Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Кафедра машинного обучения и цифровой гуманитаристики

**Направление подготовки / специальность:** 09.04.01 Информатика и вычислительная техника
**Направленность (профиль) подготовки:** Прикладная математика и информатика

# ПРОБЛЕМА ТЕМАТИЧЕСКОЙ НЕСБАЛАНСИРОВАННОСТИ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

(магистерская диссертация)

**Студент:**
Павлова Ирина Денисовна

_____
*(подпись студента)*

**Научный руководитель:**
Воронцов Константин Вячеславович,
д-р физ.-мат. наук

_____
*(подпись научного руководителя)*

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

Москва 2026

## Содержание

## 1. Аннотация

Вероятностное тематическое моделирование (probabilistic topic modeling) является одним из активно развивающихся направлений обработки естественного языка (*natural language processing, NLP*), начиная с конца 1990-х годов. Тематическая модель текстовой коллекции позволяет выявлять скрытую тематическую структуру данных, определяя, к каким темам относятся документы, а также какие слова характерны для каждой темы. На практике тематическое моделирование широко применяется для анализа больших массивов текстов, таких как новостные архивы, научные публикации, социальные сети и другое.

Одной из проблем вероятностного тематического моделирования является несбалансированность тематической структуры текстовых коллекций: реальные данные могут содержать темы, существенно различающиеся по своей представленности в объёме. Несбалансированность тематической структуры коллекции, согласно выдвинутой в работе гипотезе, приводит к двум характерным эффектам: расщеплению крупных тем и слиянию малых тем.

В рамках данной работы разработана методика количественного измерения эффектов тематической несбалансированности на основе синтетического расширения категоризированных коллекций и создана инструментальная среда для сравнения тематических моделей по устойчивости тем в условиях контролируемой несбалансированности. Проведённое экспериментальное исследование подтверждает гипотезу: степень расщепления крупных тем и слияния малых тем закономерно возрастает с увеличением дисбаланса коллекции.

## 2. Введение

### 2.1. Тематическое моделирование как метод анализа текстовых коллекций

Тематическое моделирование является одним из методов анализа текстовых коллекций, направленных на выявление скрытой тематической структуры документов [1]. В задачах анализа текстовых данных часто требуется представить каждый документ в виде набора тем, отражающих его содержа-

ние. Тематическое моделирование решает данную задачу в рамках обучения без учителя (*unsupervised learning*), автоматически выделяя набор тем на основе статистических закономерностей в данных.

Отдельный класс тематических моделей составляют вероятностные тематические модели, который определяются следующим образом: пусть $D$ — коллекция текстовых документов, $W$ — словарь терминов, $T$ — множество тем, а $n_d$ — количество терминов в документе $d$. Используя данное представление, документ $d$ может быть представлен в виде терминов $(w_1, \ldots, w_{n_d})$, $w_i \in W$. Вероятностная тематическая модель описывает условные вероятности появления терминов $w$ в документах $d$ через вероятности терминов в темах $\varphi_{wt} = p(w \mid t)$ и вероятности тем в документах $\theta_{td} = p(t \mid d)$. Несмотря на использование вероятностных представлений, такие модели не предназначены для генерации связных текстов и не могут соперничать с современными нейросетевыми языковыми моделями, такими как GPT [2], созданными специально под такой тип задач. Основное назначение тематических моделей заключается в интерпретируемом анализе тематической структуры документов, для чего они оказываются более простыми и менее ресурсоёмкими по сравнению с крупными нейросетевыми моделями.

## 2.2. Классические вероятностные методы тематического моделирования

Рассмотрим подробнее подходы вероятностного тематического моделирования. Как уже было упомянуто выше, одними из первых вероятностных подходов к тематическому моделированию стали работы Хофмана [3] и Блей и соавторов [4], основанные на принципе максимизации правдоподобия для оценки параметров модели и выявления скрытой тематической структуры текстовой коллекции. В своей работе Хофмана [3] предложил метод Probabilistic Latent Semantic Indexing (PLSI), в котором каждому документу $d \in D$ сопоставляется распределение по скрытым темам $\theta_{td} = p(t \mid d)$, а каждой теме $t \in T$ — распределение слов $\varphi_{td} = p(w \mid t)$. Тогда вероятность термина $w$ в документе $d$ выражается как

$$p(w \mid d) = \sum_{t \in T} p(w \mid t)\, p(t \mid d) = \sum_{t \in T} \varphi_{wt} \theta_{td}. \tag{1}$$

PLSI показал высокую эффективность для выявления тематической структуры небольших коллекций, однако число параметров растёт с числом документов, что ограничивает её обобщающую способность на новые данные. Метод Latent Dirichlet Allocation (LDA) [4], предложенный D. M. Blei и соавторами, решает эту проблему. Авторы вводят априорные распределения Дирихле для распределений тем в документах и слов в темах:

$$\theta_d \sim \mathrm{Dir}(\alpha), \quad \varphi_t \sim \mathrm{Dir}(\beta),$$

где $\alpha$ задаёт априорное распределение тем внутри документов, а $\beta$ определяет априорное распределение слов внутри тем. Наличие априорных распределений Дирихле наделяет LDA способностью обрабатывать новые документы без необходимости заново оценивать все параметры, а также уменьшает риск подстройки под шум (переобучения). Это позволяет получать более надёжные оценки распределений слов по темам и тем по документам по сравнению с PLSI.

### 2.3. Применение вероятностного тематического моделирования на практике

Вероятностные методы тематического моделирования, включая Latent Dirichlet Allocation (LDA) [4] и Probabilistic Latent Semantic Analysis (pLSA) [3], описанные выше, появились еще в конце 1990-х и начале 2000-х годов. Несмотря на это, данные модели до сих пор широко используются в прикладных задачах. На практике продолжают применяться как исходные алгоритмы, так и их модификации. Современные исследования (2020-2025 гг.) показывают, что вероятностные тематические модели продолжают эффективно использоваться для анализа сложных текстовых корпусов. Так, в исследовании [5], посвящённому анализу авиационных инцидентов, показано, что LDA и pLSA позволяют извлекать содержательно интерпретируемые темы, связанные с человеческими ошибками, техническими неисправностями и внешними факторами. Аналогичный подход используется в работе [6], где тематическое моделирование применяется к описаниям авиационных инцидентов для выявления повторяющихся семантических структур и группировки событий. Также множество современных публикаций посвящается методологическим аспектам выбора вероятностной тематической модели и влиянию этого вы-

бора на интерпретацию результатов. Так, например, В статье [7] проводится сравнительная оценка LDA, pLSA и других методов на датасете, собранном из отчётов об авиационных инцидентах. Показано, что классические вероятностные модели сохраняют конкурентоспособность по показателям когерентности и интерпретируемости тем. Сходные выводы получены и в более раннем исследовании [8], где LDA демонстрирует положительные результаты при анализе аварийных отчётов. Применимость вероятностного тематического моделирования подтверждается и за пределами авиационной области. Так, в работе [9] LDA используется для систематизации научной литературы в области Space Syntax (исследования в архитектуре и градостроительстве). В работе была показана универсальность вероятностного подхода для анализа специализированных научных датасетов. В обзорной статье Chauhan и Shah [10] систематизируются основные подходы к тематическому моделированию на основе LDA, рассматриваются расширения модели и текущие ограничения, такие как оценка качества тем, масштабируемость и интерпретируемость. Этот обзор подчёркивает, что, несмотря на широкое применение LDA, ряд фундаментальных проблем классических вероятностных моделей остаётся актуальным.

Дополнительно, в ряде недавних исследований вероятностные тематические модели применяются для анализа крупных и разнородных корпусов в условиях ограниченных вычислительных ресурсов. Так, в работе [11] LDA используется для анализа динамики общественных дискуссий в период пандемии COVID-19, где авторы подчёркивают преимущество классических вероятностных моделей с точки зрения интерпретируемости тем и устойчивости результатов по сравнению с более сложными нейросетевыми подходами. Аналогично, в исследовании [12] показано, что несмотря на активное развитие нейросетевых методов, вероятностные тематические модели остаются востребованными в прикладных задачах анализа текстов благодаря прозрачной вероятностной интерпретации и возможности явного контроля структуры тем.

Сейчас развитие классических моделей продолжается за счёт их расширения с сохранением вероятностной основы. Например, в статье [13] предложена модификация LDA с использованием нейронных компонентов. Эта работа подчёркивает актуальность классического вероятностного тематиче-

ского моделирования как основы для современных гибридных подходов. Таким образом, вероятностные тематические модели, их развитие и применение являются актуальной темой для исследований.

## 2.4. Проблема некорректности задачи оптимизации и регуляризация

Задача тематического моделирования формально сводится к решению некорректно поставленной задачи неотрицательного матричного разложения распределений слов и тем, поскольку обучение тематической модели формулируется как задача максимизации логарифма правдоподобия наблюдаемого корпуса документов. В вероятностных тематических моделях, таких как PLSI или LDA, этот факт обуславливает существование множества локальных оптимумов и значительную зависимость получаемых тем от начальных условий и гиперпараметров. Согласно теории регуляризации Тихонова [14], некорректную задачу оптимизации можно доопределить и сделать устойчивой. В частности, в работе [15] описана данная проблема и показано, что введение регуляризаторов позволяет комбинировать различные требования и получать более интерпретируемые и устойчивые модели по сравнению с классическими вероятностными подходами. Таким образом, задача построения тематической модели сводится к задаче регуляризованной максимизации логарифмического правдоподобия наблюдаемых данных, что демонстрирует следующее выражение:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Phi,\Theta) \quad \to \quad \max_{\Phi,\Theta},$$

где $R$ является регуляризатором, а на $\Phi$ и $\Theta$ наложены следующие ограничения:

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

Дальнейшее развитие данного подхода представлено в работе Ирхина, Булатова и Воронцова [16], где задача тематического моделирования также рассматривается как некорректно поставленная задача неотрицательного матричного разложения. Авторы показывают, что стандартная максимизация правдоподобия приводит к неустойчивым решениям и высокой вариативности

получаемых тем, что обуславливается наличием множества локальных экстремумов. В рамках аддитивной регуляризации тематических моделей (ARTM) предлагается добавлять к логарифму правдоподобия взвешенную сумму регуляризаторов, которые обеспечивают дополнительные требования к модели. Экспериментальные результаты демонстрируют, что такая постановка позволяет стабилизировать процесс обучения и получать более интерпретируемые и различимые темы по сравнению с классическими вероятностными моделями (такими как LDA и PLSI, например).

Более систематическое изложение теории аддитивной регуляризации тематических моделей и её практической реализации представлено в работе Воронцова [17]. В книге подробно рассматриваются причины нестабильности классических вероятностных моделей, принципы построения регуляризаторов в рамках ARTM, а также описывается библиотека BigARTM с открытым кодом, ориентированная на масштабируемое и воспроизводимое тематическое моделирование.

Кроме формальных проблем оптимизации, практические ограничения классических вероятностных моделей проявляются и в прикладных задачах. Систематический обзор применения тематических моделей к коротким текстам социальных сетей показал, что LDA и его модификации часто используются без учета особенностей данных [18]. Авторы отмечают, что во многих работах применяются метрики, которые плохо отражают интерпретируемость тем с семантической точки зрения.

В работе [18] также отмечается, что отсутствие единых рекомендаций по построению, оценке и интерпретации тематических моделей снижает качество выделяемых тем, что особенно заметно это проявляется при работе с короткими, разреженными или несбалансированными текстовыми корпусами.

### 2.5. Проблема тематической несбалансированности

Одной из практических проблем вероятностного тематического моделирования является несбалансированность тематической структуры текстовых коллекций. В реальных корпусах разные темы обычно представлены неравномерно. Одни темы содержат большое количество документов, тогда как другие встречаются значительно реже. Подобная неоднородность распределения документов влияет на процесс обучения тематических моделей и

может приводить к искажению структуры тематического пространства.

Проблема тематической несбалансированности особенно важна в прикладных задачах анализа научных публикаций, новостных коллекций, социальных сетей и специализированных предметных корпусов. В таких данных крупные темы начинают доминировать при обучении модели, тогда как менее представленные темы могут частично исчезать или смешиваться с другими направлениями.

В результате модель начинает хуже разделять близкие по содержанию темы. Это снижает интерпретируемость тематической структуры и усложняет анализ редких тематических областей.

### 2.5.1. Определение и эффекты расщепления/слияния

Под дисбалансом тематической структуры будем понимать ситуацию, когда различные темы представлены в текстовой коллекции в существенно разных объёмах. Иными словами, количество документов, относящихся к разным темам, может отличаться в разы или на порядки.

Такая ситуация является типичной для реальных текстовых коллекций. Например, корпус может содержать сотни документов, посвящённых одной предметной области, и лишь небольшое число текстов, относящихся к другим темам. При этом семантически эти редкие темы могут быть чётко выраженными и содержательно значимыми, несмотря на малую представленность в данных.

Так, если рассматривать коллекцию, состоящую из 1000 документов, из которых 980 посвящены биологии, а по 10 математике и социологии, то вероятностная тематическая модель будет вести себя следующим образом. Если обучить тематическую модель с тремя темами, то в большинстве случаев все три темы окажутся связанными с биологией, а документы по математике и социологии будут распределены между ними случайным образом.

Чтобы модель выделила отдельные темы для математики и социологии, потребуется значительно увеличить число тем. Однако в этом случае большая часть тем снова будет посвящена биологии и отличаться друг от друга лишь незначительно.

Таким образом, стандартные вероятностные тематические модели оказываются неспособными корректно восстанавливать редкие, но семантически

однородные темы в условиях сильного тематического дисбаланса.

