МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра математической кибернетики и компьютерных наук

АВТОМАТИЧЕСКАЯ ТЕМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ НОВОСТНОГО МАССИВА

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 451 группы
направления 09.03.04 — Программная инженерия
факультета КНиИТ
Кондрашова Даниила Владиславовича

Научный руководитель
доцент, к. ф.-м. н. _____ С. В. Папшев

Заведующий кафедрой
доцент, к. ф.-м. н. _____ С. В. Миронов

Саратов 2025

ВВЕДЕНИЕ

В настоящее время оперативный поиск информации становится критически важной задачей. Однако анализ полного массива данных невозможен из-за его масштабов, что создаёт необходимость в классификации и последующей фильтрации данных для выделения релевантной информации. Решением этой проблемы может служить тематическая классификация.

Большие объёмы данных, такие как новостные потоки, часто не имеют системной тематической разметки. Даже при наличии рубрикации, её субъективность может приводить к проблемам: некорректному присвоению тем, избыточности тематических категорий и их недостаточному охвату. Это вызывает ошибки при поиске и анализе информации. Для устранения этих недостатков требуется механизм, обеспечивающий точную тематическую классификацию с возможностью автоматической разметки новостных материалов.

Одним из инструментов для реализации такого подхода являются тематические модели в сочетании с алгоритмами глубокого обучения. Первые позволяют выявить скрытые темы в текстовых данных и подготовить разметку для обучения вторых. Алгоритмы глубокого обучения, в свою очередь, могут классифицировать новые тексты по заданным темам.

Таким образом, целью данной работы является разработка нейросетевого метода автоматической классификации новостей на основе тематической модели предметной области.

Для достижения цели необходимо решить следующие задачи:

1. Выполнить парсинг новостных данных и их текстовую предобработку;
2. Провести анализ характеристик и параметров набора данных;
3. Выполнить тематическое моделирование подготовленных данных с оптимальными параметрами;
4. Разметить данные для обучения нейронной сети-классификатора с помощью тематического моделирования;
5. Выполнить обучение нейронной сети-классификатора на размеченных данных;
6. Провести анализ качетсва обученной модели;
7. Проанализировать эффективность разработанного метода автоматической тематической классификации.

Структура и объём работы. Для решения поставленных задач выполнена

выпускная квалификационная работа, включающая в себя введение, 2 основные главы, заключение, список использованных источников из 29 наименований и 11 приложений. Работа изложена на 97 страницах.

Первая глава имеет название «Теоретические и методологические основы автоматической тематической классификации» и содержит информацию об основных подходах и инструментах используемых в автоматической тематической классификации.

Вторая глава имеет название «Практико-технологические основы автоматической тематической классификации» и содержит подробное описание процесса выполнения работы и кратткие итоги по каждому из реализованных пунктов.

Завершается выпускная квалификационная работа заключением, списком использованных источников, а также приложениями с кодом А-Л.

1 Теоретические и методологические основы автоматической тематической классификации

Сбор новостных данных. Подраздел «Сбор новостных данных» посвящён методологии формирования новостного корпуса. Проведён сравнительный анализ трёх принципиальных подходов к получению данных: ручного сбора, запросов через API и автоматизированного веб-скрапинга. Ручной метод отвергнут из-за низкой производительности, а API-интеграция — ввиду ограниченной применимости (неоперативность предоставления данных владельцами платформ). В результате обоснован выбор веб-скрапинга как оптимального метода, сочетающего высокую скорость обработки с минимальной зависимостью от внешних факторов.

Детально рассмотрены критерии отбора новостной платформы. Ключевыми требованиями выступили: единая структура HTML-документов на всём сайте, отсутствие системных блокировок HTTP-запросов и статичность контента (полная доступность информации без динамической подгрузки через JavaScript). На основе этих критериев выбран сайт НИУ ВШЭ, полностью соответствующий установленным требованиям. Особо отмечена его структурная однородность новостных материалов, что исключает необходимость ручной настройки парсера под разноформатные документы.

