МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«САРАТОВСКИЙНАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра информатики и программирования

# Разработка рекомендательной системы для приложения по агрегации текстового контента

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса, 441 группы

направления 02.03.03 «Математическое обеспечение и администрирование информационных систем»

факультета компьютерных наук и информационных технологий

Устимова Матвея Дмитриевича

Научный руководитель

к.ф.-м.н., доцент _____ Огнева М. В.

Зав. кафедрой

к.ф.-м.н., доцент _____ Огнева М. В.

Саратов 2026

# ВВЕДЕНИЕ

Стремительный рост объёмов текстовой информации в сети Интернет ставит перед разработчиками цифровых сервисов задачу эффективной фильтрации и структурирования контента. В условиях информационной перегрузки агрегаторы текстового контента приобретают особое значение как инструменты организации потоков данных. Качество пользовательского опыта в таких системах напрямую определяется способностью предлагать точные и персонализированные рекомендации, влияющие на ключевые показатели вовлечённости аудитории.

Рекомендательные системы обеспечивают персонализированный подбор контента на основе предпочтений пользователя и характеристик самих материалов. Для текстовых агрегаторов задача усложняется необходимостью обработки неструктурированных данных, извлечения семантически значимых признаков и учёта динамично меняющихся интересов. Отдельную проблему представляет оценка качества рекомендательной системы до начала её фактической эксплуатации: классические метрики автономной оценки требуют эталонной разметки релевантности, которая недоступна до накопления журнала реальных взаимодействий пользователей.

Цель выпускной квалификационной работы: разработать рекомендательную систему и сервис формирования персонализированной ленты для приложения-агрегатора текстового контента.

Задачи:

1. Провести анализ современных методов обработки естественного языка и алгоритмов построения рекомендательных систем.
2. Собрать и подготовить корпус текстовых данных из каналов мессенджера.
3. Реализовать конвейер обработки контента, включающий категоризацию и построение векторных представлений текстов.
4. Разработать шестикомпонентную формулу расчёта оценки релевантности,

объединяющую тематическую, семантическую, основанную на именованных сущностях, тональную, временную и форматную составляющие.

5. Реализовать постпроцессор разнообразия выдачи.
6. Спроектировать сервис формирования персонализированной ленты с кэшированием и резервной цепочкой обработки.
7. Разработать методику экспериментальной оценки системы в условиях отсутствия журнала реальных взаимодействий.

Основной экспериментальный корпус содержит более 152 тысяч русскоязычных публикаций из 153 каналов мессенджера Telegram за период с мая по ноябрь 2025 года.

Выпускная квалификационная работа состоит из введения, четырёх разделов, заключения, списка литературы и пяти приложений с исходными кодами. В первом разделе выполнен анализ возможных решений. Во втором разделе описано извлечение признаков из текстового контента. Третий раздел содержит экспериментальное исследование. Четвёртый раздел посвящён реализации рекомендательной системы.

## КРАТКОЕСОДЕРЖАНИЕ РАБОТЫ

В разделе 1 рассмотрены теоретические основы и современные подходы к построению рекомендательных систем для агрегаторов текстового контента. Проведена систематизация методов обработки естественного языка, алгоритмов персонализации, метрик оценки качества, архитектурных паттернов промышленных систем и способов экспериментальной оценки в условиях отсутствия эталонной разметки.

В области методов обработки естественного языка рассмотрены последовательные этапы преобразования текстовых данных: токенизация, нормализация посредством лемматизации или стемминга, а также различные методы векторизации. Для русскоязычных текстов с их развитой морфологической системой обоснован приоритет лемматизации перед стеммингом. Среди методов векторизации описаны классический подход TF-IDF с косинусной мерой сходства, нейросетевые модели слов Word2Vec и FastText, а также трансформерные модели типа BERT, формирующие контекстуализированные векторные представления с учётом двунаправленного контекста. Отдельное внимание уделено извлечению именованных сущностей и тематическому моделированию посредством методов латентного размещения Дирихле и латентно-семантического анализа.

В части алгоритмов рекомендательных систем представлена трёхчастная классификация: коллаборативная фильтрация, фильтрация на основе содержания и гибридные методы. Подробно рассмотрен метод мультикритериальной линейной свёртки, при котором итоговая оценка релевантности вычисляется как взвешенная сумма частных показателей: тематического сходства, семантической близости векторных представлений, совпадения именованных сущностей, тональности, свежести публикации и формата материала. Подчёркнуто преимущество этой схемы с точки зрения интерпретируемости по сравнению с нейросетевыми ранжирующими

моделями. Для метрик оценки качества изложена система показателей, охватывающая точностные характеристики на первых K позициях, метрики позиционного ранжирования (NDCG, MAP), а также качественные свойства выдачи — разнообразие, новизну и охват каталога.

