Федеральное государственное образовательное бюджетное учреждение высшего образования  
**«Финансовый университет при Правительстве Российской Федерации»**

Кафедра искусственного интеллекта

**ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

**К КУРСОВОЙ РАБОТЕ**

на тему: **«Автоматическая категоризация текстовых документов на основе их векторного представления»**

**Выполнила**:

Конева Виктория Викторовна

студентка группы ПМ22-3

факультета ИТиАБД

**Научный руководитель**:

к.т.н. доцент

Болтачев Эльдар Филаридович

Оглавление

[ВВЕДЕНИЕ 3](#_Toc198652740)

[ОСНОВНАЯ ЧАСТЬ 4](#_Toc198652741)

[1. Теоретическая часть 4](#_Toc198652742)

[2. Методология исследования и выбор технологического стека и архитектуры решения 5](#_Toc198652743)

[3. Практическая часть 6](#_Toc198652744)

[4. Анализ полученных результатов 11](#_Toc198652745)

[ЗАКЛЮЧЕНИЕ 13](#_Toc198652746)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 14](#_Toc198652747)

# ВВЕДЕНИЕ

В современную эпоху информационного изобилия автоматическая обработка текстов приобретает всё большую актуальность. С ростом объёмов текстовой информации возрастает необходимость в интеллектуальных системах, способных быстро и точно определять тематику документов. Автоматическая категоризация текстов — одна из ключевых задач в области обработки естественного языка (NLP), находящая применение в новостной аналитике, системах фильтрации контента, интеллектуальных помощниках и многих других сферах.

Данная курсовая работа посвящена разработке системы автоматической классификации новостных статей на основе их векторного представления. В рамках проекта реализован полный цикл обработки данных: от сбора текстов с новостного ресурса до обучения модели машинного обучения и анализа полученных результатов.

**Цель настоящей работы** – разработка системы автоматической классификации текстов новостных статей на основе векторных представлений.

**Объект исследования** — текстовые данные, представленные в естественном языке.

**Предмет исследования** — методы глубокого обучения для семантической классификации текстов.

Для достижения цели были поставлены следующие **задачи**:

1. Сбор новостных данных с сайта Lenta.ru.
2. Предобработка текстовых данных.
3. Преобразование текста в векторное представление.
4. Обучение и тестирование модели классификации.
5. Оценка эффективности и анализ результатов.

Актуальность темы обусловлена тем, что в условиях стремительного роста объемов цифровой информации, особенно в сфере новостных ресурсов, возрастает необходимость в автоматических методах обработки текстов. Ежедневно публикуются тысячи статей, обзоров и заметок, что делает ручную классификацию и структурирование контента практически невозможными. Поэтому задачи автоматической категоризации текстов становятся всё более востребованными.

Особую актуальность данное направление приобретает в контексте медиааналитики. Быстрое и точное определение тематики текстов позволяет эффективно фильтровать информацию, строить ленты новостей по интересам пользователя, а также анализировать информационные тренды.

С развитием технологий обработки естественного языка (Natural Language Processing) и появлением более совершенных алгоритмов машинного обучения появилась возможность построения высокоточных моделей категоризации на основе векторных представлений текстов. Это делает задачу, рассмотренную в данной курсовой работе, не только теоретически значимой, но и практически применимой в современных информационных системах.

# ОСНОВНАЯ ЧАСТЬ

## 1. Теоретическая часть

Методы машинного обучения для классификации текстов можно условно разделить на два поколения:

* Классические методы: наивный байесовский классификатор, логистическая регрессия, SVM, случайный лес и т.д. Эти подходы предполагают предварительное преобразование текста в числовые признаки, чаще всего с помощью эмбеддингов .
* Современные методы на основе нейросетей: рекуррентные нейронные сети (RNN), сверточные нейронные сети (CNN), а также трансформеры (Transformers).

