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.

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 выбранны 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.