Вероятностные тематические модели, такие как pLSA и другие, основанные на стохастическом неотрицательном матричном разложении, при обучении путём максимизации логарифма правдоподобия демонстрируют тенденцию к относительно равномерному использованию всех тем модели. Данное свойство обусловлено самой постановкой задачи оптимизации: для увеличения правдоподобия модели выгодно задействовать все используемые параметры [3, 4, 19]. Подобное поведение и его влияние на структуру тем отмечались и в последующих эмпирических исследованиях вероятностных тематических моделей [19]. В результате модели оказываются склонны распределять документы между всеми темами таким образом, чтобы каждая тема была задействована при описании данных, даже если в реальной коллекции соответствующая тема представлена слабо. Иначе говоря, модель стремится «использовать» все заданные темы, а не оставлять часть из них почти пустыми. Если какая-либо тема получает слишком малую долю документов, её вклад в правдоподобие становится незначительным, и модель теряет возможность эффективно использовать соответствующие параметры. Описанные эффекты приводят к двум характерным проблемам тематического моделирования: расщеплению крупных тем на несколько слабо различимых тем-дубликатов и слиянию маломощных тем в семантически неоднородные «мусорные» темы. Описанные проблемы были экспериментально продемонстрированны на больших и несбалансированных корпусах. Например, в работе [20] авторы обучали LDA на финансовом корпусе, где одни тематические области были представлены сотнями документов, а другие - всего несколькими десятками. Результаты показали, что стандартная LDA склонна распределять документы редких темы между несколькими наиболее часто встречающимися темами, что приводит к дроблению крупных тем и слиянию мелких.

В работе Wallach et al. [21] показано, что стандартная постановка LDA с симметричными априорными распределениями Дирихле существенно влияет на итоговую тематическую структуру. Авторы демонстрируют, что выбор априоров напрямую определяет баланс тем и может приводить к искусственному выравниванию их вероятностей, маскируя редкие и низкочастотные темы. Для устранения данного эффекта предлагается использовать асимметричные априоры и оценивать их гиперпараметры в процессе обучения с помо-

щью байесовского вывода. Экспериментальные результаты показывают, что такая модификация повышает интерпретируемость тем и позволяет частично учитывать дисбаланс тематической структуры корпуса, однако не устраняет полностью проблему нестабильности решений и зависимости от начальных условий.

Помимо описанных ограничений классических подходов, вероятностные тематические модели часто показывают нестабильные результаты (что будет продемонстрировано в рамках данной работы). Повторное обучение одной и той же модели на одинаковом корпусе при разных начальных инициализациях или настройках может приводить к заметно различающимся тематическим распределениям [22].

Подобная вариативность связана с особенностями численных методов оптимизации и наличием большого числа локальных оптимумов. Результатов является различное распределение документы между темами при разных запусках, особенно при наличии редких тем или сильного тематического дисбаланса.

В работе [23] показано, что стандартная модель LDA способна формировать существенно разные тематические структуры даже при повторном обучении на одной и той же коллекции документов. Наиболее заметно данный эффект проявляется для редких и слабо выраженных тем. Это дополнительно усложняет корректное выделение небольших тематических областей в несбалансированных корпусах.

### 2.5.2. Существующие подходы к компенсации дисбаланса в научной литературе

В задачах машинного обучения проблема несбалансированности классов исследована достаточно подробно. Один из наиболее известных обзоров данной области представлен в работе [24]. Классические подходы включают ресемплинг данных, взвешивание объектов и модификацию функции потерь с учётом стоимости ошибок для различных классов. Однако прямое применение данных методов в задачах тематического моделирования оказывается невозможным. В отличие от классификации, тематические модели работают в условиях отсутствия априорно заданных меток классов: тематическая принадлежность документов и слов является скрытой и восстанавливается

в процессе обучения модели. В результате невозможно заранее определить, какие документы или слова относятся к «редким» темам, а какие — к «частым», что исключает использование стандартных процедур ресемплинга и взвешивания объектов. Более того, попытки искусственного вмешательства в распределение данных без учёта латентной структуры тем могут приводить к искажению семантического содержания тем и ухудшению интерпретируемости модели.

В последние годы предложено несколько подходов, которые могут быть применены для решения задачи ухудшения интерпретируемости тем при дисбалансе:

Одним из таких методов является **Регуляризация «ёмкости» тем (TopicPriorR**
что демонстрируется в работе Веселовой и соавторов [25]. Авторы показали существование проблемы дисбаланса тем в вероятностных тематических моделях, хотя количественного измерения влияния дисбаланса на качество моделей произведено не было. Экспериментально показано, что классические модели, такие как pLSA и LDA, при обучении на несбалансированных коллекциях стремятся равномерно распределять документы между всеми темами модели. Для формального анализа дисбаланса авторы вводят понятие *ёмкости темы* (topic capacity) как суммы вероятностей темы по всем документам:

$$n_t = \sum_{d \in D} p(t \mid d)\, n_d,$$

где $n_d$ — длина документа $d$. Степень дисбаланса определяется через отношение

$$k = \frac{n_{\max}}{n_{\min}},$$

где $n_{\max}$ и $n_{\min}$ — максимальная и минимальная ёмкость среди тем. В стандартных LDA/pLSA моделях наблюдается тенденция к выравниванию $n_t$, что не отражает реального распределения тем в коллекции.

Для решения этой проблемы авторы предлагают *TopicPriorRegularizer*, который задаёт априорные предпочтения распределения тем:

$$R_{\text{TopicPrior}}(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_t \log \phi_{wt},$$

где $\beta_t$ — коэффициент, отражающий желаемую «ёмкость» темы $t$. Экспериментальные результаты показывают, что использование TopicPriorRegularizer позволяет корректно выделять редкие темы, предотвращая их «растворение», сохранять крупные темы без дробления и получать более интерпретируемые темы по документам.

Еще одним методом, известным из научной литературы, является **семантическая инициация через ключевые слова (Keyword-Assisted Topic Models)**. В работе [26] предложена модель *Keyword Assisted Topic Models* (keyATM), в которой к классической LDA добавляется набор ключевых слов, заранее сформированный для известных тем.

Авторы отмечают, что полностью автоматические методы тематического моделирования нередко формируют темы с пересекающимся содержанием или объединяют разные концепты в одну тему. Для уменьшения подобных эффектов в модели keyATM исследователь задаёт ключевые слова до начала обучения. Тем самым часть информации о тематической структуре вводится вручную.

Эксперименты показывают, что использование ключевых слов улучшает качество тематического моделирования. Модель keyATM формирует более интерпретируемые темы, уменьшает пересечения между ними и снижает вероятность объединения разных тем. Кроме того, модель оказывается менее чувствительной к выбору числа тем и настройке гиперпараметров по сравнению со стандартной LDA.

Метод **выявления редких тем через графовые связи слов** также продемонстрировал способность решить данную задачу.Одним из последствий тематической несбалансированности является слабое выделение редких тем или их полное исчезновение из итоговой тематической структуры. Для решения данной проблемы был предложен ряд графовых тематических моделей. В таких подходах текст представляется в виде сети совстречаемости слов. Это позволяет учитывать связи между словами и усиливать влияние редких тем при обучении модели.

В работе *Word Network Topic Model* (WNTM) [27] предложено использо-

вать графовое представление текста для повышения чувствительности модели к редким темам. В отличие от классической LDA, модель WNTM строит сеть совстречаемости слов и оценивает распределение тем не для документов, а для отдельных слов.

Такой подход делает семантическое пространство более плотным и позволяет лучше выделять мало представленные темы даже в несбалансированных корпусах. Авторы показывают, что WNTM эффективно обнаруживает редкие темы и может применяться в различных сценариях без существенного роста вычислительных затрат.

При этом модель имеет ряд ограничений. WNTM ориентирована преимущественно на небольшие корпуса и короткие тексты. Кроме того, модель не полностью устраняет проблему пересечения тем, когда редкие темы частично смешиваются с более популярными направлениями.

В работе [28] предложена модель *CWIBTD* (Co-occurrence Word Network Based Rare Topic Discovery), ориентированная на обнаружение редких тем в коротких и несбалансированных текстовых наборах. В отличие от классического LDA, модель строит сеть совстречаемости слов и использует активность узлов для моделирования распределения тем, что повышает плотность семантического пространства и чувствительность к мало представленным темам. Экспериментальные результаты показывают, что CWIBTD превосходит стандартные методы в задаче выявления редких тем, сохраняя простоту вывода благодаря использованию сэмплирования Гиббса, аналогичного LDA [28].

Похожую идею развивает модель *CWUTM*, предложенная в [29]. В ней переработано определение активности узлов и введено дополнительное нормирование представлений как редких, так и частых тем, что позволяет ещё более эффективно выделять низкочастотные тематические структуры. Валидация на несбалансированных корпусах коротких текстов подтверждает превосходство CWUTM над базовыми подходами LDA в задаче обнаружения редких тем [29].

CWIBTD и CWUTM хорошо работают на коротких и относительно небольших корпусах, но имеют ряд ограничений: модели ориентированы на короткие тексты; из-за построения графов слов и вычисления активности узлов вычислительная сложность растёт с числом уникальных слов, что ограничивает масштабируемость; несмотря на улучшение выделения редких тем,

модели не решают полностью проблему семантического перекрытия тем, особенно когда редкие темы сильно пересекаются с популярными.

**Выделение новых тем через foreground-background подход.** Ещё одним способом борьбы с эффектами дисбаланса тематической структуры является выявление так называемых новых или возникающих тем через сравнение двух подмножеств документов: фонового (background) и фокусного (foreground). В рамках данного подхода корпус разделяется на два набора, где фоновая часть отражает уже устоявшуюся тематическую структуру, а фокусная содержит документы, в которых предполагается появление новых или слабо представленных тем. Темы, характерные преимущественно для фокусного поднабора и отсутствующие либо слабо выраженные в фоновой части, интерпретируются как новые или редкие.

Идея противопоставления фонового и фокусного распределений тем восходит к работе [30], в которой была предложена модель для выявления эволюции тематической структуры текстовых коллекций во времени. Авторы рассматривают фоновые темы как отражение общего, устойчивого контекста корпуса, тогда как фокусные темы соответствуют новым или усиливающимся направлениям. Сравнение тематических распределений позволяет выявлять темы, значимость которых возрастает в фокусной части по сравнению с фоном, что делает данный подход особенно полезным для анализа динамических и несбалансированных корпусов.

Подобные foreground-background подходы могут быть реализованы на основе вероятностных тематических моделей, в частности LDA, путём отдельного обучения моделей на фоновом и фокусном поднаборах и последующего сопоставления полученных тем. Если тема, выявленная в фокусной части, существенно отличается от всех фоновых тем по семантическому составу, она рассматривается как новая тематическая структура. Такой подход широко применяется при анализе корпусов с временной разметкой, где фокусная выборка соответствует более позднему периоду, а фоновая — более раннему.

Несмотря на способность выявлять новые и редкие темы, foreground-background методы имеют ряд ограничений. Их эффективность существенно зависит от способа разбиения корпуса на фоновую и фокусную части, а также от выбо-

ра числа тем. Кроме того, при частичном пересечении лексического состава фоновых и фокусных тем возможны ложные срабатывания, при которых фоновые слова включаются в состав «новых» тем. Тем не менее, данные подходы представляют собой полезный инструмент анализа несбалансированных коллекций, в которых новые тематические структуры формируются постепенно и могут быть утрачены при стандартном обучении LDA на всём корпусе.

### 2.5.3. Современные направления развития тематического моделирования

Помимо классических вероятностных моделей [31], в последние годы активно развиваются нейросетевые тематические модели, использующие распределённые представления слов и документов, а также непараметрические и полууправляемые подходы, такие как HDP [32], STM [33, 34] и GuidedLDA [35]. Эти методы могут автоматически определять число тем, учитывать метаданные документов или использовать априорную информацию в виде ключевых слов. Однако их подробное рассмотрение выходит за рамки настоящей работы, которая сосредоточена на классических вероятностных моделях (pLSA, LDA) и их регуляризованных версиях (ARTM). Выбор обусловлен тем, что целью исследования является анализ влияния тематического дисбаланса на структуру тем в наиболее распространённых и широко используемых вероятностных моделях. Кроме того, HDP ориентирован прежде всего на автоматическое определение числа тем, тогда как STM и GuidedLDA используют дополнительную информацию о документах или темах и решают несколько иную задачу по сравнению с полностью неуправляемым тематическим моделированием. Несмотря на различия в механизмах построения тем, предложенная в работе методика оценки эффектов дисбаланса потенциально может быть распространена и на данные классы моделей, что представляет интерес для дальнейших исследований.

## 2.6. Подходы к оценке качества тематических моделей

### 2.6.1. Традиционные метрики качества тематических моделей

Оценка качества тематических моделей является важной задачей, поскольку тематическое моделирование представляет собой задачу без учителя

и не предполагает наличия единственно правильного решения. В связи с этим в литературе было предложено множество подходов к оценке качества полученных тем. Наиболее распространёнными являются вероятностные метрики качества, основанные на способности модели описывать данные, а также метрики семантической согласованности тем.

Одной из наиболее известных вероятностных метрик является *перплексия* (perplexity), широко используемая при сравнении тематических моделей и подборе их параметров [4, 19]. Перплексия характеризует способность модели предсказывать слова в ранее не наблюдавшихся документах. Чем ниже значение перплексии, тем лучше модель описывает вероятностную структуру корпуса. Однако многочисленные исследования показывают, что снижение перплексии не всегда приводит к повышению интерпретируемости тем для человека [36, 37].

В связи с этим широкое распространение получили метрики *когерентности тем* (topic coherence), оценивающие степень семантической связанности слов внутри темы. Наиболее известными являются меры UMass, UCI, PMI и NPMI [37, 38, 39]. Предполагается, что тема является качественной, если её наиболее вероятные слова часто встречаются совместно и образуют семантически согласованный набор. Исследования показывают, что показатели когерентности значительно лучше коррелируют с человеческими оценками качества тем по сравнению с вероятностными метриками [37, 39].

Несмотря на широкое использование перплексии и когерентности, данные показатели оценивают преимущественно внутренние свойства тематической модели. Перплексия характеризует качество вероятностного описания документов, а когерентность — согласованность слов внутри отдельных тем. При этом данные метрики не позволяют определить, насколько полученные темы соответствуют реальной структуре предметной области, а также не выявляют возможные структурные искажения тематической модели. В частности, высокая когерентность темы не гарантирует отсутствия эффектов слияния нескольких тем в одну тему или расщепления одной темы на несколько новых.