Особое внимание в работе уделяется архитектуре Transformer, предложенной Vaswani et al. в 2017 году. Эта модель положила начало новому направлению в NLP, обеспечив качественные улучшения в задачах перевода, суммаризации, генерации и классификации текста.

Модель BERT (Bidirectional Encoder Representations from Transformers) — одна из наиболее известных реализаций трансформеров. Она использует двунаправленное обучение на основе маскирования слов, позволяя модели учитывать контекст с обеих сторон от целевого слова. Это приводит к более глубокому пониманию смысла текста.

В курсовой работе рассматривается классификация текстов с использованием и классических, и современных методов.

## 2. Методология исследования и выбор технологического стека и архитектуры решения

1. Исследование начинается с этапа сбора и подготовки данных. В качестве исходных данных были выбраны новостные статьи с сайта Lenta.ru, а именно были спарсены такие категории новостей, как:
2. «Экономика»
3. «МИР»
4. «Бывший СССР»
5. «Наука и техника»
6. «Культура»
7. «Интернет и СМИ»
8. «Спорт»
9. «Из жизни»

Они представляют собой разнообразные тексты c разнообразными категориями, что делает их интересным объектом для анализа. Для обеспечения корректности анализа текст был предварительно обработан:

* Токенизация — разбиение текста на отдельные слова.
* Приведение слов к нижнему регистру.
* Удаление пунктуации и неинформативных символов.
* Очистка от стоп-слов.
* Лемматизация — приведение слов к начальной форме.

1. Следующая часть исследования содержала получение эмбеддингов. Использовались два метода:
2. Классический метод tf-idf, который вычисляет важность каждого слова в документе относительно количества его употреблений в данном документе и во всей коллекции текстов.
3. С помощью модели “XLM-RoBERTa-base”.
4. Основной частью исследования стало обучение различных моделей и их последующее сравнение. В рамках работы были использованы следующие алгоритмы:

* LogisticRegression — базовая линейная модель классификации, хорошо интерпретируемая и способная давать стабильные результаты при использовании разреженных признаковых представлений, таких как TF-IDF.
* MLP (Multilayer Perceptron) — полносвязная нейронная сеть, способная выявлять нелинейные зависимости в данных. Использование данной модели позволяло учитывать более сложные взаимодействия между признаками.
* CatBoostClassifier — градиентный бустинг на решающих деревьях, эффективно работающий с категориальными признаками и устойчивый к переобучению. Отличается высокой скоростью обучения и качеством предсказаний.
* XLM-RoBERTa-base — трансформерная модель, предобученная на многоязычных корпусах. Использовалась как в качестве эмбеддинга (предварительного представления текста), так и в виде самостоятельной модели с дообучением на задаче классификации.
* GNN (Graph Neural Network) — графовая нейронная сеть, применяемая для анализа структурированных данных в виде графов. В рамках данного исследования она позволила учитывать взаимосвязи между сущностями, представленными в текстах.

Для моделей LogisticRegression, MLP и CatBoostClassifier использовались различные векторные представления текста: классический TF-IDF и контекстуальные эмбеддинги, полученные с помощью XLM-RoBERTa-base. Это обеспечивало возможность сравнить влияние различных типов признаков на качество классификации. В свою очередь, модель XLM-RoBERTa-base применялась также как end-to-end решение, включающее как извлечение признаков, так и классификацию. Графовая модель (GNN) потребовала отдельной подготовки данных в виде графовой структуры, что позволило учитывать не только семантику, но и структурные связи в текстах.

Такой комплексный подход позволил всесторонне оценить эффективность различных архитектур и методов представления текста, выявить их сильные и слабые стороны, а также подобрать оптимальные решения для решаемой задачи.

