МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»**

Кафедра Дифференциальных уравнений и математической экономики

Применение методов машинного обучения для автоматической классификации русскоязычных текстов

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 247 группы

направления 09.04.03 — Прикладная информатика

механико-математического факультета

Салтыкова Никиты Александровича

Научный руководитель
профессор, д.э.н., профессор В.А.Балаш

Заведующий кафедрой
зав. кафедрой, д.ф.-м.н.,
профессор С.И. Дудов

Саратов 2025

Введение. Современный цифровой мир характеризуется стремительным ростом объёмов текстовой информации. Новостные агрегаторы, социальные сети, корпоративные документы и научные публикации требуют эффективных методов автоматической обработки и классификации. В условиях избытка данных ручная разметка становится непрактичной, что делает методы машинного обучения и обработки естественного языка (NLP) критически важными для автоматизации этих процессов.

Особую сложность представляет классификация русскоязычных текстов из-за морфологической сложности языка, наличия синонимии и многозначности слов, а также недостатка предобученных моделей по сравнению с английским языком. Кроме того, в зависимости от задачи (тематическая категоризация, анализ тональности, определение жанра) требуются разные подходы к обработке текста и выбору алгоритмов.

В данной работе исследуются методы автоматической классификации русскоязычных новостных статей на основе датасета Lenta.ru текстов. Рассматриваются как традиционные алгоритмы машинного обучения (логистическая регрессия, K-ближайших соседей), так и современные нейросетевые архитектуры (LSTM, CNN, BERT). Особое внимание уделяется сравнению их эффективности и анализу ошибок.

Основная цель работы - Провести сравнительный анализ эффективности различных методов машинного обучения для задачи автоматической классификации русскоязычных новостных текстов на примере корпуса документов полученных с сайта Lenta.Ru.

Для достижения цели были поставлены следующие задачи:

1. Провести предобработку данных с учётом морфологии: лемматизация, удаление стоп-слов, токенизация.
2. Реализовать методы классификации.
3. Провести сравнительный анализ эффективности.

Структура работы. В данной работе содержится введение, 3 раздела и заключение:

- Во введении рассматривается актуальность темы работы и краткое описание самой работы.

- В первом разделе работы приведено описание корпуса текстов, использованного при выполнении выпускной работы, а также методов загрузки и предварительной обработки данных.
- Во втором разделе представлен систематизированный обзор методов машинного обучения, применяемых для классификации текстов и результат их выполнения.
- В третьем разделе проведён сравнительный анализ эффективности рассмотренных методов.
- В заключении подводятся итоги работы.

Основное содержание работы. В начале работы описывается набор данных, затем производится предобработка и подготовка данных для ML. Загружается датасет новостных статей из CSV-файла, оставляя только два столбца: text (текст статьи) и topic (тематика). Затем выполняется очистка текста и обработка редких тем. Датасет довольно большой(800 000+ строчек), поэтому для ускорения работы извлечена выборка объемом 10 000 документов. Далее при помощи функции clean_text производится очистка текста от HTML-тегов, знаков пунктуации, а также все слова приводятся к нижнему регистру. Темы, которые встречаются реже 100 раз удаляются, чтобы проблема распределения классов была не так ярко выражена, т.к. редкие классы могут негативно влиять на качество модели. Далее происходит подготовка данных для обучения. Текстовые названия тем (например, «Политика», «Экономика») преобразуются в числа (0, 1, 2...), так как модели работают только с числовыми данными. Затем данные делятся на обучающую и тестовую выборки 70% и 30% соответственно.

Далее рассматриваются традиционные методы машинного обучения.

Первой рассматривается модель логистической регрессии, использующая векторное представление TF-IDF. TF-IDF (Term Frequency-Inverse Document Frequency) преобразует тексты в числовые векторы, учитывая важность слов в документе относительно всей коллекции. Максимальное количество признаков ограничено 5000 для уменьшения размерности пространства признаков. Количество итераций обучения увеличено до 1000 для гарантии сходимости алгоритма.