Дополнительным направлением развития методов оценки тематических моделей являются подходы, основанные на сопоставлении тем с внешними источниками знаний и предметной областью. В частности, в работе Chuang

и соавт. [40] предлагается метод *topical alignment*, позволяющий оценивать соответствие тем модели заранее заданной структуре предметной области. Однако даже такие подходы не дают полного представления о структурных искажениях тематического пространства, возникающих при изменении распределения документов и наличии тематической несбалансированности.

Следовательно, традиционные и внешние метрики качества оказываются недостаточными для анализа структурных эффектов, возникающих в тематических моделях при дисбалансе тем.

### 2.6.2. Методы сопоставления и сравнения тематических моделей

Задача сопоставления тематических моделей возникает в различных сценариях, включая сравнение моделей с разными параметрами, анализ устойчивости тематической структуры, а также исследование динамики тем во времени или при изменении характеристик корпуса. В отличие от задач внутренней оценки качества, здесь рассматривается проблема установления соответствия между темами, полученными в разных моделях, и анализа их взаимного сходства.

Одним из наиболее простых и широко используемых подходов является измерение попарного сходства тематических распределений. Темы в вероятностных моделях, таких как LDA, представляются в виде распределений вероятностей слов, что позволяет применять стандартные меры близости векторных представлений. Наиболее распространённой является косинусная мера сходства, используемая для оценки близости тематических векторов в пространстве слов. Данный подход широко применяется в исследованиях тематического моделирования благодаря своей вычислительной эффективности и интерпретируемости.

Альтернативным направлением является использование дивергенций между распределениями, таких как KL-дивергенция и её симметризованные варианты, включая Jensen–Shannon дивергенцию. Подобные меры позволяют учитывать вероятностную природу тематических распределений и применяются для более точного измерения различий между темами [41, 42]. По сравнению с косинусным сходством дивергенции сильнее реагируют на различия в маловероятных частях распределений. Это делает их полезными при анализе тонких различий между темами.

После построения матрицы попарных сходств или расстояний возникает задача сопоставления тем двух моделей. Для этого часто используется задача оптимального назначения (assignment problem), которая может решаться с помощью венгерского алгоритма (Hungarian algorithm). Такой подход позволяет находить взаимное соответствие между темами двух моделей с максимальным суммарным сходством [43].

Отдельное направление исследований связано с методами topic alignment и topic matching. В подобных подходах сопоставление тем рассматривается как задача поиска соответствий между тематическими распределениями разных моделей или между темами модели и внешними категориями корпуса. В частности, рассматриваются подходы, основанные на максимизации сходства между темами или на вероятностных моделях соответствия, позволяющих учитывать неоднозначность сопоставления и возможное расщепление тем.

Дополнительным простым подходом является сравнение тем на основе пересечения наиболее вероятных слов (top-N words overlap). В этом случае тема представляется как множество наиболее вероятных терминов, а степень сходства определяется количеством общих слов между темами. Несмотря на свою простоту, данный метод широко используется в прикладных исследованиях благодаря высокой интерпретируемости и отсутствию необходимости учитывать полные распределения слов.

Несмотря на разнообразие существующих методов сопоставления тематических моделей, большинство из них направлено на установление парного соответствия между темами различных моделей. При этом они не позволяют явно выделять структурные эффекты, возникающие при изменении тематического распределения корпуса, такие как расщепление одной темы на несколько или слияние нескольких тем в одну. В данной работе предлагается подход к анализу таких эффектов на основе матрицы попарного сходства тематических распределений и порогового выделения отношений correspondence, split и merge, что позволяет количественно оценивать влияние тематической несбалансированности на структуру тематического пространства.

## 2.7. Выводы по главе

В данной главе рассмотрены основные подходы к тематическому моделированию текстовых коллекций, начиная с классических вероятностных моделей (pLSA, LDA) и заканчивая их современными расширениями и альтернативными направлениями. Показано, что несмотря на развитие нейросетевых и непараметрических методов, классические вероятностные модели остаются широко используемым инструментом анализа текстов благодаря своей интерпретируемости, устойчивости и сравнительной вычислительной эффективности.

Отдельное внимание уделено проблемам оптимизации и регуляризации тематических моделей, которые формально сводятся к некорректно поставленным задачам неотрицательного матричного разложения. Показано, что наличие множества локальных оптимумов и зависимость от гиперпараметров приводит к нестабильности результатов и необходимости введения регуляризаторов, как в рамках аддитивной регуляризации тематических моделей (ARTM).

Далее рассмотрена проблема тематической несбалансированности, проявляющаяся в неравномерном распределении тем в корпусе и приводящая к характерным искажениям тематической структуры, таким как расщепление крупных тем и слияние редких тем. Показано, что стандартные вероятностные модели склонны к выравниванию тематических распределений, что затрудняет корректное выделение слабо представленных тем.

Также были рассмотрены существующие подходы к уменьшению эффектов тематического дисбаланса. К ним относятся регуляризация тематической ёмкости, семантическая инициализация с использованием ключевых слов, графовые модели для выделения редких тем и foreground background подходы. Несмотря на положительные результаты отдельных методов, ни один из них не решает проблему тематической несбалансированности в общем случае. Большинство существующих подходов ориентировано на конкретные сценарии или требует дополнительных предположений о структуре данных.

Дополнительно были проанализированы методы оценки качества и сопоставления тематических моделей. Показано, что традиционные метрики, включая перплексию и когерентность, оценивают преимущественно внутрен-

ние свойства модели и не позволяют напрямую выявлять структурные изменения тематического пространства. Методы сопоставления тем, основанные на мерах сходства распределений, дивергенциях, assignment подходах и topic alignment, позволяют устанавливать соответствия между темами различных моделей, однако не описывают явно эффекты расщепления и слияния тем.

Таким образом, существующие методы в основном ориентированы либо на локальную оценку качества отдельных тем, либо на поиск соответствий между ними. При этом задача количественного анализа структурных изменений тематического пространства в условиях дисбаланса данных остаётся недостаточно исследованной.

В связи с этим актуальной является задача разработки методов, позволяющих одновременно учитывать соответствие тем между моделями и фиксировать структурные эффекты их преобразования, такие как split и merge. Именно этой задаче посвящена последующая часть работы, в которой предлагается подход к количественной оценке влияния тематической несбалансированности на структуру тематических моделей.

# 3. Методика исследования влияния тематической несбалансированности

## 3.1. Постановка исследовательской задачи

Проведённый обзор литературы показал, что проблема тематической несбалансированности является одним из факторов, влияющих на качество вероятностных тематических моделей. При наличии существенных различий в объёмах тем, представленных в корпусе, качество тематических моделей может ухудшаться, а получаемая тематическая структура – искажаться. В литературе отмечаются сложности выделения редких тем, снижение устойчивости результатов и ухудшение интерпретируемости моделей. Однако вопрос о том, каким образом данные эффекты проявляются на уровне структуры корпуса и приводят ли они к явлениям расщепления, слияния или исчезновения тем, остаётся недостаточно изученным. Несмотря на существование подходов, позволяющих частично смягчить отдельные проявления тематической несбалансированности, в литературе практически отсутствуют методы количественного анализа того, как именно дисбаланс влияет на структуру

тематической модели. В частности, отсутствуют общепринятые способы измерения таких эффектов, как расщепление, слияние и исчезновение тем, а также методы сравнения тематических моделей по степени устойчивости к данным явлениям.

Дополнительную сложность представляет отсутствие эталонной тематической структуры для реальных текстовых коллекций. В результате затрудняется анализ того, каким образом изменение распределения документов между темами влияет на итоговую структуру тематической модели и насколько устойчивыми оказываются различные методы тематического моделирования.

Целью данной работы является разработка методики экспериментального исследования влияния тематической несбалансированности на структуру тематических моделей и проведение сравнительного анализа устойчивости различных подходов к тематическому моделированию в условиях контролируемого дисбаланса тем.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Выполнить анализ существующих работ, посвящённых проблеме тематической несбалансированности, устойчивости тематических моделей и методам оценки качества тематического моделирования.

2. Разработать методику количественной оценки эффектов тематической несбалансированности на основе сравнения тематических структур моделей, обученных на коллекциях с различной степенью дисбаланса.

3. Разработать программную инструментальную среду для автоматизации экспериментов по обучению тематических моделей, генерации несбалансированных корпусов и анализу изменений тематической структуры.

4. Сформировать набор экспериментальных коллекций на основе датасета 20 Newsgroups с контролируемой степенью тематической несбалансированности.

5. Провести экспериментальное исследование влияния дисбаланса на вероятностные тематические модели и количественно оценить возникающие эффекты расщепления, слияния и исчезновения тем.

6. Выполнить сравнительный анализ базовой модели pLSA, регуляризованных тематических моделей и моделей, учитывающих локальный контекст слов, с точки зрения их устойчивости к тематической несбалансированности.

Решение поставленных задач позволит разработать методику количественного анализа влияния тематической несбалансированности на структуру тематических моделей, провести её экспериментальную апробацию и выполнить сравнительное исследование устойчивости различных подходов к тематическому моделированию в условиях контролируемого дисбаланса тем.

## 3.2. Формирование экспериментальных корпусов

Для исследования влияния тематической несбалансированности на структуру тематических моделей необходим корпус, для которого заранее известна тематическая принадлежность документов. Это позволяет контролируемым образом изменять распределение документов между темами и анализировать возникающие изменения тематической структуры.

Использование реальных текстовых коллекций для такой задачи затруднено. В большинстве практических сценариев истинная тематическая структура корпуса неизвестна, а сами документы могут одновременно относиться к нескольким темам. В результате становится сложно определить, вызваны ли наблюдаемые изменения особенностями модели или объективной неоднородностью данных. По этой причине в данной работе используются синтетически сформированные корпуса с контролируемой степенью тематической несбалансированности.

В качестве исходного набора данных выбран датасет *20 Newsgroups* [44], являющийся одним из наиболее распространённых бенчмарков для задач тематического моделирования, кластеризации и классификации текстов. Корпус содержит около 20 тысяч сообщений из двадцати тематических групп новостей и обладает заранее известной категоризацией документов.

Для уменьшения семантического пересечения тем в экспериментах использовалось подмножество категорий датасета, относящихся к различным предметным областям и имеющим минимальное лексическое пересечение. В частности, были выбраны категории *comp.sys.mac.hardware, comp.windows.x,*

*rec.autos, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, soc.religion.christ talk.politics.guns, talk.politics.mideast*. Такой подход позволяет более надёжно интерпретировать изменения тематической структуры модели и снижает влияние неоднозначности категорий на результаты экспериментов.

На основе выбранных категорий формировалась базовая сбалансированная коллекция, в которой каждая тема была представлена одинаковым количеством документов. Полученная коллекция использовалась в качестве эталонного корпуса при дальнейшем сравнении тематических моделей.

В работе рассматривались два сценария тематической несбалансированности. В первом сценарии увеличивалась представленность одной категории, тогда как количество документов в остальных категориях оставалось неизменным. Такой сценарий моделирует ситуацию существования одной доминирующей темы в коллекции. Во втором сценарии одновременно увеличивалась представленность нескольких категорий. При этом все доминирующие категории увеличивались одинаково, что позволяло моделировать наличие нескольких крупных тематических кластеров в корпусе. Остальные категории сохраняли исходный объём документов. Следует отметить, что коэффициент дисбаланса определялся как отношение количества документов в доминирующих категориях к количеству документов в остальных категориях. Во всех экспериментах число документов в недоминирующих категориях оставалось постоянным, что позволяло исследовать влияние исключительно роста представленности отдельных категорий на структуру тематической модели.

Таблица 1 – Пример распределения документов при одной доминирующей категории

| Коэффициент дисбаланса $k$ | Доминирующая категория | Остальные категории |
|---|---|---|
| 1 | 100 | 100 |
| 2 | 200 | 100 |
| 4 | 400 | 100 |
| 8 | 800 | 100 |
| 16 | 1600 | 100 |

Таблица 2 – Пример распределения документов при нескольких доминирующих категориях

| Коэффициент дисбаланса $k$ | Доминирующие категории | Остальные категории |
|---|---|---|
| 1 | 100 : 100 : 100 | 100 |
| 2 | 200 : 200 : 200 | 100 |
| 4 | 400 : 400 : 400 | 100 |
| 8 | 800 : 800 : 800 | 100 |
| 16 | 1600 : 1600 : 1600 | 100 |

Для моделирования тематической несбалансированности формировалась серия производных корпусов. В каждом эксперименте одна или несколько из категорий (каждая из категорий характеризует заранее известную тему) выбирались в качестве преобладающей. Для увеличения представленности данных категорий использовалась генерация дополнительных документов с помощью большой языковой модели DeepSeek V4 Flash.

При генерации модели передавались несколько документов выбранной категории из датасета 20 Newsgroups, а также указание на соответствующую тему категории. На основе этих примеров модель формировала новые тексты, тематически близкие к исходным документам и сохраняющие характерную лексику предметной области. После формирования дополнительных документов они объединялись с исходной коллекцией, что позволяло получать корпуса с различной степенью тематической несбалансированности при сохранении общей тематической структуры исходного датасета.

Степень дисбаланса определялась коэффициентом

$$k = \frac{N_{\text{major}}}{N_{\text{minor}}},$$

где $N_{\text{major}}$ обозначает число документов в преобладающей категории, а $N_{\text{minor}}$ — число документов в остальных категориях.

В работе рассматривались несколько уровней дисбаланса, начиная от сбалансированного корпуса ($k = 1$) и заканчивая сильно несбалансированными коллекциями, в которых объём доминирующей категории (или нескольких доминирующих) превышал объём остальных категорий в несколько десятков раз. Для каждого значения коэффициента дисбаланса формировался отдельный корпус, на котором независимо обучались тематические модели.