1. Для оценки качества обученных моделей были использованы стандартные метрики классификации: **accuracy**, **F1-мера** (в вариантах macro, micro и weighted) и **матрица ошибок (confusion matrix)**. Эти метрики позволяют комплексно оценить точность предсказаний как по всем классам в целом, так и по каждому отдельному классу. Такой подход дает возможность выявить, насколько хорошо модель различает категории новостных текстов, особенно при наличии классов с разным количеством примеров.
2. Последним этапом исследования стал сравнительный анализ моделей. Результаты были визуализированы в виде графика и таблицы, что позволило наглядно сопоставить показатели каждой модели. Такой подход не только демонстрировал эффективность методов, но и позволял выявить их сильные и слабые стороны.

## Практическая часть

1. Сбор данных.

Первым и важным этапом стал сбор данных. Для проведения исследования были собраны новостные статьи с сайта Lenta.ru с помощью парсинга HTML-страниц. Использовались библиотеки requests и BeautifulSoup для автоматического извлечения текстов. Данные охватывали различные рубрики, что обеспечило тематическое и временное разнообразие корпуса. Полученные тексты были сохранены в формате CSV для последующей предобработки и анализа.

1. Подготовка данных.

Следующий этапом стала подготовка текстовых данных. Качество предобработки напрямую влияет на конечный результат. Тексты прошли несколько шагов очистки:

* Удаление ненужных признаков
* Все слова были приведены к нижнему регистру.
* Токенизация текста — разбиение на отдельные слова.
* Удаление стоп-слов (артикли, предлоги и другие служебные части речи).
* Удаление знаков пунктуации и специальных символов, которые не влияют на смысл текста.
* Лемматизация слов (приведение к начальной форме) с использованием библиотеки NLTK, что позволило уменьшить количество уникальных слов и улучшить качество моделей.

1. Получение эмбеддингов.

На данном этапе текстовые данные были преобразованы в числовую форму, пригодную для подачи на вход алгоритмам машинного обучения. В работе использовались два подхода к векторизации текста: метод TF-IDF и эмбеддинги на основе модели XLM-RoBERTa.

Для построения векторного представления применялся TF-IDF с ограничением на максимальное количество признаков(10000) и учётом как отдельных слов (униграмм), так и пар последовательных слов (биграмм). Это позволило сохранить наиболее информативные элементы текста, снизив размерность признакового пространства. Данный подход сформировал разреженные векторы фиксированной размерности, пригодные для линейных моделей и нейросетей.

Второй метод основан на использовании предобученной трансформерной модели XLM-RoBERTa-base. Эта модель предоставляет контекстно-зависимые эмбеддинги, то есть учитывает не только сами слова, но и их значение в конкретном окружении. Для каждого текста извлекался специальный [CLS]-вектор размерности 768, который служит обобщённым представлением всего документа. Такой тип эмбеддинга является плотным и более информативным по сравнению с TF-IDF, особенно в задачах, где важно учитывать семантику и контекст.

Оба подхода — TF-IDF и XLM-RoBERTa — применялись параллельно, что позволило провести сравнительный анализ их эффективности при решении задачи классификации текстов.

1. Разделение данных

Перед обучением модели классификации корпус новостных текстов был дважды разбит на обучающую и тестовую выборки. Это было сделано отдельно для каждого типа векторного представления — TF-IDF и эмбеддингов XLM-RoBERTa. Такой подход позволил независимо оценить качество классификации при использовании различных способов представления текста.

1. Обучение LogisticRegression

Для классификации текстов была использована модель логистической регрессии с параметром max\_iter=1000, что обеспечило достаточное количество итераций для сходимости алгоритма. Обучение проводилось отдельно на двух типах векторных представлений: TF-IDF и XLM-RoBERTa.

Лучшие результаты модель показала при использовании TF-IDF: достигнута высокая точность и F1-мера по большинству категорий. Эмбеддинги XLM-RoBERTa, несмотря на свою контекстную природу, дали менее стабильные показатели на данной выборке. Это может быть связано с тем, что для логистической регрессии более подходят разреженные признаки, такие как TF-IDF. На рисунке 1 можно увидеть classification\_report на эмбеддингах TF-IDF.