В разделе об архитектурных паттернах описана трёхслойная декомпозиция промышленных рекомендательных систем: слой лингвистической обработки, слой подбора кандидатов и ранжирования, слой доставки контента. Рассмотрена двухстадийная схема подбора кандидатов на основе приближённого поиска ближайших соседей с векторными индексами. Завершающий подраздел посвящён пяти методам оценки в условиях отсутствия эталонной разметки, образующим систему методической триангуляции: косвенные объективные метрики выдачи, сценарное тестирование на синтетических профилях, оценка языковой моделью в роли независимого асессора, проверка контрактных свойств и параметрические переборы гиперпараметров.

В разделе 2 описана методология извлечения признаков из текстового контента и представлены результаты её применения к реальному корпусу данных. Исходный корпус формировался автоматизированными средствами сбора публичного контента в период с мая по ноябрь 2025 года и насчитывает 153 919 текстовых записей из 48 информационных каналов. Медианная длина текста составляет 381 символ, что типично для новостных каналов в мессенджерах. Перед обработкой тексты проходили очистку: нормализацию пробельных символов, удаление служебной разметки, проверку кодировки; записи с нулевой или критически малой длиной исключались.

Разработанный конвейер обработки является многоэтапной системой, интегрирующей методы обработки естественного языка и машинного обучения. На первом этапе извлекаются базовые статистические характеристики: длина текста, количество слов и предложений, доля заглавных букв, лексическое разнообразие и расчётное время прочтения. Второй этап предполагает глубокий лингвистический анализ с

предобученной моделью для русского языка: токенизация, морфологический и синтаксический разбор, распознавание именованных сущностей трёх типов — персон, организаций и географических локаций. Для оценки сложности восприятия применяется адаптированная для русского языка формула удобочитаемости.

Тематическая классификация, определение типа и стиля контента реализованы методом классификации без обучающей разметки на основе модели RuBERT-NLI: публикации сопоставляются с 18 тематическими категориями, 10 типами контента и 5 стилистическими характеристиками. Определение тональности осуществляется специализированной моделью, дообученной на задаче классификации эмоциональной окраски русскоязычных текстов, по трём классам: позитивный, нейтральный, негативный.

Для извлечения ключевых слов сопоставлены два метода тематического моделирования. Метод скрытого размещения Дирихле реализован в режиме инкрементального обучения на пакетах данных, что обеспечивает работу с корпусами, превышающими объём оперативной памяти. Альтернативный метод BERTopic строит тематическое моделирование на основе плотных контекстных векторных представлений с последующей кластеризацией. Сравнительный анализ показал, что метод скрытого размещения Дирихле формирует более разнообразные и интерпретируемые ключевые слова на уровне отдельных документов, в то время как BERTopic склонен присваивать документам одного кластера идентичные наборы ключевых слов. По совокупности факторов для финальной реализации выбран метод скрытого размещения Дирихле.

Применение конвейера к полному корпусу показало, что средняя сложность восприятия текстов составляет около 0,47 по нормализованной шкале, доминирующей тематикой являются международные новости, а в распределении тональности преобладает негативная окраска (51,3 %), что закономерно для новостного контента. В итоговое признаковое описание

каждого документа входят: 312-мерное семантическое векторное представление, вектор оценок по 18 тематическим меткам, класс тональности, множество именованных сущностей, тип и стиль контента, временна́я метка.

В разделе 3 проведено экспериментальное исследование шестикомпонентной формулы расчёта оценки релевантности, фильтра разнообразия выдачи и классификатора поведенческих сигналов. Формула имеет вид линейной взвешенной свёртки шести составляющих: тематического совпадения, семантической близости векторных представлений, совпадения именованных сущностей, тональной близости, временной свежести и близости формата подачи.

Эксперименты выполнены на корпусе из 152 332 русскоязычных публикаций из 153 каналов мессенджера. Все посты обработаны производственным конвейером, объединяющим пять моделей: формирование 312-мерных векторных представлений, классификацию тем по 18 меткам, анализ тональности, извлечение именованных сущностей и вычисление текстовых признаков. Параллельно построен граф похожих публикаций: для 152 332 постов выявлено 835 198 рёбер сходства трёх уровней. Поскольку платформа не имеет журнала реальных взаимодействий пользователей, применялась методическая триангуляция: одно утверждение проверялось пятью независимыми способами. Для оценки построено шесть синтетических пользовательских персон с различными профилями интересов.