Подобная схема построения экспериментальных данных позволяет контролируемо изменять только распределение документов между темами, сохраняя неизменными остальные характеристики корпуса. Благодаря этому наблюдаемые изменения тематической структуры могут быть непосредственно связаны с влиянием тематической несбалансированности.

## 3.3. Предобработка текстов

Качество тематического моделирования в значительной степени зависит от предварительной обработки текстов. Несмотря на широкое использование датасета 20 Newsgroups в задачах тематического моделирования, анализ исходных данных показал наличие ряда особенностей, способных негативно влиять на качество выделяемых тем.

Во-первых, часть документов содержит значительный объём технического шума, не связанного с тематическим содержанием сообщений. Наиболее характерным примером являются фрагменты в формате uuencode, использовавшиеся в группах Usenet для передачи бинарных файлов через текстовые сообщения. Такие последовательности состоят из длинных строк специальных символов и практически не несут семантической информации. Для оценки распространённости данной проблемы был проведён анализ обучающей выборки датасета 20 Newsgroups. Результаты представлены в таблице 3.

Как видно из таблицы, признаки uuencoded-контента присутствуют более чем в 18% документов обучающей выборки. Для отдельных категорий доля таких сообщений превышает 25%. Наличие подобных артефактов способно приводить к появлению искусственных тем, отражающих особенности кодирования данных вместо содержательной структуры корпуса.

Во-вторых, после удаления технического шума, стоп-слов и других нерелевантных элементов обнаружилось значительное количество документов с крайне малым объёмом содержательного текста. Для тематического моделирования такие документы малоинформативны, поскольку содержат недостаточно данных для устойчивого определения тематической принадлежности.

Статистика длины документов после предобработки представлена в таблице 4.

Поскольку документы длиной менее шести содержательных слов не

Таблица 3 – Количество документов, содержащих uuencoded-фрагменты

| Категория | Всего | С мусором | Доля, % |
|---|---|---|---|
| alt.atheism | 480 | 78 | 16.2 |
| comp.graphics | 584 | 112 | 19.2 |
| comp.os.ms-windows.misc | 591 | 153 | 25.9 |
| comp.sys.ibm.pc.hardware | 590 | 110 | 18.6 |
| comp.sys.mac.hardware | 578 | 107 | 18.5 |
| comp.windows.x | 593 | 145 | 24.5 |
| misc.forsale | 585 | 97 | 16.6 |
| rec.autos | 594 | 71 | 12.0 |
| rec.motorcycles | 598 | 151 | 25.3 |
| rec.sport.baseball | 597 | 87 | 14.6 |
| rec.sport.hockey | 600 | 128 | 21.3 |
| sci.crypt | 595 | 69 | 11.6 |
| sci.electronics | 591 | 106 | 17.9 |
| sci.med | 594 | 153 | 25.8 |
| sci.space | 593 | 109 | 18.4 |
| soc.religion.christian | 599 | 79 | 13.2 |
| talk.politics.guns | 546 | 140 | 25.6 |
| talk.politics.mideast | 564 | 67 | 11.9 |
| talk.politics.misc | 465 | 86 | 18.5 |
| talk.religion.misc | 377 | 38 | 10.1 |
| Всего | 11314 | 2086 | 18.4 |

позволяют надёжно оценивать тематические распределения, такие документы исключались из дальнейшего анализа.

Итоговая процедура предобработки включала следующие этапы:

1. удаление адресов электронной почты;
2. удаление uuencoded-фрагментов и других технических артефактов;
3. приведение текста к нижнему регистру;
4. токенизацию текста;
5. удаление чисел и знаков пунктуации;
6. лемматизацию слов с использованием библиотеки spaCy;
7. удаление общеязыковых стоп-слов из библиотеки Scikit-Learn;

Таблица 4 – Статистика коротких документов после предобработки

| Условие | Количество документов |
|---|---|
| Количество слов = 0 | 19 (0.19%) |
| Количество слов < 2 | 87 (0.86%) |
| Количество слов < 6 | 596 (5.87%) |

8. удаление дополнительного набора пользовательских стоп-слов, характерных для сообщений Usenet;

9. исключение коротких документов, содержащих менее шести содержательных слов.

Для формирования словаря использовались только термины, встречающиеся в коллекции не менее 20 раз. Такое ограничение позволяет уменьшить размер словаря и исключить случайные или редко встречающиеся слова, не оказывающие существенного влияния на тематическую структуру корпуса.

В результате была получена очищенная версия корпуса, пригодная для проведения дальнейших экспериментов по исследованию влияния тематической несбалансированности на качество тематических моделей.

### 3.4. Построение словаря и представление корпуса

Для обеспечения сопоставимости экспериментов во всех моделях использовался единый процесс построения словаря и предобработки текстов. Словарь формировался на основе объединённой обучающей выборки корпуса 20 Newsgroups после этапа предобработки, описанного ранее. В словарь включались только те термины, которые встречались в корпусе не менее 20 раз, что позволило исключить случайные и шумовые слова и уменьшить размер признакового пространства.

При формировании корпуса для обработки моделью, для всех текстов применялась одинаковая схема предобработки, описанная в разделе выше. Таким образом, различия между моделями не связаны с различиями в исходной текстовой обработке, а определяются исключительно форматом представления корпуса и особенностями самих моделей.

Для моделей, основанных на реализации BigARTM, в частности PLSA, использовался стандартный формат представления корпуса UCI Bag-of-Words.

В данном формате корпус представляется в виде двух объектов: словаря, матрицы `docword`. Каждая запись `docword` содержит тройку $(d, w, n_{dw})$, где $d$ – идентификатор документа, $w$ – идентификатор слова, $n_{dw}$ – частота слова в документе.

UCI представление корпуса формировалось на основе предварительно отфильтрованной текстовой коллекции с заданным уровнем тематического дисбаланса. Для этого из исходного корпуса выбиралось необходимое количество документов в требуемом соотношении между темами. После этого выполнялась конвертация корпуса в формат `docword` с использованием общего словаря терминов.

Для моделей, использующих контекстные представления слов (CARTM), применялся другой формат хранения данных. В данном случае каждый документ сохранялся как последовательность идентификаторов токенов. Такой способ представления позволяет учитывать локальный порядок слов и использовать контекстную информацию при построении тематических распределений.

Дополнительно сохранялся бинарный массив границ документов и информация о принадлежности документов к категориям корпуса.

Несмотря на различия форматов хранения, для всех моделей использовались одинаковый словарь и единая процедура предобработки текстов. Это позволяет корректно сравнивать результаты моделей и исключает влияние различий текстовой нормализации на итоговые различия тематических структур.

## 3.5. Построение тематических моделей

В рамках экспериментального исследования рассматривались три типа тематических моделей, отличающиеся как базовой вероятностной постановкой, так и учетом регуляризации и локального контекста. Такой выбор позволяет провести сравнительный анализ устойчивости моделей к тематическому дисбалансу при фиксированной предобработке данных и едином словаре.

### 3.5.1. Вероятностная модель PLSA без регуляризации

В качестве базовой модели использовалась вероятностная латентно-семантическая модель PLSA (Probabilistic Latent Semantic Analysis), реализо-

ванная в библиотеке BIGARTM.

Формально вероятность слова в документе задается как:

$$P(w \mid d) = \sum_{t=1}^{T} P(w \mid t) P(t \mid d),$$

где $T$ – число тем, $P(w \mid t)$ – распределение слов в теме, а $P(t \mid d)$ – распределение тем в документе.

Данная модель используется как базовая точка сравнения и не содержит дополнительных механизмов регуляризации, что делает ее чувствительной к дисбалансу тематической структуры.

### 3.5.2. PLSA с регуляризацией (DecorrelatorPhi)

Вторая модель представляет собой расширение PLSA с использованием регуляризаторов библиотеки BIGARTM. В частности, применялся регуляризатор *DecorrelatorPhiRegularizer*, направленный на уменьшение корреляции между тематическими распределениями слов. Основная идея регуляризации заключается в снижении схожести тем за счет штрафа за перекрытие высоковероятных слов в различных темах, что формально повышает различимость тематических распределений $P(w \mid t)$. Использование регуляризации должно позволить частично компенсировать эффект слияния тем, возникающий при наличии дисбаланса в корпусе.

### 3.5.3. Модель локальных контекстов CARTM

Третья модель – CARTM (Context-Aware Recurrent Topic Model) – относится к классу моделей, учитывающих локальный контекст слов через механизм внимания. В отличие от классических тематических моделей, где порядок слов игнорируется, CARTM использует информацию о соседних токенах для уточнения тематического представления.

Реализация модели основана на исходном репозитории [45]. Внутри реализации модель строится как комбинация:

- контекстного энкодинга последовательности токенов,
- механизма внимания для взвешивания соседних слов,

- и последующего тематического распределения.

Таким образом, CARTM позволяет учитывать локальные зависимости в тексте, что потенциально повышает устойчивость модели к шуму и неоднородности тематической структуры корпуса.

## 3.6. Методика оценки устойчивости тематических моделей

Для количественного анализа влияния тематической несбалансированности на структуру тематических моделей была разработана единая методика оценки устойчивости. Методика включает следующие этапы:

1. формирование экспериментальных корпусов с контролируемым уровнем дисбаланса категорий;
2. обучение тематических моделей при фиксированных гиперпараметрах;
3. сопоставление тема модели, обученной на исследуемом корпусе, с темами эталонной модели, обученной на сбалансированном корпусе;
4. вычисление метрик, характеризующих структурные искажения тематического пространства.

В качестве точки отсчёта использовалась эталонная модель, обученная на сбалансированном корпусе. Для оценки устойчивости тематических моделей выполнялось сопоставление тем исследуемых моделей с темами эталонной модели. При анализе воспроизводимости сравнивались модели, обученные на одном и том же сбалансированном корпусе, но с различными случайными инициализациями параметров. При анализе влияния дисбаланса сравнивались модели, обученные на корпусах с различной степенью тематической несбалансированности, с эталонной моделью. Такой подход позволяет разделить эффекты, связанные со случайностью процедуры обучения, и эффекты, обусловленные изменением структуры коллекции документов.

### 3.6.1. Сопоставление тем

Пусть задана эталонная тематическая модель, обученная на сбалансированном корпусе, и исследуемая тематическая модель, обученная на корпусе

с некоторой степенью дисбаланса. Модели представлены матрицами тематических распределений слов

$$\Phi^{(ref)} = \left(\varphi_{wt}^{(ref)}\right), \qquad \Phi^{(exp)} = \left(\varphi_{wt}^{(exp)}\right),$$

где элемент $\varphi_{wt}$ соответствует вероятности слова $w$ в теме $t$, а каждый столбец матрицы представляет собой вероятностное распределение слов внутри темы.

Для оценки близости тем используется косинусное сходство между тематическими распределениями слов.

Пусть

$$\varphi_{\cdot i}^{(ref)}$$

и

$$\varphi_{\cdot j}^{(exp)}$$

обозначают распределения слов для тем эталонной и исследуемой моделей соответственно.

Тогда косинусное сходство между темами определяется как

$$\cos(\varphi_i, \varphi_j) = \frac{\sum_{w \in W} \varphi_{wi}^{(ref)} \varphi_{wj}^{(exp)}}{\sqrt{\sum_{w \in W} (\varphi_{wi}^{(ref)})^2} \sqrt{\sum_{w \in W} (\varphi_{wj}^{(exp)})^2}}.$$

Значения косинусного сходства принадлежат диапазону

$$[0, 1],$$

где большие значения соответствуют более близким темам.

Для всех пар тем формируется матрица сходства

$$S = (s_{ij}), \qquad s_{ij} = \cos(\varphi_i, \varphi_j).$$

На основе матрицы сходства для каждой темы эталонной модели определяется множество тем исследуемой модели, сходство с которыми превышает заданное пороговое значение.

На рисунке 1 приведён пример матрицы сходства между темами эталонной модели и модели, обученной на корпусе с многоклассовым дисбалансом при коэффициенте диспропорции $k = 2$.

Каждая строка матрицы соответствует теме эталонной модели, а каждый столбец — теме исследуемой модели. Значения элементов матрицы отражают величину косинусного сходства между распределениями слов соответствующих тем.

Высокие значения сходства образуют локальные максимумы, соответствующие наиболее близким темам. При этом наличие нескольких высоких значений в одной строке может свидетельствовать о расщеплении исходной темы на несколько тем новой модели. Аналогично, несколько высоких значений в одном столбце могут указывать на слияние нескольких тем эталонной модели.

Подобное представление позволяет визуально анализировать структурные изменения тематического пространства и служит основой для вычисления метрик correspondence, split и merge.