Рисунок 1 - classification\_report на эмбеддингах TF-IDF (LogisticRegression)

1. Обучение MLPClassifier

Следующим этапом стала классификация с использованием полносвязной нейронной сети (MLP). Модель была настроена с двумя скрытыми слоями размерностью 256 и 128 нейронов, максимальным числом итераций 300 и фиксированным случайным состоянием для воспроизводимости (random\_state=42).

Обучение проводилось как на TF-IDF-векторах, так и на эмбеддингах XLM-RoBERTa. И в этом случае результаты оказались лучше на TF-IDF: модель достигла более высокой точности и F1-меры. Это подтверждает, что для базовых архитектур, таких как MLP, традиционные разреженные представления текста остаются более эффективными, чем плотные эмбеддинги, особенно без дополнительной настройки. На рисунке 2 можно увидеть classification\_report на эмбеддингах TF-IDF.



Рисунок 2 - classification\_report на эмбеддингах TF-IDF (MLPClassifier)

1. Обучение CatBoostClassifier

Для решения задачи классификации также была использована градиентный бустинг — модель CatBoostClassifier. Настройки включали 10 000 итераций, скорость обучения 0.01, глубину деревьев 6, регуляризацию и раннюю остановку на 200 итерациях без улучшений. Обучение производилось с использованием GPU, что значительно ускорило процесс.

Модель обучалась на обоих типах эмбеддингов — TF-IDF и XLM-RoBERTa. Как и в предыдущих экспериментах, CatBoost продемонстрировал лучшие результаты при работе с TF-IDF: метрики точности и F1 оказались выше по всем основным категориям. Это ещё раз подтвердило эффективность классических текстовых признаков для табличных моделей. На рисунке 3 можно увидеть classification\_report на эмбеддингах TF-IDF.

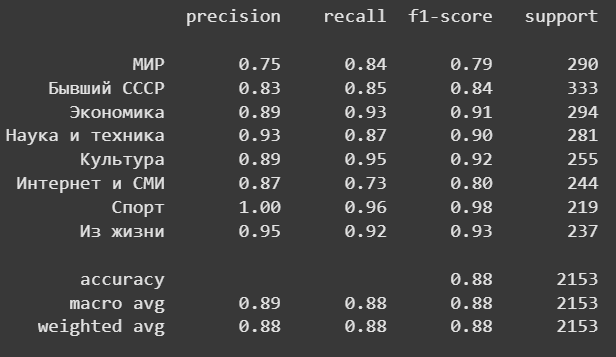


Рисунок 3 - classification\_report на эмбеддингах TF-IDF (CatBoostClassifier)

1. Обучение XLM-RoBERTa

На следующем этапе была проведена тонкая настройка предобученной трансформерной модели XLM-RoBERTa-base для задачи классификации. Модель была загружена с помощью библиотеки Hugging Face и дообучалась на размеченных новостных данных.

Для обучения был создан собственный класс Dataset, в котором тексты токенизировались с максимальной длиной 256 токенов и сопоставлялись с соответствующими метками классов. Данные подавались в модель батчами по 16 экземпляров. В качестве оптимизатора использовался AdamW с learning rate 2e-5.

Обучение проходило в течение максимум 100 эпох с включённым механизмом ранней остановки, который прерывает тренировку при отсутствии улучшений в течение 300 батчей. Модель остановилась на 4-й эпохе, что говорит о быстрой сходимости.

Графики значений функции потерь (loss) по батчам и эпохам были построены для наглядной оценки хода обучения и представлены ниже рисунок 4.