Подготовка собранных данных. Подраздел «Подготовка собранных данных» детализирует методологию предобработки новостного корпуса, где ключевой задачей выступает устранение шума при сохранении семантической целостности текстов. Основой обработки стала лингвистическая нормализация, требующая выбора метода приведения слов к словарной форме. После сравнительного анализа стемминга и лемматизации предпочтение отдано лемматизации как методике, гарантирующей корректное сохранение семантики терминов. Это критически важно для последующего тематического моделирования, где искажение смысла слов приводит к ошибочной интерпретации контекста.

Математические основы тематического моделирования. Подраздел «Математические основы тематического моделирования» обосновывает применение тематического моделирования как ключевого инструмента для решения задачи автоматической разметки новостных данных. Центральная гипотеза метода заключается в том, что тексты формируются из скрытых тематических распределений, а слова порождаются конкретными темами, а не документами.

Это позволяет выявлять латентные семантические структуры в больших текстовых коллекциях без ручной аннотации.

Основная практическая цель использования тематического моделирования в работе — подготовка размеченных данных для обучения нейросетевого классификатора. Традиционная ручная разметка неприменима из-за масштабов новостных потоков и субъективности человеческой оценки. Тематические модели решают эту проблему, автоматически присваивая документам вероятностные распределения по темам на основе частотных закономерностей терминов.

Критически важным аспектом является обеспечение интерпретируемости и устойчивости моделей. Для этого применена аддитивная регуляризация (ARTM), позволяющая контролировать свойства матриц распределений слов и тем. Регуляризаторы решают проблемы избыточного перекрытия тематик и повышают семантическую согласованность терминов внутри тем.

Раздел также включает анализ метрик оценки качества моделей. Рассмотрены внутренние критерии: перплексия (степень предсказуемости текста), когерентность тем (семантическая связность топовых слов) и разреженность распределений. Эти показатели позволяют объективно сравнивать эффективность различных конфигураций модели без привлечения экспертов.

Методы обработки текста с помощью нейросетей. Подраздел «Методы обработки текста с помощью нейросетей» посвящён обоснованию выбора нейросетевой архитектуры для тематической классификации. Проведён критический анализ методов преобразования текста в векторные представления. Рассмотрены классические подходы (Bag-of-Words, TF-IDF), отвергнутые из-за принципиальных ограничений: высокая размерность, игнорирование контекста и семантических связей. Обоснован приоритет семантических эмбеддингов — плотных векторных представлений слов, сохраняющих смысловую близость терминов и адаптируемых к новым лексическим единицам.

Детально исследованы архитектуры нейронных сетей для обработки последовательностей. Выявлены недостатки рекуррентных моделей (RNN/LSTM): низкая скорость из-за последовательных вычислений, проблемы с длинными зависимостями и затухание градиентов. В качестве решения предложены трансформеры, чей механизм self-attention обеспечивает:

1. параллельную обработку токенов;
2. учёт глобального контекста через анализ взаимовлияния слов;

3. эффективную работу с протяжёнными текстами.

Вывод по первому разделу: проведён комплексный анализ инструментальных средств и методологий для реализации автоматической тематической классификации новостных массивов. На системном уровне рассмотрены альтернативные подходы к сбору данных, предобработке текстов, тематическому моделированию и нейросетевой классификации, после чего обоснован выбор оптимальных решений для каждого этапа. Для ключевых компонентов системы (веб-скрапинг, лингвистическая нормализация, регуляризация тематических моделей, трансформерные архитектуры) детализированы критерии выбора, подтверждающие эффективность отобранных инструментов. Дополнительно исследованы математические принципы работы тематического моделирования как фундаментального элемента системы, с акцентом на механизмы регуляризации и оценки качества моделей. Сформированная методологическая база обеспечивает теоретическую основу для практической реализации алгоритма.

2 Практико-технологические основы автоматической тематической классификации

Получение новостного массива путём веб-скраппинга. Подраздел «Получение новостного массива путём веб-скраппинга» посвящён разработке программного компонента для сбора новостного массива, а также рассмотрению количественных характеристик собранных данных.