Второй подход сочетает алгоритм CatBoost с векторными представлениями Doc2Vec. Doc2Vec создает семантические векторные представления документов фиксированной размерности (100 компонент), учитывая контекстную информацию. Обучение Doc2Vec проводится с параметрами window=3 (размер контекстного окна) и min_count=5 (игнорирование редких слов). CatBoostClassifier реализует градиентный бустинг на деревьях решений с параметрами: 300 итераций, скорость обучения 0.1, глубина деревьев 5. Используется механизм ранней остановки при отсутствии улучшения качества на протяжении 20 итераций.

Для комплексного сравнения методов классификации дополнительно реализованы:

1. Random Forest с TF-IDF представлениями (100 деревьев с балансировкой классов)
2. Метод градиентного бустинга (XGBoost + TF-IDF)
3. Метод k-ближайших соседей (k-NN) с векторными представлениями Doc2Vec (k=5)
4. Метод LogisticRegression с библиотекой FastText

Каждая модель оценивается с помощью метрик precision, recall и F1-score, вычисленных для каждого класса отдельно и усредненных по всем классам. Отчеты о классификации включают названия исходных тематических категорий благодаря сохраненному преобразованию LabelEncoder.

Использование различных комбинаций методов векторизации текста и алгоритмов классификации позволяет провести всесторонний анализ эффективности разных подходов к решению задачи тематической классификации новостных статей. Особое внимание уделено обработке несбалансированных данных через параметр class_weight.

В соответствии с рисунком 1, приведены результат работы CatBoost + Doc2Vec на тестовой выборке.

	precision	recall	f1-score	support
Интернет и СМИ	0.74	0.79	0.76	281
Мир	0.75	0.79	0.77	605
Россия	0.91	0.84	0.87	1541
Спорт	0.94	0.92	0.93	108
Экономика	0.72	0.82	0.77	464
accuracy			0.82	2999
macro avg	0.81	0.83	0.82	2999
weighted avg	0.83	0.82	0.83	2999

Рисунок 1 — Результат работы CatBoost + Doc2Vec

Далее рассматривается реализация и сравнение современных подходов к классификации текстовых данных, включая трансформерные архитектуры (BERT) и нейронные сети (LSTM, CNN).

Для решения задачи тематической классификации новостных статей была применена предобученная модель BERT (Bidirectional Encoder Representations from Transformers), а именно её компактная русскоязычная версия «rubert-tiny2». Процесс обработки данных включал токенизацию текстов с ограничением максимальной длины последовательности до 128 токенов, автоматическое дополнение (padding) и усечение (truncation) текстов для приведения к единому размеру.

Архитектура решения включает создание специализированного класса NewsDataset, наследующего от torch.utils.data.Dataset, который обеспечивает корректную работу с входными данными. Модель последовательной классификации инициализируется с количеством выходных нейронов, соответствующим числу тематических классов в датасете.

Для оценки качества модели в процессе обучения реализована функция compute_metrics, вычисляющая accuracy и macro-averaged F1-score. Обучение проводится с помощью класса Trainer из библиотеки transformers, что позволяет организовать весь процесс обучения с минимальным количеством кода.

В качестве альтернативы трансформерным архитектурам были реализованы классические нейросетевые подходы LSTM и CNN. Предварительная подготовка данных включала: создание словаря токенов (10000 наиболее ча-

стых слов), преобразование текстов в последовательности индексов, приведение последовательностей к единой длине (200 токенов) через паддинг.

Обе модели обучались в течение 5 эпох с размером батча 64. В качестве функции потерь использовалась sparse_categorical_crossentropy, что позволяет работать с метками классов в виде целых чисел.

В соответствии с рисунком 2, приведены результат работы алгоритма BERT на тестовой выборке.

	precision	recall	f1-score	support
Интернет и СМИ	0.79	0.82	0.80	281
Мир	0.88	0.78	0.83	605
Россия	0.90	0.94	0.92	1541
Спорт	0.94	0.93	0.93	108
Экономика	0.83	0.82	0.82	464
accuracy			0.88	2999
macro avg	0.87	0.86	0.86	2999
weighted avg	0.88	0.88	0.88	2999

Рисунок 2 — Результат работы BERT

Ансамблевый метод был реализован для комбинирования сильных сторон трех различных алгоритмов классификации: логистической регрессии, случайного леса и метода k-ближайших соседей (KNN). Данный подход реализует гибридный гетерогенный ансамбль, где используются принципиально разные алгоритмы: линейный, ансамблевый, метрический, а также совмещаются преимущества нескольких парадигм машинного обучения.