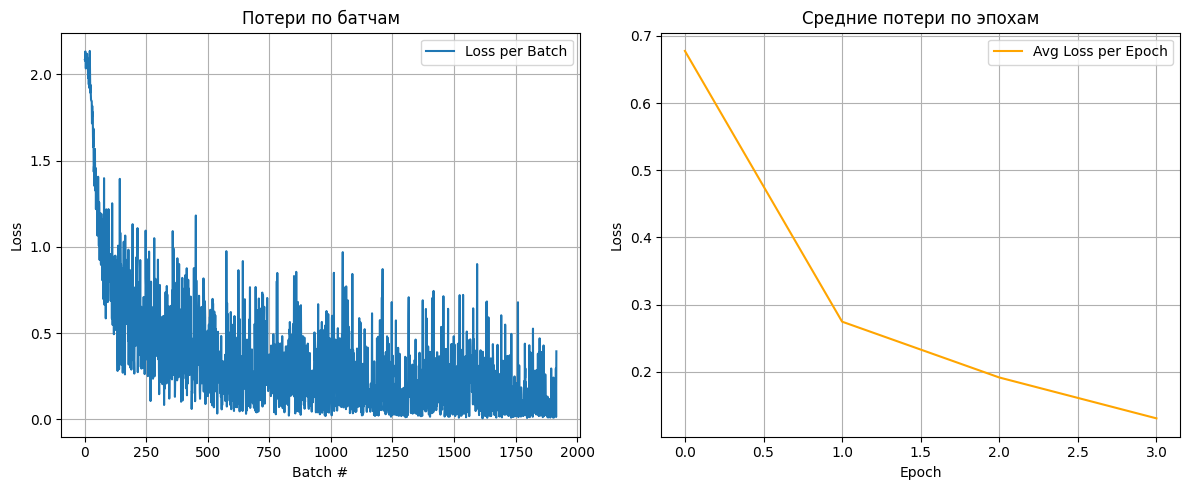


Рисунок 4 - графики значений функции потерь (loss) по батчам и эпохам

1. Обучение GNN

Для использования структурных связей между документами была построена графовая нейронная сеть на основе модели GCN (Graph Convolutional Network). Документы были представлены как узлы графа, а связи между ними формировались с использованием косинусной близости эмбеддингов: каждый узел соединялся с пятью ближайшими по смыслу.

В качестве входных признаков использовались плотные эмбеддинги размерности 768, полученные из модели XLM-RoBERTa. Архитектура сети состояла из двух сверточных слоёв с размером скрытого слоя 64 нейрона. Обучение проводилось с использованием функции потерь NLLLoss и оптимизатора Adam при скорости обучения 0.01.

Для оценки качества применялись метрики accuracy и F1-мера. Обучение шло с ранней остановкой по лучшему значению F1, и было завершено автоматически при отсутствии улучшений в течение 300 эпох. В процессе обучения отслеживалась динамика функции потерь, точности и F1, которые визуализированы на графиках ниже (Рисунок 5).