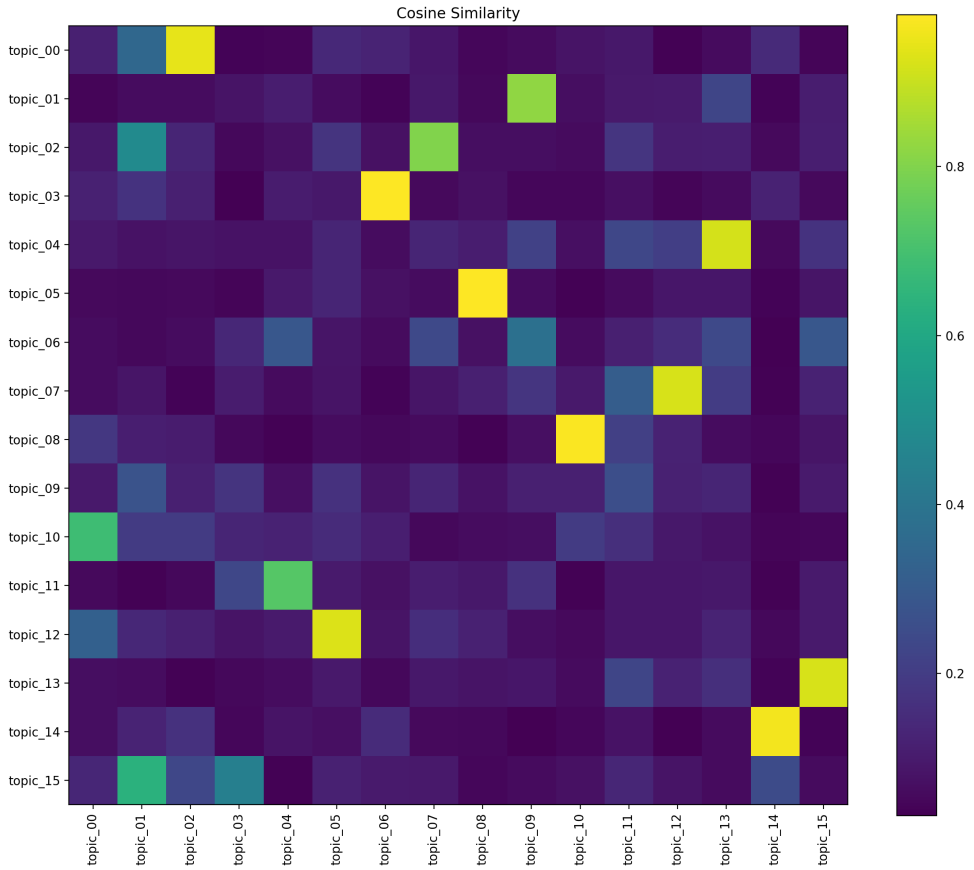



Рис. 1 – Матрица косинусного сходства между темами эталонной модели и модели при многоклассовом дисбалансе ($k = 2$)

### 3.6.2. Выявление структурных искажений

Для каждой темы $t_i^{(1)}$ определяется множество тем второй модели, расстояние до которых не превышает заданный порог соответствия:

$$M_{\text{corr}}(t_i^{(1)}) = \left\{ t_j^{(2)} \;\middle|\; s_{ij} \geq \tau_{\text{corr}} \right\}.$$

На основе полученных соответствий выделяются следующие типы структурных преобразований.

- **Соответствие (correspondence)** — тема первой модели имеет единственную близкую тему во второй модели, и данное соответствие является взаимно однозначным.
- **Расщепление (split)** — одной теме первой модели соответствуют две или более темы второй модели:

$$|M_{\text{corr}}(t_i^{(1)})| \geq 2.$$

- **Слияние (merge)** — нескольким темам первой модели соответствует одна тема второй модели.
- **Исчезновение темы (disappeared topic)** — тема не имеет ни одного соответствия, удовлетворяющего пороговому условию.

Подобная классификация позволяет анализировать не только изменение качества тематической модели, но и характер структурных преобразований тематического пространства при росте дисбаланса.

На рис. 2 приведена схема, иллюстрирующая эффекты расщепления и слияния тем.

### 3.6.3. Метрики устойчивости и структурных искажений

Для количественной оценки поведения моделей используются следующие показатели:

- количество соответствий (*correspondence count*);
- количество расщеплений (*split count*);

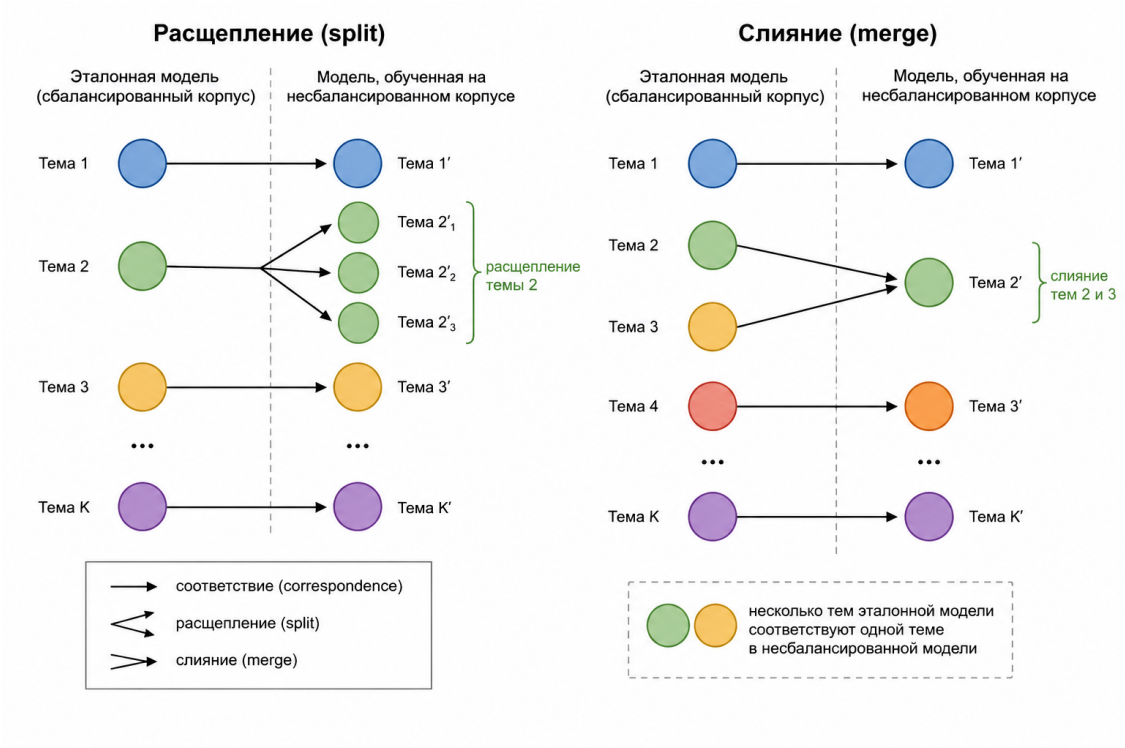



Рис. 2 – Схема расщепления и слияния тем

- степень расщепления (*split degree*);
- количество слияний (*merge count*);
- степень слияния (*merge degree*);
- количество исчезнувших тем (*disappeared count*).

Степень расщепления определяется как суммарное число тем, участвующих в эффектах расщепления:

$$\text{split\_degree} = \sum_{t_i \in Split} |M_{\text{corr}}(t_i)|.$$

Аналогично степень слияния вычисляется как суммарное число тем, участвующих в эффектах слияния:

$$\text{merge\_degree} = \sum_{t_j \in Merge} |M_{\text{merge}}(t_j)|.$$

Для оценки устойчивости каждая модель обучалась несколько раз при различных случайных инициализациях параметров. После этого метрики вычислялись для всех пар запусков, а итоговые значения определялись усред-

нением результатов. Такой подход позволяет отделить эффекты, вызванные тематической несбалансированностью, от случайной вариативности процесса обучения.

На рисунке 3 представлен общий пайплайн проведения экспериментов по исследованию влияния тематической несбалансированности на структуру тематических моделей. Схема включает этапы предобработки текстового корпуса, генерации несбалансированных выборок, обучения тематических моделей, построения матрицы косинусного сходства тем и последующего вычисления метрик структурных искажений тематического пространства.

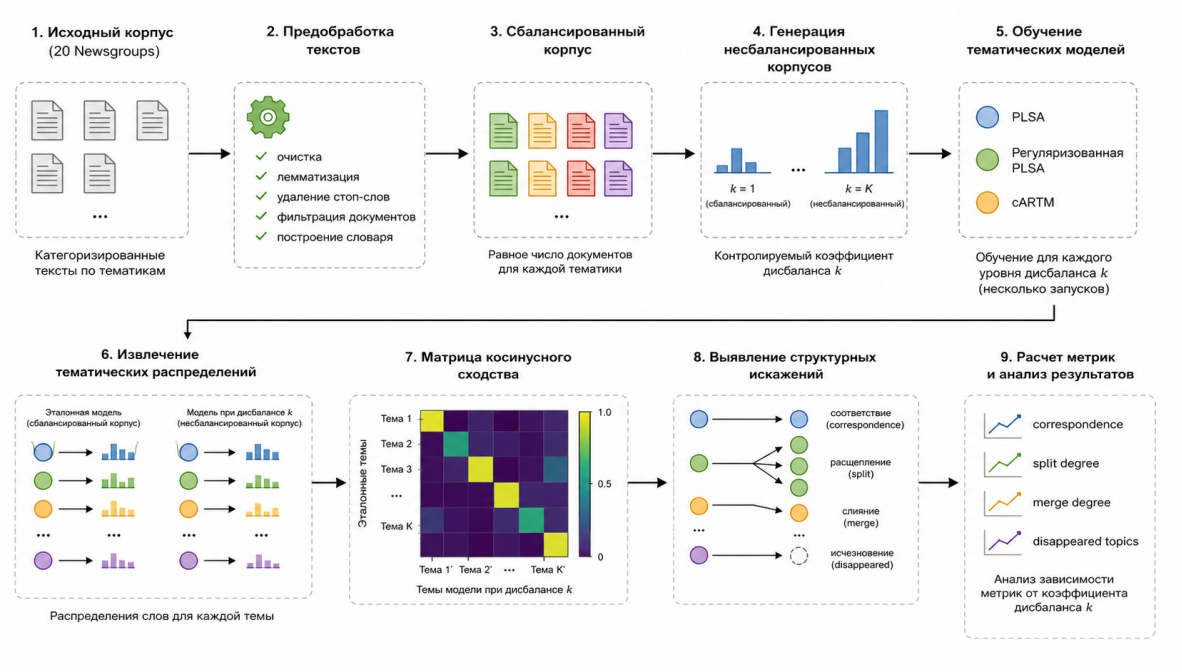



Рис. 3 – Матрица косинусного сходства между темами эталонной модели и модели при многоклассовом дисбалансе ($k = 2$)

## 4. Экспериментальное исследование

В рамках экспериментального исследования была построена и оценена тематическая модель, выступающая в роли эталонного (baseline) решения. Данная модель используется для последующего сравнения с модифицированными подходами, разрабатываемыми в работе.

### 4.1. Эталонная модель и методика оценки ее качества

В качестве эталонной модели была использована базовая тематическая модель, обученная на корпусе документов 20 Newsgroups, включающем 11

тематических категорий. Основная цель данного этапа заключалась в получении устойчивого тематического разбиения корпуса, которое может быть использовано как опорная точка для анализа качества дальнейших модификаций.

Оценка качества эталонной модели проводилась в полуформальном режиме и включала два основных компонента:

- Интерпретируемость тематических распределений, то есть ручной анализ топ-N слов каждой темы с точки зрения их семантической согласованности и соответствия предметной области.
- Согласованность с известными метками документов, то есть проверка того, насколько документы с заранее известной тематической принадлежностью концентрируются в соответствующих темах модели.

Таким образом, оценка носила как качественный (экспертная интерпретация), так и количественно-структурный характер.

На рисунке 4 представлено распределение предсказанных тем по документам корпуса. Вертикальные линии разделяют документы по истинным категориям датасета.

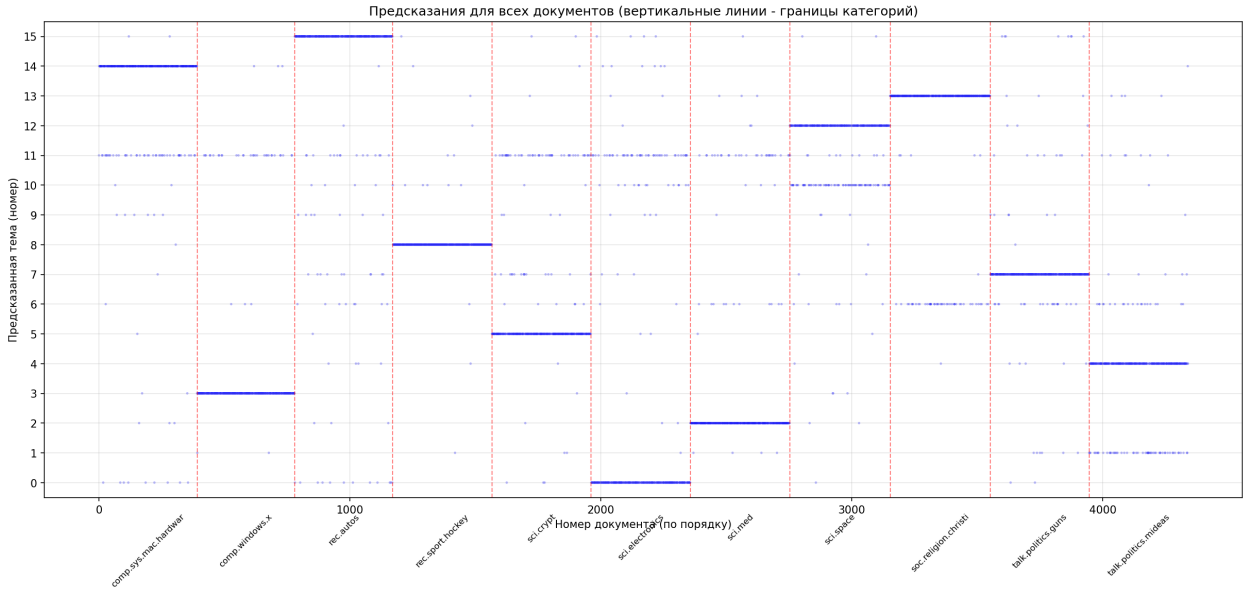



Рис. 4 – Распределение предсказанных тем по документам корпуса 20 Newsgroups

Из представленного распределения видно, что модель в целом формирует устойчивые кластеры документов, соответствующие исходным категориям. Для каждой категории сформировалась наиболее явно выраженная тем.

Данная визуализация подтверждает способность модели выделять семантически однородные группы текстов.

Дополнительно была проведена ручная интерпретация тематических распределений на основе топ-слов. В таблицах ниже представлены наиболее вероятные слова для каждой темы, а также их веса. Например, для темы 0 доминирующими являются термины, связанные с электроникой и сигналами (power, circuit, voltage, signal), что позволяет интерпретировать её как тему, связанную с электронными устройствами и схемотехникой.

