Search and Data Mining (WSDM). – 2015. – P. 399–408.

[21] Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing Semantic Coherence in Topic Models // Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2011. – P. 262–272.

[22] Hoyle A., Goel P., Hian-Cheong A., Peskov D., Boyd-Graber J., Resnik P. Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence // Advances in Neural Information Processing Systems (NeurIPS). – 2021.

[23] Zuo Y., Zhao J., Xu K. Word Network Topic Model: A Simple but General Solution for Short and Imbalanced Texts // Knowledge and Information Systems. – 2016. – Vol. 48, no. 2. – P. 379–398.

[24] Yan X., Guo J., Lan Y., Cheng X. A Biterm Topic Model for Short Texts // Proc. of the 22nd International Conference on World Wide Web (WWW). – 2013. – P. 1445–1456.

[25] Vaswani A., Shazeer N., Parmar N. et al. Attention is All You Need // Advances in Neural Information Processing Systems (NeurIPS). – 2017.

[26] Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proc. of the 2019 Conference of the NAACL. – 2019. – P. 4171–4186.