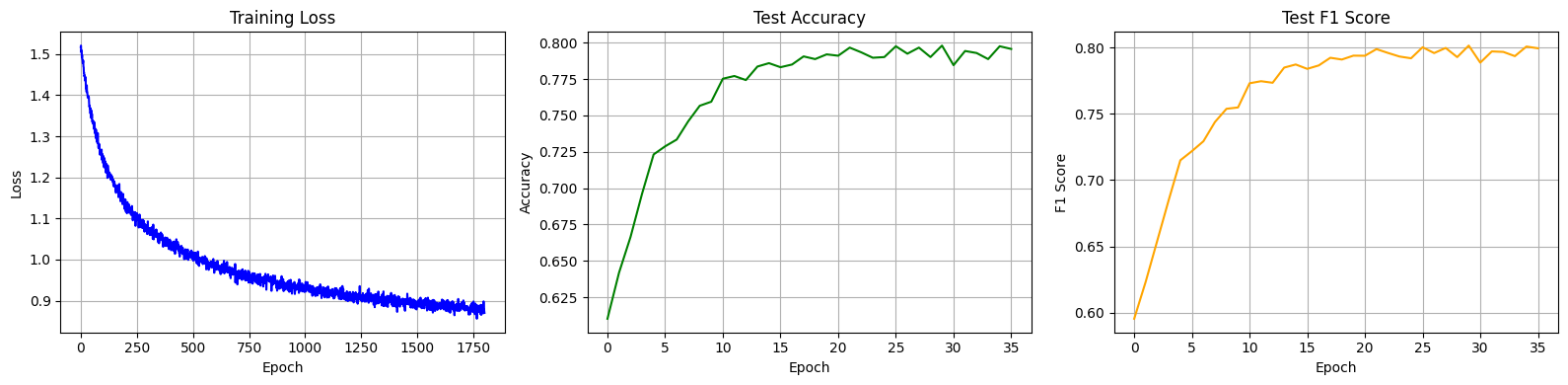


Рисунок 5 - динамика функции потерь, точности и F1

## 4. Анализ полученных результатов

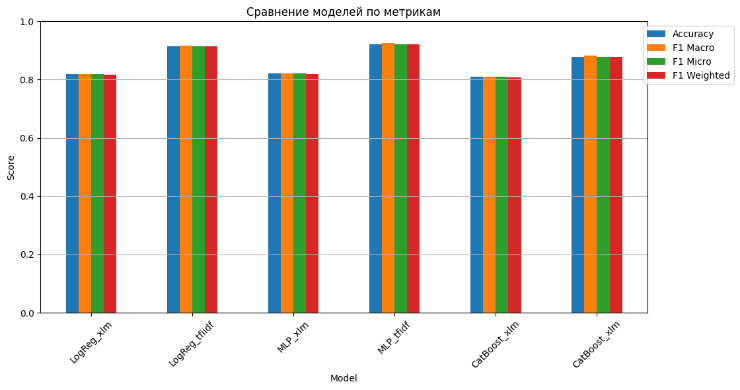
Для оценки качества классификации использовались четыре метрики: Accuracy, F1 Macro, F1 Micro и F1 Weighted. Наибольший интерес представляет F1 Macro, поскольку она учитывает баланс между классами и особенно важна при наличии классов с разным количеством примеров.

1. **Модели с TF-IDF и XLM-RoBERTa**

В таблице 1 сравниваются классические модели (логистическая регрессия, MLP, CatBoost), обученные на TF-IDF и XLM-RoBERTa:

MLP на TF-IDF показал наилучший результат среди всех табличных моделей, с F1 Macro = 0.9248.

Логистическая регрессия на TF-IDF также показала высокий результат — F1 Macro = 0.9179.

Для всех моделей результаты на TF-IDF оказались значительно выше, чем на XLM-RoBERTa-эмбеддингах, что подчёркивает эффективность разреженных признаков для базовых классификаторов.

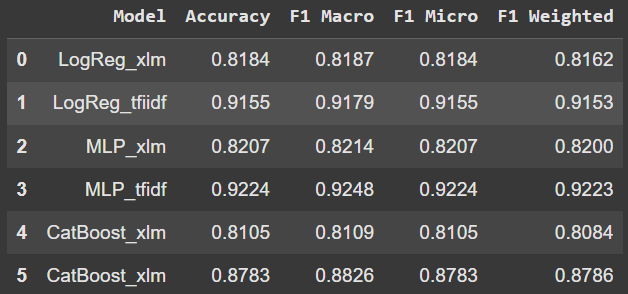
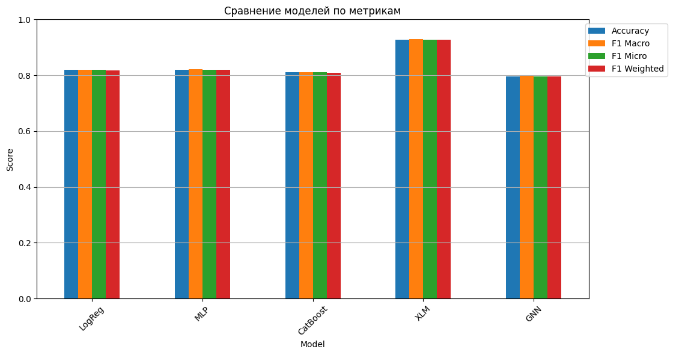