У модели логистической регрессии использован пайплайн с масштабированием данных, настроены параметры - количество итераций, балансировка классов, solver 'saga'. Она выбрана как модель, хорошо работающая с линейно разделимыми данными.

Случайный лес настреон на 200 деревьев с ограничением глубины (max_depth=15). Применена балансировка классов ('balanced_subsample'). Был выбран для учета нелинейных зависимостей в данных

KNN использует косинусную метрику расстояния. Учтены веса соседей по расстоянию. Применен для учета локальных закономерностей в данных.

Для построения ансамбля был использован VotingClassifier с мягким голосованием ('soft'). Назначены веса моделям: 0.4 для логистической регрессии, 0.3 для случайного леса и KNN. Включена параллельная обработка (n_jobs=-1), ансамбль был обучен на TF-IDF представлении текстов.

Данный подход компенсирует слабые стороны отдельных моделей, у него повышенная устойчивость к переобучению, а также есть возможность тонкой настройки через весовые коэффициенты. Ансамблевая модель демонстрирует высокое качество классификации, превосходя базовые решения.

В конце работы приводится сравнительный анализ эффективности методов классификации. Показатели точности классификации приведены в таблице 1 и 2.

Таблица 1 — Сравнительный анализ методов классификации

Метод	**Accuracy**	**F1 (weighted)**	**Время обучения**
Logistic Regression + TF-IDF	0.89	0.89	<1 мин
CatBoost + Doc2Vec	0.82	0.83	<1 мин
RandomForest + TF-IDF	0.80	0.80	<1 мин
XGBoost	0.85	0.85	2 мин
KNN + Doc2Vec	0.83	0.83	<1 мин
LogisticRegression + FastText	0.82	0.82	<1 мин
LSTM + FastText	0.84	0.84	3 мин
CNN	0.83	0.83	1 мин
Ансамбль (LR+RF+KNN)	0.88	0.88	2 мин
BERT	0.88	0.88	3 мин

Таблица 2 — Анализ эффективности методов по классам

Метод	**Лучший класс (F1)**	**Худший класс (F1)**
Logistic Regression + TF-IDF	Спорт (0.95)	Мир (0.84)
CatBoost + Doc2Vec	Спорт (0.93)	Интернет и СМИ (0.76)
RandomForest + TF-IDF	Спорт (0.95)	Экономика (0.73)
XGBoost	Спорт (0.93)	Экономика (0.74)
KNN + Doc2Vec	Спорт (0.95)	Мир (0.72)
LogisticRegression + FastText	Спорт (0.93)	Мир (0.74)
LSTM + FastText	Спорт (0.95)	Экономика (0.76)
CNN	Россия (0.91)	Экономика (0.74)
Ансамбль (LR+RF+KNN)	Спорт (0.96)	Интернет и СМИ (0.83)
BERT	Спорт (0.93)	Интернет и СМИ (0.80)

Метрика F1-score для лучшего и худшего классов в каждой модели. Данные получены на тестовой выборке (n=2999).

Заключение. В данной работе был проведен сравнительный анализ эффективности традиционных, современных и ансамблевого методов машинного обучения для автоматической классификации текстов. Неожиданным лидером стала Logistic Regression + TF-IDF, модель показала наивысшую точность (Accuracy=0.89, F1=0.89), превзойдя даже нейросетевые подходы (LSTM, CNN, BERT).

Ансамблевый метод и BERT разделили второе место с Accuracy = 0.88. Ансамбль требует в 2 раза меньше ресурсов, чем BERT, и обеспечивает лучшую интерпретируемость.

Категория «Спорт» легко классифицируется всеми методами (F1=0.93–0.96) благодаря уникальной лексике. А «Экономика» и «Интернет и СМИ» — самые сложные классы (F1=0.72–0.85) из-за пересекающейся терминологии.