Реализация сбора данных выполнена на языке Python с использованием специализированных библиотек. Для отправки HTTP-запросов применена библиотека Requests, обеспечивающая эффективное получение HTML-кода страниц. Парсинг извлечённого контента осуществлён средствами BeautifulSoup4, позволяющей анализировать структуру документов через поиск по тегам и CSS-классам. Отказ от Selenium обоснован статичностью сайта-источника (НИУ ВШЭ), не требующей эмуляции браузерного взаимодействия.

Разработан алгоритм, включающий: анализ структуры новостных карточек, извлечение метаданных (заголовков, дат, анонсов), рекурсивное получение полных текстов по внутренним ссылкам. Оптимизация производительности достигнута за счёт реализации многопоточности через стандартные средства Python, что ускорило обработку архивных страниц. Результатом работы алгоритма стал структурированный набор данных в формате Excel, содержащий 17 430 документов с полными текстами и атрибутами. Количественные характеристики корпуса подтверждают репрезентативность выборки: общий объём превышает 12 миллионов токенов при средней длине документа 695 слов.

Подготовка новостного массива. Подраздел «Подготовка новостного массива» посвящен разработке программного компонента для предобработки новостного массива, а также рассмотрению экспериментальных результатов подготовки данных.

Реализация предобработки данных выполнена на Python с использованием библиотеки SpaCy для лингвистической нормализации. Выбор SpaCy обоснован её способностью к контекстному анализу, интегрированному конвейеру токенизации и лемматизации, а также высокой точности предобученных моделей для русского и английского языков. Однако выявлено ограничение: встроенные словари стоп-слов недостаточно эффективны для специализированных новостных корпусов, где частотные термины могут быть тематически значимыми (например, «котировка» в финансовых текстах).

Для решения этой проблемы разработан и реализован гибридный подход. Он дополняет стандартную фильтрацию SpaCy метрикой TF-IDF, рассчитываемой средствами библиотеки Gensim. Данная метрика автоматически идентифицирует и удаляет низкозначимые термины на основе их распределения в корпусе, обеспечивая адаптивность к тематической специфике. Дополнительно реализованы: алгоритмы очистки от нетекстовых элементов с пороговым контролем (50% неалфавитных символов), обработка мультиязычных фрагментов через разделение русско-английских сегментов, удаление документов с недостаточным содержанием (<80 токенов).

Результатом обработки стал очищенный корпус из 11 860 документов. Ключевой эффект — сокращение словаря на 93% (с 278 724 до 18 707 уникальных токенов) при сохранении семантической целостности текстов.

Построение тематической модели. Подраздел «Построение тематической модели» посвящен разработке программных компонентов для осуществления тематического моделирования, а также рассмотрению экспериментальных результатов проведённого тематического моделирования.

Библиотека BigARTM выбрана как основной инструмент тематического моделирования благодаря её поддержке аддитивной регуляризации (ARTM), позволяющей гибко контролировать свойства тем через комбинацию сглаживающих, декоррелирующих и разреживающих регуляризаторов. Для расширения функциональности разработаны два специализированных класса:
My_BigARTM_model (добавляет расчёт когерентности, визуализацию динамики метрик и упрощённый интерфейс регуляризации) и Hyperparameter_optimizer (реализует интеллектуальный подбор параметров через Optuna).

Эксперименты на 11 конфигурациях данных дали следующие результаты. С одной стороны, достигнуты высокие значения метрик: когерентность до 0.537 (TF-IDF порог 10%), перплексия до 2810 (TF-IDF 1% + декорреляция). С другой, все модели показали критически высокое перекрытие тематических распределений, что подтверждается визуальным анализом и отклонением от эталонной разметки на 84%. Гипотезы причин включают недостаточный объём данных и ограниченность гиперпараметрического поиска.

Практическим итогом этапа стало создание автоматизированного конвейера тематической разметки на основе матрицы θ. Полученный датасет с вероятностными распределениями тем по документам использован для обучения

нейросетевого классификатора, хотя его низкое качество (F1 < 0.25) указывает на фундаментальные ограничения подхода для выбранного корпуса.

Обучение модели классификатора. Подраздел «Обучение модели классификатора» посвящен разработке программных компонентов для осуществления обучения нейросетевого классификатора, а также рассмотрению экспериментальных результатов обучения и проведению их анализа.