Таблица 5 – Интерпретация тем модели по ключевым словам

| Тема | Интерпретация | Ключевые слова | 20newsgroup |
|---|---|---|---|
| 5 | Криптография | encryption, cipher, pgp, key | sci.crypt |
| 8 | Спорт (хоккей) | game, hockey, goal, playoff | rec.sport.hockey |
| 12 | Космос | space, orbit, satellite, nasa | sci.space |
| 15 | Автомобиль | car, save, engine, drive | rec.autos |

Всего было выделено 11 наиболее интерпретируемых тем, которые демонстрируют хорошую семантическую согласованность и соответствие ожидаемым доменам корпуса.

Полученные результаты показывают, что базовая модель демонстрирует удовлетворительное качество тематического разбиения корпуса. Несмотря на наличие пересечений между отдельными темами и некоторую шумность распределений, модель формирует достаточно интерпретируемые и устойчивые тематические кластеры, что позволяет использовать её в качестве эталона для дальнейшего сравнения с модифицированными подходами.

### 4.2. Результаты с дисбалансом одного класса

На рисунке 5a представлены результаты для модели PLSA, на рисунке 5b – для PLSA с регуляризацией, а на рисунке 5c – для контекстно-зависимой модели cARTM.

### 4.2.1. Результаты базовой модели PLSA при одноклассовом дисбалансе

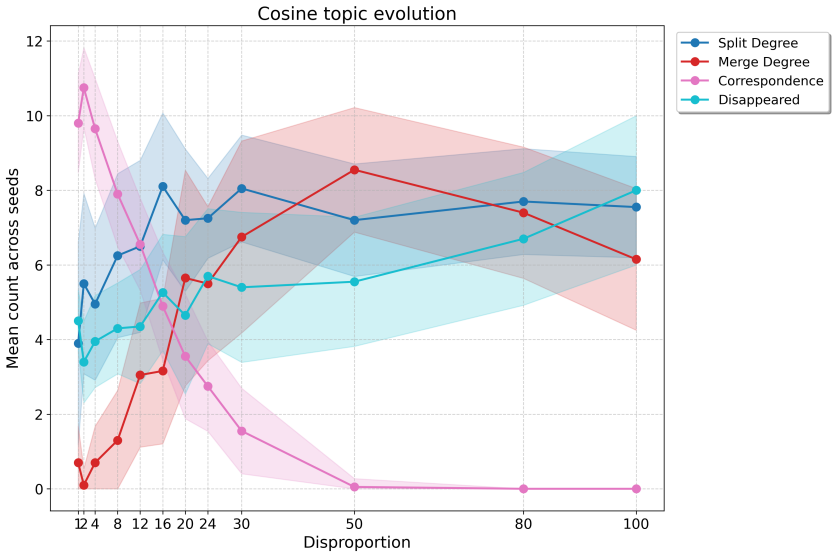
Для базовой модели PLSA наблюдается ожидаемое ухудшение соответствия тематических структур по мере роста дисбаланса классов (рисунок 5а).

Наиболее заметным эффектом является быстрое уменьшение показателя Correspondence. Если при небольших значениях дисбаланса сохраняется около 10–11 взаимно соответствующих тем, то уже при дисбалансе порядка 30 данный показатель снижается до 1–2 тем, а начиная с уровня 50 практически достигает нуля. Это свидетельствует о том, что первоначальная тематическая структура перестаёт воспроизводиться моделью.

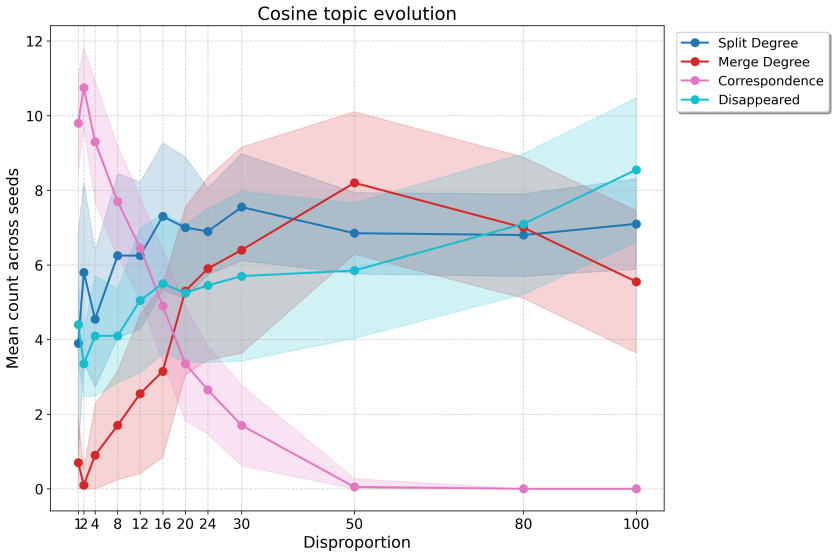
Одновременно наблюдается рост показателей Split Degree и Merge Degree. Среднее число разбиений возрастает примерно с 4-5 до 7-8, а число объединений увеличивается практически до аналогичных значений. Таким образом, вместо стабильного сохранения тем происходит их активная перестройка: одни темы дробятся на несколько новых, тогда как другие объединяются между собой.

Дополнительным подтверждением деградации структуры является увеличение числа исчезнувших тем. При максимальном исследуемом дисбалансе количество тем, не имеющих соответствия в новой модели, достигает примерно восьми. Это означает, что около половины исходных тем фактически теряются в процессе обучения на несбалансированных данных.

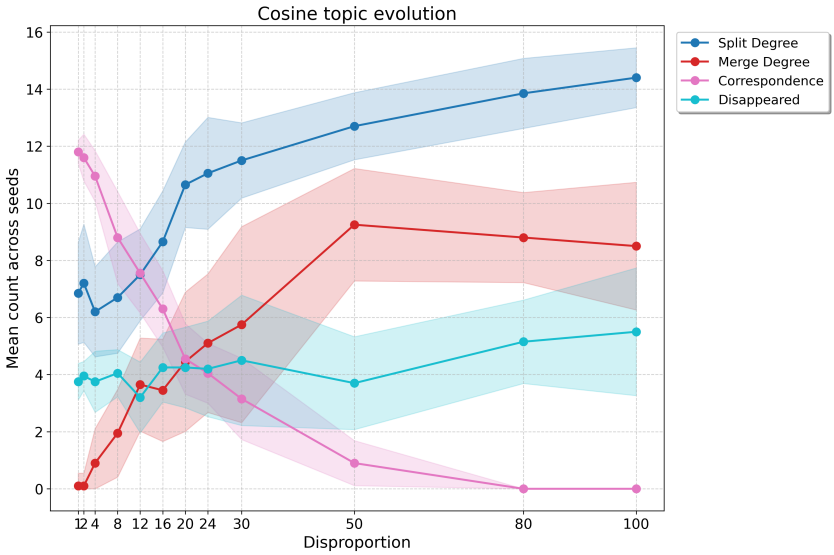
В целом результаты показывают высокую чувствительность классической модели PLSA к нарушению баланса классов.




(a) PLSA




(b) Регуляризованная PLSA




(c) cARTM

Рис. 5 – Изменение метрик эволюции тем при одноклассовом дисбалансе для различных тематических моделей

### 4.2.2. Результаты регуляризованной модели PLSA при одноклассовом дисбалансе

Применение регуляризации не приводит к качественному изменению характера зависимости (рисунок 5b). Основные тенденции сохраняются: Corresponde монотонно уменьшается с ростом дисбаланса, а показатели Split Degree, Merge Degree и Disappeared постепенно возрастают.

Тем не менее можно отметить некоторое снижение вариативности результатов между различными инициализациями модели. Доверительные интервалы для большинства метрик становятся несколько уже по сравнению с обычной PLSA, что указывает на более стабильное поведение алгоритма.

При этом абсолютные значения показателей остаются близкими к результатам базовой модели. Уже при высоких уровнях дисбаланса количество соответствующих тем также стремится к нулю, а число исчезнувших тем достигает значений порядка 8–9. Следовательно, используемая регуляризация способствует стабилизации обучения, однако не решает проблему разрушения тематической структуры при сильном дисбалансе данных.

### 4.2.3. Результаты модели cARTM при одноклассовом дисбалансе

Иная картина наблюдается для контекстно-зависимой модели cARTM (рисунок 5c).

В отличие от моделей PLSA, показатель Split Degree здесь растёт значительно быстрее и достигает значений 14–15 при больших уровнях дисбаланса. Аналогичная тенденция наблюдается и для Merge Degree. На первый взгляд это может свидетельствовать о более сильной перестройке тематической структуры.

Однако одновременно показатель Disappeared остаётся заметно ниже, чем у моделей PLSA. Даже при максимальном уровне дисбаланса число полностью исчезнувших тем не превышает 5–6, тогда как для PLSA данный показатель достигает 8–9. Кроме того, снижение Correspondence происходит более плавно.

Полученные результаты позволяют предположить, что в условиях дисбаланса cARTM чаще сохраняет информацию об исходных темах за счёт их перераспределения между несколькими новыми темами, тогда как PLSA чаще

полностью утрачивает отдельные темы. Иными словами, для cARTM характерна трансформация тематической структуры через процессы расщепления и переорганизации тем, а не через их исчезновение.

### 4.2.4. Сравнение результатов трех моделей при одноклассовом дисбалансе

Сравнение результатов моделей PLSA, регуляризованной PLSA и cARTM показывает, что все исследуемые подходы чувствительны к увеличению дисбаланса классов, однако характер изменения тематической структуры существенно различается.

Для всех моделей наблюдается снижение показателя Correspondence по мере роста дисбаланса, что свидетельствует об уменьшении числа тем, сохраняющих взаимно-однозначное соответствие с темами эталонной модели. При высоких уровнях дисбаланса данный показатель стремится к нулю независимо от используемого алгоритма, что указывает на значительную перестройку тематического пространства.

Базовая модель PLSA и её регуляризованная версия демонстрируют близкое поведение по всем рассматриваемым метрикам. В обоих случаях рост дисбаланса сопровождается увеличением показателей Split Degree и Merge Degree до значений порядка 6–8, а также существенным ростом количества исчезнувших тем. При максимальных значениях дисбаланса число тем, не имеющих значимых соответствий в новой модели, достигает 8–9, что составляет примерно половину от общего числа тем.

Использование регуляризации не приводит к принципиальному изменению динамики метрик. Основным эффектом является некоторое снижение разброса результатов между различными инициализациями модели, что проявляется в более узких доверительных интервалах. Таким образом, регуляризация способствует стабилизации процесса обучения, однако не устраняет влияние дисбаланса на тематическую структуру.

Наиболее заметные отличия наблюдаются для модели cARTM. В отличие от моделей PLSA, значения Split Degree и Merge Degree для неё оказываются существенно выше и продолжают расти по мере увеличения дисбаланса. При этом количество исчезнувших тем остаётся ниже, чем для обеих версий PLSA, даже при максимальных уровнях дисбаланса.

Полученные результаты позволяют сделать вывод о различном характере трансформации тематической структуры в рассматриваемых моделях. Для PLSA увеличение дисбаланса преимущественно сопровождается потерей соответствий между темами и ростом числа исчезнувших тем. Для cARTM изменения проявляются главным образом в увеличении количества связей между темами эталонной и исследуемой моделей, что отражается в высоких значениях Split Degree и Merge Degree при относительно меньшем числе исчезнувших тем.

Таким образом, при одноклассовом дисбалансе контекстно-зависимая модель cARTM демонстрирует более устойчивое сохранение тематической информации по сравнению с классическими вероятностными тематическими моделями.

## 4.3. Результаты с дисбалансом 4 классов

На рисунке 6a представлены результаты для модели PLSA, на рисунке 6b – для PLSA с регуляризацией, а на рисунке 6c – для контекстно-зависимой модели cARTM.

### 4.3.1. Результаты базовой модели PLSA при многоклассовом дисбалансе

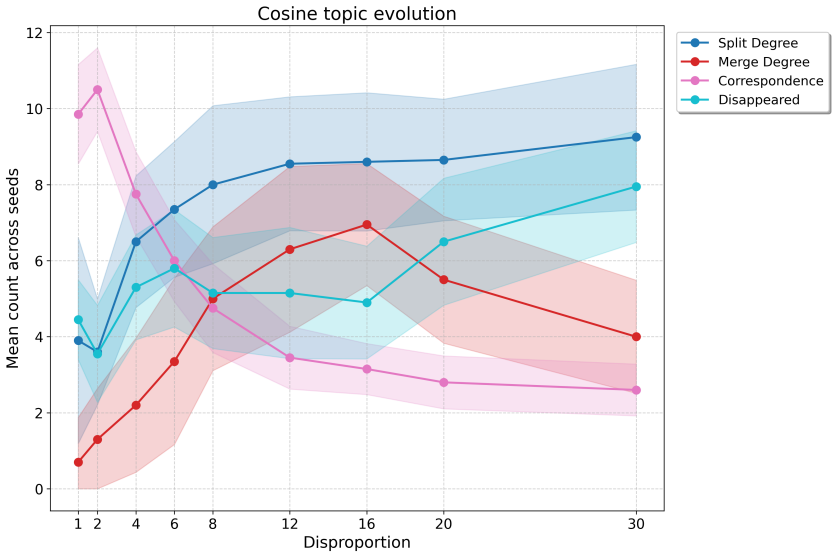
Для базовой модели PLSA при увеличении коэффициента многоклассового дисбаланса наблюдается постепенное изменение структуры тематического пространства (рисунок 6a).

Показатель Correspondence монотонно уменьшается по мере роста дисбаланса. Если при сбалансированном корпусе число взаимно соответствующих тем составляет около 10–11, то при максимальных значениях коэффициента дисбаланса данный показатель снижается приблизительно до 3–4. При этом снижение происходит постепенно, а значения метрики стабилизируются при высоких уровнях дисбаланса.