При сравнении пяти стратегий ранжирования полная формула показала среднее сходство выдачи с профилем 0,789, что на 11 % выше случайного ранжирования (0,714). Стратегия на основе только косинусной близости достигла 0,831, однако при этом охват тем снизился до 7 из 18, тогда как полная формула обеспечивает баланс между релевантностью и разнообразием ленты. В анализе чувствительности методом поочерёдного изменения весов наиболее чувствительной оказалась составляющая близости

формата подачи, а наименьшую чувствительность показала составляющая тональной близости.

Исследование вклада составляющих показало, что отключение тематического совпадения изменяет состав первых 30 позиций выдачи на 70 % при сдвиге распределения тем в норме L1 на 0,800. Отключение семантической близости векторных представлений меняет состав на 57 %, отключение формата подачи — на 46 %. Проверка качества векторных представлений показала, что применяемая модель rubert-tiny2 уверенно отличает дубликаты от случайных пар (медианная близость 1,00 против 0,57), однако слабо разделяет публикации разных тем в 312-мерном пространстве.

Фильтр разнообразия применяет три ограничения: предельная доля одной темы (не более 0,40), предельная длина серии подряд идущих постов одной темы (не более 3) и минимальный промежуток между связанными публикациями. На профилях с узким распределением интересов фильтр существенно повышает тематическое разнообразие ленты при снижении средней релевантности менее чем на 4 %.

Независимая оценка языковой моделью в роли судьи проводилась по 720 парам «персона, публикация» по шкале от 0 до 5. Полная формула получила средний балл 2,49, стратегия косинусной близости получила 2,27, случайное ранжирование получило 0,82, ранжирование по популярности получило 0,68. По критерию Манна–Уитни превосходство полной формулы над случайным ранжированием подтверждено с уровнем значимости p (вероятность ошибочного отклонения нулевой гипотезы) = $2{,}7 \times 10^{22}$, над ранжированием по популярности с уровнем значимости $p = 6{,}2 \times 10^{2}$. Параметрический перебор по предельной доле одной темы выявил погрешность округления при нецелых произведениях параметра на длину ленты, устраняемую переходом к целочисленному вычислению.

В разделе 4 описана реализация рекомендательной системы. В рамках разработки созданы две взаимосвязанные подсистемы: рекомендательная подсистема на основе FastAPI, реализующая конвейер обработки

естественного языка, формулу расчёта оценки релевантности и постпроцессор разнообразия, а также сервис формирования персонализированной ленты на Spring, обеспечивающий доставку рекомендаций клиенту с кэшированием и резервной цепочкой обработки отказов.

Платформа реализована как совокупность подсистем, обменивающихся данными двумя способами. Синхронное взаимодействие через HTTP применяется для запрос-ответных сценариев: сервис формирования ленты обращается к рекомендательной подсистеме за идентификаторами постов, к сервису агрегации контента за фактическим содержимым публикаций. Все остальные взаимодействия являются асинхронными и осуществляются через шину сообщений Apache Kafka, что позволяет выполнять ресурсоёмкие операции без блокировки клиента.

PostgreSQL с расширением pgvector обеспечивает хранение векторных представлений постов и поиск ближайших соседей с помощью индекса на основе иерархического графа навигируемых малых миров, что исключает необходимость в отдельной специализированной базе данных. Valkey применяется как хранилище предсобранной ленты, кэш карточек контента и механизм блокировки. Apache Kafka в режиме без внешнего координатора обслуживает топики публикации нового контента, регистрации пользователя и агрегированных взаимодействий.

Подбор кандидатов реализован как двухпоточная процедура: первый поток отбирает публикации не старше 48 часов, второй выполняет приближённый поиск 500 ближайших к профилю публикаций по индексу иерархического графа. Медианная задержка стадии подбора на корпусе из 152 000 публикаций составляет около 40 мс. Сервис формирования ленты при наличии предсобранной ленты в кэше возвращает её с курсорной нумерацией, при отсутствии последовательно обращается к рекомендательной подсистеме, при недоступности переключается на выдачу для нового пользователя, а при повторном отказе использует подборку

популярных материалов. Обработка поведенческих сигналов пользователя организована через выделенный топик сообщений: каждое событие (показ, открытие, положительная или отрицательная оценка, добавление в закладки) получает численный вес, а обновление профиля выполняется методом экспоненциального скользящего среднего.