Реализация этапа основана на предобученной модели RoBERTa, выбранной за её эффективность в анализе контекстных зависимостей. Интеграция выполнена через платформу Hugging Face, обеспечившую доступ к инструментарию обучения. Для обработки текстов применён специализированный токенизатор.

Разработан класс TopicClassifier, автоматизирующий конвейер обработки: преобразование меток, токенизацию данных, настройку параметров обучения (скорость обучения 2e-5, размер батча 32) и мониторинг метрик качества.

Эксперименты на 12 конфигурациях данных показали точность ниже ожидаемой: accuracy в диапазоне 0.166-0.291, F1-мера — 0.035-0.252. Наилучший результат достигнут без TF-IDF обработки. Контрольный тест с ручной разметкой ВШЭ подтвердил потенциал архитектуры, продемонстрировав accuracy 0.71. Это указывает, что тематическая разметка требует уточнения, а не замены классификатора. Дополнительные оптимизации (сокращение словаря, биграммы) не дали значимого улучшения.

Перспективным направлением признано совершенствование тематического моделирования: увеличение объёма данных, расширение перебора гиперпараметров и раннее применение регуляризации способны повысить точность разметки и, как следствие, качество классификации.

Вывод по второму разделу: разработан комплекс программных компонентов, реализующих полный цикл автоматической тематической классификации — от сбора данных до нейросетевого предсказания. Практическая реализация включает: создание парсера на Python с использованием библиотек Requests и BeautifulSoup4, разработку гибридного подхода к предобработке текста (SpaCy + TF-IDF фильтрация), построение расширенного решения для тематического моделирования на базе BigARTM с классами-обёртками, а также внедрение классификатора RoBERTa через Hugging Face API.

Экспериментальные результаты подтвердили работоспособность систе-

мы: успешно собран корпус из 17 430 документов, сокращение словаря на 93% при сохранении семантической целостности, автоматическая тематическая разметка данных. При этом выявлены направления для улучшения: точность классификации (0.166-0.291 accuracy) оказалась ниже ожидаемой, что связано главным образом с неточностью тематической разметки. Контрольный эксперимент с ручной разметкой (accuracy 0.71) подтвердил адекватность архитектуры классификатора и обозначил перспективу улучшения через совершенствование тематического моделирования.

Ключевой практический итог — создание модульной системы, готовой к интеграции и дальнейшей оптимизации путём увеличения объёма данных, расширенного подбора гиперпараметров и модификации регуляризационных стратегий.

ЗАКЛЮЧЕНИЕ

В ходе данной дипломной работы был разработан алгоритм автоматической классификации новостей на основе тематической модели предметной области.

Для этого было выполнено следующее:

1. Проведён анализ инструментов по сбору данных и выбраны наиболее удобные из них (BeautifulSoup4, requests);
2. Проведён сбор данных;
3. Проанализированы способы обработки текстовых данных и выбраны наиболее удобные из них;
4. Проанализированы популярные инструменты для обработки текстовых данных (NLTK, Pymorphy3, SpaCy) и выбран наиболее удобный и точный из них (SpaCy);
5. Проведена подготовка данных для тематического моделирования и проведён анализ её результатов;
6. Изучен механизм тематического моделирования с помошью аддитивной тематической регуляризации;
7. Разработаны инструменты для тематической классификации с помощью библиотеки BigARTM;
8. Проведены эксперименты по проведению тематической классификации над подготовленными различными способами данными, а также проведён анализ результатов экспериментов;
9. Рассмотрены различные способы обработки текстовых данных нейронными сетями и выбран наиболее подходящий из них (семантическое векторное представление);
10. Проведён анализ архитектур подходящих типов нейронных сетей и выбрана наиболее подходящая из них (transformer);
11. Проведён анализ доступных предобученных сетей и сервисов, которые их предоставляют, в ходе которого выбран наиболее удобный из них (Hugging Face и Roberta);
12. Проведены эксперименты по обучению тематического классификатора новостей, а также выполнен анализ результатов и сделаны соответствующие выводы.

Основной вывод по итогам работы: предложенный метод автоматической

классификации имеет перспективу применения при более тщательном тематическом моделировании исходного набора данных.