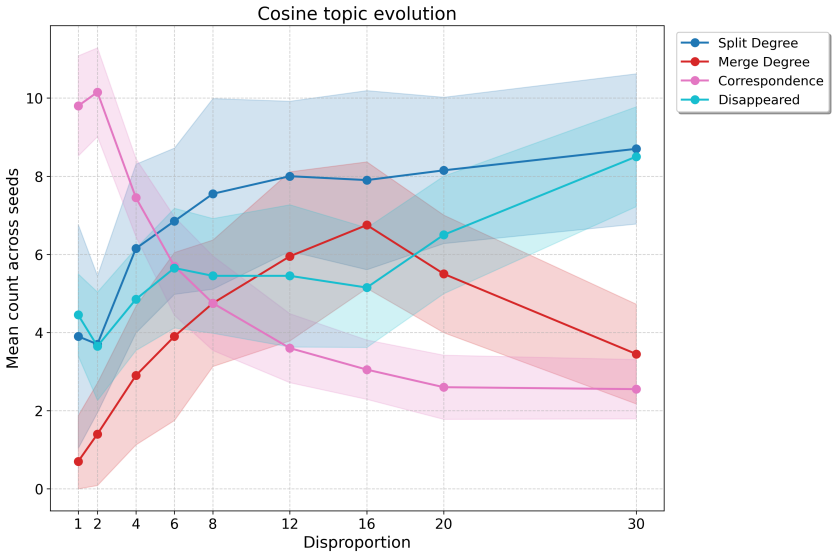
Одновременно наблюдается рост показателей Split Degree и Merge Degree. Значение Split Degree увеличивается примерно с 7–8 до 15, что указывает на активное расщепление исходных тем на несколько новых тем. Показатель Merge Degree также возрастает и достигает значений порядка 5–6, отражая увеличение числа объединений тем.

Количество исчезнувших тем постепенно возрастает с увеличением дисбаланса и достигает примерно шести при максимальном коэффициенте диспропорции. Это указывает на постепенную утрату части тем эталонной модели.

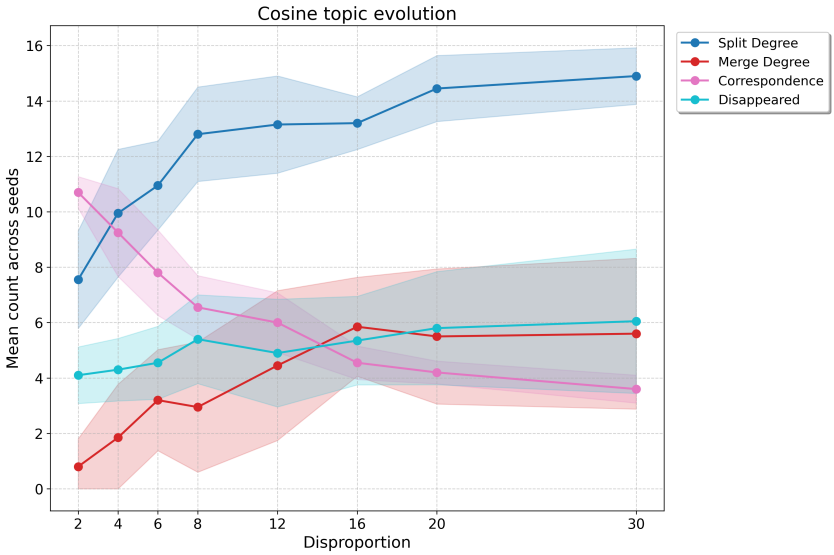
В целом результаты показывают, что рост многоклассового дисбаланса сопровождается существенной перестройкой тематической структуры модели PLSA, проявляющейся в одновременном увеличении числа расщеплений, слияний и исчезнувших тем.




(a) PLSA




(b) Регуляризованная PLSA




(c) cARTM

Рис. 6 – Изменение метрик эволюции тем при многоклассовом дисбалансе для различных тематических моделей

### 4.3.2. Результаты регуляризованной модели PLSA при многоклассовом дисбалансе

Для регуляризованной версии PLSA наблюдаются аналогичные закономерности изменения тематической структуры (рисунок 6b).

Показатель Correspondence постепенно уменьшается с ростом коэффициента дисбаланса и при максимальных значениях диспропорции достигает уровня около 2–3. При этом динамика снижения носит достаточно плавный характер.

Показатели Split Degree и Merge Degree увеличиваются по мере роста дисбаланса. Значения Split Degree возрастают примерно до 9, тогда как Merge Degree достигает значений порядка 7. Это указывает на усиление процессов расщепления и объединения тем при нарушении баланса классов.

Количество исчезнувших тем также возрастает и достигает примерно 8 при максимальном уровне дисбаланса. Следовательно, часть тем эталонной модели перестаёт воспроизводиться при обучении на несбалансированных данных.

Следует отметить, что доверительные интервалы для большинства метрик остаются сравнительно стабильными, что указывает на более устойчивое поведение модели между различными запусками.

### 4.3.3. Результаты модели cARTM при многоклассовом дисбалансе

Контекстно-зависимая модель cARTM также демонстрирует изменение тематической структуры по мере роста многоклассового дисбаланса (рисунок 6b).

Показатель Correspondence постепенно уменьшается с увеличением коэффициента дисбаланса и при максимальных значениях принимает значения порядка 2–3. Снижение происходит плавно, без резких скачков метрики.

Показатели Split Degree и Merge Degree постепенно возрастают. Значение Split Degree увеличивается примерно до 8–9, тогда как Merge Degree достигает значений порядка 6–7. Это свидетельствует о постепенном увеличении числа расщеплений и слияний тем при росте дисбаланса.

Количество исчезнувших тем также возрастает и при максимальном коэффициенте дисбаланса достигает примерно 8–9. Таким образом, при силь-

ном многоклассовом дисбалансе часть тем эталонной модели перестаёт иметь устойчивые соответствия.

Следует отметить, что изменение всех рассматриваемых метрик происходит достаточно плавно, а доверительные интервалы остаются относительно стабильными на различных уровнях дисбаланса.

### 4.3.4. Сравнение результатов трех моделей при многоклассовом дисбалансе

Результаты экспериментов показывают, что увеличение многоклассового дисбаланса приводит к существенному изменению тематической структуры для всех исследуемых моделей.

Для всех моделей наблюдается уменьшение показателя Correspondence по мере роста коэффициента дисбаланса, что указывает на сокращение числа взаимно соответствующих тем между эталонной и исследуемыми моделями.

Одновременно для всех моделей возрастают показатели Split Degree и Merge Degree, отражающие усиление процессов расщепления и объединения тем. Также наблюдается постепенный рост количества исчезнувших тем.

Наиболее высокие значения Split Degree демонстрирует базовая модель PLSA, для которой показатель достигает значений порядка 15 при максимальных уровнях дисбаланса. Регуляризованная версия PLSA показывает сходную динамику метрик, однако характеризуется несколько меньшей вариативностью результатов между различными запусками.

Модель cARTM демонстрирует более плавное изменение показателей и сравнительно стабильные доверительные интервалы. При этом для всех моделей сохраняется общая тенденция: увеличение тематической несбалансированности сопровождается перестройкой тематического пространства, ростом числа расщеплений и слияний тем, а также постепенным исчезновением части тем эталонной модели.

# 5. Обсуждение результатов

## 5.1. Интерпретация показателей эволюции тем при одноклассовом дисбалансе

При анализе полученных результатов необходимо учитывать особенности используемых метрик. В отличие от традиционных показателей качества тематического моделирования, метрики Split Degree и Merge Degree характеризуют не качество модели как таковое, а степень изменения тематической структуры относительно эталонной модели.

Рост показателя Split Degree не обязательно свидетельствует об ухудшении качества тематической модели. Высокое значение данной метрики означает, что одной теме эталонной модели соответствует несколько тем в исследуемой модели. Такая ситуация может возникать как при разрушении темы, так и при её детализации, когда исходная тема сохраняется, но распределяется между несколькими более специализированными темами.

Аналогично, увеличение Merge Degree показывает, что несколько тем эталонной модели оказались представлены одной темой новой модели. Данный эффект отражает укрупнение тематических областей, однако сам по себе не означает полной потери содержащейся в них информации.

Поэтому показатели Split Degree и Merge Degree следует рассматривать совместно с метриками Correspondence и Disappeared. Особый интерес представляет показатель Disappeared, отражающий количество тем эталонной модели, не имеющих значимых соответствий в исследуемой модели. В отличие от операций разделения и объединения, исчезновение темы означает фактическую потерю соответствующего семантического направления.

Именно поэтому результаты модели cARTM требуют отдельной интерпретации. Несмотря на то, что значения Split Degree и Merge Degree для неё существенно выше, чем для моделей PLSA, количество исчезнувших тем остаётся заметно ниже. Это указывает на то, что при увеличении дисбаланса cARTM чаще сохраняет информацию об исходных темах, перераспределяя её между несколькими новыми темами, тогда как модели PLSA чаще полностью утрачивают часть тематической структуры.

Таким образом, увеличение показателей Split Degree и Merge Degree в случае cARTM следует интерпретировать скорее как признак активной

реорганизации тематического пространства, чем как непосредственное ухудшение качества модели. С точки зрения устойчивости к дисбалансу более важным является сохранение семантического содержания тем, что косвенно подтверждается меньшим количеством исчезнувших тем.

## 5.2. Интерпретация показателей эволюции тем при многоклассовом дисбалансе

Полученные результаты показывают, что многоклассовый дисбаланс приводит к постепенной перестройке тематического пространства для всех исследуемых моделей. При увеличении коэффициента дисбаланса уменьшается количество взаимно соответствующих тем, одновременно возрастает число расщеплений, слияний и исчезнувших тем.

При интерпретации результатов необходимо учитывать особенности используемых метрик. Показатели Split Degree и Merge Degree характеризуют не качество тематической модели как таковое, а степень изменения структуры тематического пространства относительно эталонной модели.

Рост Split Degree означает, что одной теме эталонной модели начинает соответствовать несколько тем исследуемой модели. Подобное поведение отражает расщепление исходных тем и увеличение числа связей между тематическими распределениями. Аналогично рост Merge Degree показывает, что несколько тем эталонной модели начинают отображаться в одну тему новой модели, что соответствует укрупнению тематических областей.

Особый интерес представляет поведение показателя Correspondence. В отличие от сценария с одной доминирующей темой, при многоклассовом дисбалансе значение Correspondence не стремится к нулю даже при высоких уровнях дисбаланса, а постепенно выходит на плато. Это связано с тем, что в корпусе одновременно присутствует несколько крупных тематических кластеров, которые продолжают устойчиво воспроизводиться моделью.

Таким образом, сохраняющиеся соответствия преимущественно отражают наиболее представленные тем корпуса. При этом менее представленные темы постепенно теряют устойчивость, расщепляются, объединяются с другими темами либо полностью исчезают.

Полученные результаты позволяют предположить, что многоклассовый дисбаланс приводит не к полному разрушению тематического пространства,

а к его структурной поляризации. В процессе обучения модели начинают преимущественно сохранять и воспроизводить наиболее крупные тематические области, тогда как малые темы постепенно вытесняются из структуры модели.

Дополнительно следует отметить различие характера перестройки тематического пространства у различных моделей. Для базовой PLSA наиболее характерен интенсивный рост Split Degree, что указывает на сильное дробление тем. Регуляризованная модель демонстрирует схожее поведение, однако характеризуется меньшей вариативностью между различными запусками. Модель cARTM показывает более плавное изменение метрик, что может свидетельствовать о более устойчивом перераспределении тематической информации за счёт использования локального контекста слов.

## 5.3. Сравнение результатов одноклассового и многоклассового дисбаланса

Сравнение результатов показывает, что характер изменения тематической структуры существенно зависит от типа тематического дисбаланса.

В случае одноклассового дисбаланса в корпусе формируется одна доминирующая тема, которая начинает постепенно вытеснять остальные темы. При увеличении коэффициента дисбаланса наблюдается быстрое уменьшение показателя Correspondence, а также существенный рост количества исчезнувших тем. Для моделей PLSA и регуляризованной PLSA при высоких уровнях дисбаланса число взаимно соответствующих тем практически стремится к нулю. Это свидетельствует о почти полном разрушении исходной тематической структуры.

При многоклассовом дисбалансе наблюдается иной характер поведения моделей. Поскольку в корпусе одновременно присутствует несколько крупных тематических кластеров, тематическое пространство сохраняет несколько устойчивых центров. В результате показатель Correspondence уменьшается значительно более плавно и при высоких уровнях дисбаланса выходит на плато, соответствующее числу устойчиво воспроизводимых доминирующих тем.

Таким образом, при многоклассовом дисбалансе происходит не полное разрушение тематической структуры, а её структурная поляризация. Наиболее представленные темы продолжают устойчиво воспроизводиться моделью, тогда как менее представленные темы постепенно исчезают или поглощаются

более крупными тематическими кластерами.

Следует учитывать, что в сценариях одноклассового и многоклассового дисбаланса использовались различные диапазоны коэффициента дисбаланса. Для одноклассового сценария исследовались значения до (k=100), тогда как для многоклассового дисбаланса максимальное значение составляло (k=30). Поэтому прямое сравнение абсолютных значений метрик между двумя экспериментами требует осторожности. Тем не менее качественный характер изменения тематической структуры в обоих сценариях демонстрирует общую тенденцию: увеличение тематической несбалансированности приводит к усилению процессов расщепления, слияния и исчезновения тем.

Для всех исследуемых моделей наиболее устойчивое поведение демонстрирует cARTM. В обоих сценариях модель показывает более плавное изменение метрик и меньшую вариативность результатов между различными запусками. Это может свидетельствовать о том, что использование локального контекста слов позволяет модели дольше сохранять тематическую информацию при росте дисбаланса.

В целом результаты экспериментов подтверждают выдвинутую гипотезу о том, что тематическая несбалансированность приводит к систематическим структурным искажениям тематических моделей. Основными проявлениями данных искажений являются расщепление крупных тем, слияние малых тем и постепенное исчезновение части тематического пространства.

## 5.4. Ограничения работы

Представленная методика и полученные результаты имеют ряд ограничений, которые следует учитывать при интерпретации выводов и планировании будущих исследований.