Рисунок 5 - сравнение классических моделей

1. **Модели с глубоким обучением**

На рисунке сравниваются более сложные архитектуры, а также простые модели с эмбеддингами на трансформере XLM-RoBERTa:

Трансформер XLM-RoBERTa, дообученный в режиме классификации, показал наилучшее качество среди всех моделей: F1 Macro = 0.9286.

GCN, несмотря на использование графовой структуры, уступил трансформеру и даже базовым моделям на TF-IDF, что может быть связано с ограничением структуры графа и зависимостью от качества эмбеддингов.

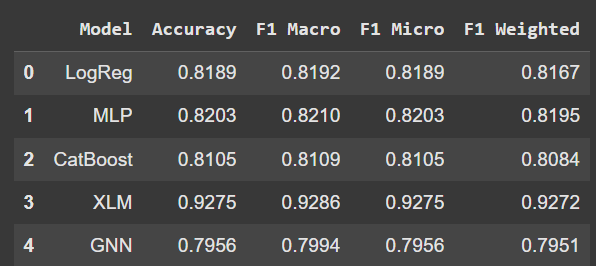


Рисунок 5 - сравнение более сложных моделей

1. **Общий вывод**

Лучший результат достигнут с использованием дообученного трансформера XLM-RoBERTa.

Среди классических моделей лидером стала MLP на основе TF-IDF.

Использование эмбеддингов XLM-RoBERTa без дообучения дало худшие результаты в сочетании с простыми моделями.

Графовая нейросеть показала умеренные результаты, но не превзошла трансформер и TF-IDF-подходы.

# ЗАКЛЮЧЕНИЕ

В данной работе была рассмотрена задача автоматической классификации новостных текстов на основе их векторного представления. Основное внимание было уделено сравнительному анализу различных методов векторизации текста и моделей классификации.

В рамках исследования тексты прошли полную предобработку и были преобразованы в два типа признакового представления — TF-IDF и эмбеддинги, полученные с использованием модели XLM-RoBERTa. На их основе были обучены и протестированы различные модели: логистическая регрессия, MLP, CatBoost, трансформер XLM-RoBERTa в режиме дообучения и графовая нейронная сеть (GCN).

Результаты показали, что наилучшее качество классификации достигается при использовании дообученного трансформера XLM-RoBERTa. Среди классических подходов лидером стала нейронная сеть MLP, обученная на TF-IDF-векторах. Простые модели на эмбеддингах XLM-RoBERTa, а также графовая сеть, показали более скромные результаты.

Проведённый анализ подтверждает, что эффективность моделей напрямую зависит от способа представления текста и архитектуры алгоритма. Универсального решения не существует — выбор подхода определяется конкретной задачей, доступными ресурсами и требованиями к точности.

Настоящее исследование демонстрирует практическую применимость современных методов обработки естественного языка и машинного обучения для анализа и классификации новостного контента.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Йылдырым С., Асгари-Ченаглу М. Осваиваем архитектуру Transformer: Разработка современных моделей с помощью передовых методов обработки естественного языка. — М.: Диалектика, 2022.
2. Parsing News from RBC & Lenta.ru / Kaggle [Электронный ресурс]. — Режим доступа: https://www.kaggle.com/code/hardtype/parsing-news-from-rbc-lenta-ru (дата обращения: 20.04.2025).
3. Извлечение признаков из текстовых данных с использованием TF-IDF / Хабр [Электронный ресурс]. - Режим доступа: https://habr.com/ru/companies/otus/articles/755772/ (дата обращения: 20.04.2025)