# ЗАКЛЮЧЕНИЕ

В ходе выполнения выпускной квалификационной работы проведено исследование и разработка рекомендательной подсистемы и сервиса формирования персонализированной ленты для приложения-агрегатора текстового контента из каналов мессенджера.

Систематизированы методы обработки текстовых данных, алгоритмы построения рекомендательных систем и подходы к оценке качества рекомендаций в условиях отсутствия эталонной разметки релевантности. Подготовлен корпус русскоязычных публикаций объёмом более 152 тысяч записей из 153 каналов. Реализован конвейер извлечения признаков, включающий тематическую классификацию, определение тональности, выделение именованных сущностей и построение векторных представлений.

Основным результатом экспериментальной части является подтверждение работоспособности предложенной шестикомпонентной формулы расчёта релевантности. Формула статистически значимо превосходит случайное и трендовое ранжирование. Все 26 контрактных утверждений на спецификацию системы выполнены. При параметрическом анализе выявлен и локализован дефект постпроцессора разнообразия, сформулировано его исправление.

На основе полученных результатов реализованы рекомендательная подсистема на языке Python и сервис формирования персонализированной ленты на языке Kotlin. Обе подсистемы интегрированы в общее приложение. Зафиксированы веса компонент формулы, пороговые значения постпроцессора разнообразия, схема базы данных и механизм обработки сигналов пользователя.

**Отдельные части бакалаврской работы были представлены на конференции:** Устимов М.Д. РАЗРАБОТКА СИСТЕМЫ ДЛЯ СБОРА И ДЕДУПЛИКАЦИИ КОНТЕНТА ДЛЯ СИСТЕМЫ РЕКОМЕНДАЦИЙ В ПРИЛОЖЕНИИ ДЛЯ АГРЕГАЦИИ КОНТЕНТА // «СНК СГУ-2026»

**Основные источники информации:**


1. Фальк К. Рекомендательные системы на практике / К. Фальк ; пер. с англ. Д. М. Павлова. Москва : ДМК Пресс, 2020. 448 с. ISBN 978-5-97060-774-9.
2. Кутянин А. Р. Рекомендательные системы: обзор основных постановок и результатов // Интеллектуальные системы. Теория и приложения. 2017. Т. 21, № 4. С. 18–30. URL: http://www.mathnet.ru/ista26 (дата обращения: 07.05.2026).
3. Козлова М. Г., Германчук М. С. Разработка гибридной системы рекомендаций // Таврический вестник информатики и математики. 2022. № 3. С. 30–52. URL: https://www.mathnet.ru/rus/tvim149 (дата обращения: 05.04.2026).
4. Воронцов К. В. Машинное обучение : курс лекций. Ранжирование (Learning to Rank) : [презентация] / Московский государственный университет имени М. В. Ломоносова. URL: http://www.machinelearning.ru/wiki/images/8/89/Voron-ML-Ranking-slides.pdf (дата обращения: 07.05.2026).
5. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение / пер. с англ. А. А. Слинкина. Москва : ДМК Пресс, 2018. 652 с. ISBN 978-5-97060-554-7.
6. Юферев В. И., Разин Н. А. Векторизация текстов на основе word-embedding моделей с использованием кластеризации // Моделирование и анализ информационных систем. 2021. Т. 28, № 3. С. 292–311. DOI 10.18255/1818-1015-2021-3-292-311. URL: http://www.mathnet.ru/mais751 (дата обращения: 05.04.2026).
7. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment / Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore : Association for Computational Linguistics, 2023. P. 2511–2522. URL: https://aclanthology.org/2023.emnlp-main.153/ (дата обращения: 05.04.2026).
8. Global Sensitivity Analysis: The Primer / A. Saltelli, M. Ratto, T. Andres, F.

Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola. Chichester : John Wiley & Sons, 2008. 304 p. ISBN 978-0-470-05997-5. DOI 10.1002/9780470725184.

9. Property-Based Testing in Practice / H. Goldstein, J. W. Cutler, D. Dickstein, B. C. Pierce, A. Head // Proceedings of the IEEE/ACM 46th International Conference on Software Engineering (ICSE '24). New York : ACM, 2024. Art. 90. DOI 10.1145/3597503.3639581.
10. SciPy 1.0: fundamental algorithms for scientific computing in Python / P. Virtanen, R. Gommers, T. E. Oliphant [et al.] // Nature Methods. 2020. Vol. 17, № 3. P. 261–272. DOI 10.1038/s41592-019-0686-2.