Прежде всего, эксперименты выполнены на корпусах, искусственно сбалансированных или несбалансированных на основе категорий `20 Newsgroups`. Такой подход позволяет строго контролировать степень дисбаланса, но не полностью воспроизводит все особенности реальных текстов (например, естественное перекрытие тем, неоднородность длины документов, шум). Дополнительная валидация на других реалистичных корпусах необходима для подтверждения общности выводов.

Кроме того, используемые метрики расщепления, слияния и соответ-

ствия чувствительны к выбранным порогам косинусной близости. В работе применены фиксированные значения, обоснованные экспериментально, однако автоматическая процедура подбора порогов не предлагается. Для других коллекций пороги, скорее всего, потребуют перенастройки. Анализ устойчивости (воспроизводимости) тем проводился для одного типа инициализации модели и фиксированного числа тем. Влияние различных способов инициализации, алгоритмов оптимизации и вариаций числа тем на стабильность в данной работе не исследовалось.

Также стоит учитывать, что все эксперименты выполнены на англоязычных текстах. Применение разработанной методики к другим языкам (в том числе к русскому) потребует адаптации предобработки (стемминг, лемматизация, стоп-слова) и, возможно, калибровки метрик.

## 5.5. Направления дальнейших исследований

Разработанная методика и полученные результаты открывают несколько направлений для продолжения работы, непосредственно связанных с проблемой тематической несбалансированности.

Прежде всего, представляет интерес исследование влияния различных регуляризаторов, доступных в библиотеке BigARTM, на устойчивость модели к дисбалансу. В данной работе использовался регуляризатор `DecorrelatorPhi`, направленный на разреживание матрицы тем. Однако существуют и другие регуляризаторы, например, `SmoothPhi`, `SmoothTheta` или `TopicsPrior`, которые потенциально могут по-разному влиять на расщепление крупных тем и слияние мелких. Сравнительный анализ их эффективности в условиях контролируемого дисбаланса позволил бы выработать практические рекомендации по выбору регуляризующего штрафа.

Другим перспективным направлением является применение моделей с выделением фоновых тем (background topics). Такие модели отделяют общую лексику (фоновые распределения слов) от содержательных тем, что может снизить влияние дисбаланса: крупная тема перестанет «захватывать» фоновые слова, создавая дублирующие темы-спутники. В рамках разработанной инструментальной среды можно экспериментально проверить, уменьшает ли явное моделирование фоновой лексики степень расщепления.

Возможно также модифицировать веса слов при построении матрицы

сходства тем. Например, можно понижать вклад высокочастотных слов, характерных для доминирующей темы, чтобы слабые темы получали большее влияние при расчёте близости распределений. Альтернативно, можно применять TF-IDF-взвешивание к вероятностям $\varphi_{wt}$ или использовать другие схемы перевзвешивания, которые делают редкие слова более значимыми при сопоставлении тем.

Наконец, требует изучения вопрос о выборе метрики близости между распределениями слов. В текущей работе использовалось косинусное расстояние как простая и симметричная мера. Однако для вероятностных распределений более естественной является KL-дивергенция или её симметричная версия (Jensen–Shannon). Переход на KL может изменить пороги срабатывания расщеплений и слияний, а также повлиять на чувствительность метода к редким темам. Сравнение результатов, полученных с разными метриками, позволило бы выбрать наиболее адекватный инструмент для оценки структурных искажений.

## 6. Заключение

В рамках данной работы была исследована проблема тематической несбалансированности вероятностных тематических моделей. Проведённый обзор литературы показал, что несмотря на широкое распространение тематического моделирования, влияние дисбаланса тем на структуру тематического пространства остаётся недостаточно изученным. Большинство существующих подходов ориентировано либо на повышение интерпретируемости тем, либо на улучшение вероятностных характеристик модели, тогда как структурные эффекты, возникающие при нарушении баланса тем, исследованы существенно меньше.

Для исследования данной проблемы была разработана методика количественного анализа структурных изменений тематических моделей. Предложенный подход основан на сопоставлении тематических распределений моделей с использованием косинусного сходства и последующем анализе эффектов correspondence, split, merge и disappeared topics. В отличие от традиционных метрик качества тематического моделирования, используемый подход позволяет анализировать не только внутреннее качество тем, но и изменение структуры тематического пространства при изменении распределения доку-

ментов между темами.

В работе была разработана программная инструментальная среда для генерации несбалансированных корпусов, обучения тематических моделей и автоматического анализа изменений тематической структуры. Эксперименты проводились на корпусах, сформированных на основе датасета 20 Newsgroups с контролируемой степенью тематической несбалансированности. Для моделирования дисбаланса использовались как сценарии с одной доминирующей категорией, так и сценарии с несколькими крупными тематическими кластерами.

Дополнительно была проведена предобработка корпуса, включавшая удаление технического шума, uuencoded-фрагментов, лемматизацию, фильтрацию коротких документов и построение общего словаря терминов. Это позволило минимизировать влияние шумовых факторов на результаты тематического моделирования и обеспечить сопоставимость экспериментов.

В ходе исследования были проанализированы три типа тематических моделей: базовая модель PLSA, регуляризованная версия PLSA и контекстно-зависимая модель cARTM. Для всех моделей была проведена серия экспериментов при различных уровнях тематического дисбаланса.

Полученные результаты подтверждают выдвинутую гипотезу о том, что увеличение тематической несбалансированности приводит к систематическим структурным искажениям тематического пространства. Основными проявлениями данных искажений являются:

- расщепление крупных тем;
- слияние малых тем;
- постепенное исчезновение части тематического пространства.

Показано, что характер деградации тематической структуры зависит от типа дисбаланса. При одноклассовом дисбалансе наблюдается почти полное разрушение исходной тематической структуры при высоких уровнях диспропорции. Доминирующая тема постепенно вытесняет остальные темы, что приводит к исчезновению значительной части исходного тематического пространства.

При многоклассовом дисбалансе наблюдается иной характер поведения моделей. Вместо полного разрушения тематической структуры происходит её

структурная поляризация: наиболее представленные темы продолжают устойчиво воспроизводиться моделью, тогда как малые темы постепенно исчезают или поглощаются более крупными тематическими кластерами.

Сравнительный анализ моделей показал, что использование локального контекста слов в модели cARTM позволяет более плавно сохранять тематическую информацию при росте дисбаланса по сравнению с классическими вероятностными моделями PLSA. При этом даже контекстно-зависимые модели не устраняют полностью эффекты тематической несбалансированности, а лишь изменяют характер перестройки тематического пространства.

Отдельно следует отметить, что регуляризация в рамках ARTM позволяет снизить вариативность результатов между различными запусками модели, однако сама по себе не предотвращает структурную деградацию тематического пространства при сильном тематическом дисбалансе.

Практическая значимость работы заключается в том, что предложенная методика позволяет количественно анализировать устойчивость тематических моделей и выявлять структурные эффекты, которые не фиксируются традиционными метриками качества, такими как perplexity или topic coherence. Разработанный подход может применяться для анализа устойчивости тематических моделей в задачах обработки научных публикаций, новостных коллекций, социальных сетей и других несбалансированных текстовых корпусов.

Полученные результаты демонстрируют, что тематическая несбалансированность является важным фактором, влияющим на устойчивость и интерпретируемость тематических моделей. Разработанная методика может быть использована для дальнейшего анализа устойчивости тематических моделей, а также для исследования новых методов компенсации эффектов тематического дисбаланса.

В качестве направлений дальнейших исследований представляет интерес анализ более сложных сценариев тематической несбалансированности, исследование нейросетевых тематических моделей, а также разработка специализированных методов регуляризации, направленных непосредственно на компенсацию эффектов split и merge в тематическом пространстве.

**Основные результаты, выносимые на защиту**

- Методика количественного измерения эффектов несбалансированно-

сти и неустойчивости тематических моделей на основе синтетического расширения коллекции категоризированных документов.

- Инструментальная среда для сравнения тематических моделей по устойчивости (воспроизводимости) тем в условиях контролируемой тематической несбалансированности коллекции.
- Модель, улучшающая устойчивость (воспроизводимость) тем в условиях тематической несбалансированности коллекции

## Список литературы


[1] David Blei, Lawrence Carin, and David Dunson. Probabilistic topic models. *IEEE signal processing magazine*, 27(6):55–65, 2010.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and ... Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[3] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57. ACM, 1999.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] Aziida Nanyonga and Graham Wild. Analyzing aviation safety narratives with lda, nmf and plsa: A case study using socrata datasets. *arXiv preprint arXiv:2501.01690*, 2025.

[6] Aziida Nanyonga, Hassan Wasswa, Ugur Turhan, Keith Joiner, and Graham Wild. Exploring aviation incident narratives using topic modeling and clustering techniques. In *2024 IEEE Region 10 Symposium (TENSYMP)*, pages 1–6. IEEE, 2024.

[7] Aziida Nanyonga, Keith Joiner, Ugur Turhan, and Graham Wild. Does the choice of topic modeling technique impact the interpretation of aviation incident reports? a methodological assessment. *Technologies*, 13(5):209, 2025.

[8] Aziida Nanyonga, Hassan Wasswa, and Graham Wild. Topic modeling analysis of aviation accident reports: A comparative study between lda and nmf models. In *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, pages 1–2. IEEE, 2023.

[9] Ju Hyun Lee and Michael J Ostwald. Latent dirichlet allocation (lda) topic models for space syntax studies on spatial experience. *City, Territory and Architecture*, 11(1):3, 2024.

[10] Uttam Chauhan and Apurva Shah. Topic modeling using latent dirichlet allocation: A survey. *ACM Computing Surveys (CSUR)*, 54(7):1–35, 2021.

[11] Hui Yin, Xiangyu Song, Shuiqiao Yang, and Jianxin Li. Sentiment analysis and topic modeling for covid-19 vaccine discussions. *World Wide Web*, 25(3):1067–1083, 2022.

[12] Rubayyi Alghamdi and Khalid Alfalqi. A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications*, 6(1), 2015.

[13] Biyi Fang, Truong Vo, Kripa Rajshekhar, and Diego Klabjan. Topic analysis with side information: A neural-augmented lda approach. *arXiv preprint arXiv:2510.24918*, 2025.

[14] А. Н. Тихонов and В. Я. Арсенин. *Методы решения некорректных задач*. Наука, Москва, 1986. Учебное пособие, 288 с. (третье издание).

[15] К. В. Воронцов and А. А. Потапенко. Аддитивная регуляризация тематических моделей коллекций текстовых документов. *Доклады РАН*, 455, 2014.

[16] И. А. Ирхин, В. Г. Булатов, and К. В. Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста. *Computer Research and Modeling*, 12(6):1515–1528, 2020.

[17] К. В. Воронцов. *Вероятностное тематическое моделирование: Теория регуляризации ARTM и библиотека с открытым кодом BigARTM*. URSS, Москва, 2025.

[18] Caitlin Doogan, Wray Buntine, and Henry Linger. A systematic review of the use of topic models for short text social media analysis. *Artificial Intelligence Review*, 56:14223–14255, 2023.

[19] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*, pages 1105–1112. ACM, 2009.

[20] Corentin Masson and Patrick Paroubek. Evaluating topic model on asymmetric and multi-domain financial corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6515–6529, 2024.

[21] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *23rd Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada*, pages 1973–1981, 2009.

[22] Amin Hosseiny Marani and Eric PS Baumer. A review of stability in topic modeling: Metrics for assessing and techniques for improving stability. *ACM Computing Surveys*, 56(5):1–32, 2023.

[23] Saranzaya Magsarjav, Melissa Humphries, Jonathan Tuke, and Lewis Mitchell. Quantifying consistency and accuracy of latent dirichlet allocation. *arXiv preprint arXiv:2511.12850*, 2025.

[24] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 2009.

[25] Eugeniia Veselova and Konstantin Vorontsov. Topic balancing with additive regularization of topic models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 59–65, 2020.

[26] Shusei Eshima, Kosuke Imai, and Tomoya Sasaki. Keyword-assisted topic models. *American Journal of Political Science*, 68(2):730–750, 2024.

[27] Yuan Zuo, Jichang Zhao, and Ke Xu. Word network topic model: A simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, 48(2):379–398, 2016.

[28] Chengjie Ma, Junping Du, Yingxia Shao, Ang Li, and Zeli Guan. A rare topic discovery model for short texts based on co-occurrence word network. *arXiv preprint*, 2022.

[29] Chengjie Ma, Junping Du, Meiyu Liang, and Zeli Guan. Topic model based on co-occurrence word networks for unbalanced short text datasets. *arXiv preprint*, 2023.

[30] Qiaozhu Mei and ChengXiang Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 198–207, 2005.

[31] Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. A survey on neural topic models: methods, applications, and challenges. *arXiv preprint arXiv:2401.15351*, 2024.

[32] Yee Whye Teh and Michael I Jordan. Hierarchical bayesian nonparametric models with applications. *Bayesian nonparametrics*, 1:158–207, 2010.

[33] Margaret E Roberts, Brandon M Stewart, and Dustin Tingley. Stm: An r package for structural topic models. *Journal of statistical software*, 91:1–40, 2019.

[34] Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. Structural topic models for open-ended survey responses. *American journal of political science*, 58(4):1064–1082, 2014.

[35] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213, 2012.

[36] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22, 2009.

[37] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108, 2010.

[38] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, 2014.

[39] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408, 2015.

[40] Jason Chuang, Sonal Gupta, Christopher Manning, and Jeffrey Heer. Topic model diagnostics: Assessing domain relevance via topical alignment. In *International conference on machine learning*, pages 612–620. PMLR, 2013.

[41] Christopher D Manning. *Introduction to information retrieval*. Syngress Publishing,, 2008.

[42] Fazlollah M Reza. *An introduction to information theory*. Courier Corporation, 1994.

[43] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[44] Ken Lang. Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995*, pages 331–339. Elsevier, 1995.

[45] Revit3D. Topic modeling with attention (cartm). `https://github.com/revit3d/topic-modelling-attention`, 2023. Accessed: 2026-06-04.