МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра дискретной математики и информационных технологий

РАЗРАБОТКА ПРИЛОЖЕНИЯ ДЛЯ КЛАССИФИКАЦИИ НОВОСТНЫХ СТАТЕЙ С ПОМОЩЬЮ BERT-МОДЕЛЕЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Морозова Никиты Андреевича

Научный руководитель
к. э. н., доцент _____ Г. Ю. Чернышова

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2025

ВВЕДЕНИЕ

Классификация текстов является одной из ключевых задач обработки естественного языка (Natural Language Processing, NLP) и находит применение в таких областях, как анализ тональности отзывов, автоматическая модерация контента, категоризация документов, распознавание интентов в чат-ботах и прочих. Традиционные методы, такие как алгоритмы на основе векторизации текстов и применении методов машинного обучения, часто сталкиваются с ограничениями в работе с контекстуальной неоднозначностью, синонимией и сложными семантическими связями. Это приводит к снижению качества предикативных моделей, особенно в сценариях, требующих глубокого понимания смысла текста.

С появлением сложных нейросетевых архитектур, в частности таких как трансформеры, произошел прорыв в NLP. Модели на основе BERT (Bidirectional Encoder Representations from Transformers), представленные в 2018 году, стали революционными благодаря своей способности учитывать двунаправленный контекст слов. В отличие от предыдущих подходов (например, LSTM), BERT обучается на предсказании пропущенных токенов (задача Masked Language Modeling) и анализе связей между предложениями (Next Sentence Prediction). Это позволяет модели формировать глубокие семантические представления текста, учитывая как локальные, так и глобальные зависимости.

В промышленных решениях BERT и его производные (например, RoBERTa, DistilBERT) успешно применяются для определения эмоциональной окраски текста (сентимент-анализ), автоматической категоризации новостных статей, фильтрации спама и токсичных сообщений, извлечения ключевой информации из юридических и медицинских документов.

Ключевым преимуществом BERT является возможность дообучения: модель, предобученная на больших корпусах текстов, может быть быстро дообучена на узкоспециализированных данных даже при ограниченном размере выборки. Это сокращает затраты на разметку данных и делает технологию доступной для малого бизнеса и научных проектов.

Рост объема цифровых текстовых данных и потребность в их автоматической обработке подчеркивают актуальность исследований в области применения BERT. В работе рассматриваются практические аспекты адаптации

модели для конкретных сценариев классификации, а именно для классификации текстов по 17 целям устойчивого развития. Эта задача имеет ограниченное количество обучающих выборок, что влияет на качество дообученных моделей.

Для русскоязычных текстовых документов эта задача широко не исследовалась, представляется актуальным исследовать возможность формирования обучающей выборки, обеспечивающей повышение точности классификационных моделей для русскоязычного контента в контексте целей устойчивого развития.

Целью данной бакалаврской работы является разработка приложения для классификации текстов в соответствии с целями устойчивого развития с помощью BERT-моделей. Для достижения поставленной цели требуется решить следующие задачи:

— отобрать и адаптировать наборы данных для обучения моделей;
— построить BERT-модели для классификации русскоязычных новостных сообщения по тематике, связанной с целями устойчивого развития;
— реализовать интерфейс для использования моделей.

Объектом работы является усовершенствование технологии классификации текстовых документов. Предметом работы являются методы текстового анализа, глубокого обучения для решения прикладных задач.

Теоретическая значимость заключается в доказательстве эффективности дообучения BERT-моделей для классификации русскоязычных текстов по ЦУР при ограниченных размеченных данных. Практическая значимость заключается в создании русскоязычной выборки для обучения BERT-моделей и в создании web-приложения с открытым исходным кодом для пакетной классификации новостей по ЦУР.

Бакалаврская работа имеет следующую структуру. В первом разделе рассмотрены методы решения задач классификации текстов и трансформерные модели, в частности архитектура BERT. Во втором разделе описано обучение BERT-моделей для классификации новостных сообщений. В третьем разделе представлена реализация интерфейса для классификации текстов.

Бакалаврская работа содержит 9 рисунков, 3 таблицы и 30 источников.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Анализ методов решения задачи классификации текстов» исследует эволюцию подходов в обработке естественного языка. Начальные методы, основанные на ручном подборе ключевых слов и экспертных правилах, страдали низкой адаптивностью к новым данным и не учитывали контекстную неоднозначность. С развитием машинного обучения получили распространение статистические методы, такие как Bag of Words и TF-IDF, в сочетании с алгоритмами типа SVM и наивного Байеса. Однако эти подходы игнорировали семантические связи и порядок слов [1].

Прорывом стали распределённые представления слов (Word2Vec, GloVe), кодирующие лексические единицы в векторные пространства с сохранением семантических отношений. Несмотря на это, они не могли отразить многозначность слов в разных контекстах. Дальнейшее развитие нейросетевых архитектур, включая свёрточные (CNN) и рекуррентные сети (RNN), улучшило учёт последовательностей, но столкнулось с проблемой исчезающих градиентов и однонаправленности контекста [2].

Ключевым этапом стало появление трансформеров и BERT-моделей. Их механизм самовнимания позволяет анализировать зависимости между всеми словами текста одновременно, а двунаправленное обучение (задачи Masked Language Modeling и Next Sentence Prediction) формирует глубокие контекстуальные представления. Это устранило ограничения предыдущих методов и обеспечило прорыв в задачах, требующих понимания сложных семантических связей, таких как классификация по ЦУР [3].

BERT доказал превосходство над традиционными подходами благодаря способности моделировать полный контекст, поддерживать трансферное обучение и эффективно работать с ограниченными размеченными данными. Это обосновывает его применение для классификации русскоязычных новостей в рамках целей устойчивого развития.

Второй раздел «Обучение BERT-моделей для классификации новостных сообщений» представляет собой детальное исследование методологии адаптации современных NLP-архитектур для классификации русскоязычных текстов по 17 целям устойчивого развития[4].

Анализ исследований показал, что исследования столкнулись с ограничениями, связанными с данными. В работе [5] исходный датасет официаль-

ных текстов ООН был узкотематическим и недостаточным по объему, что потребовало трудоемкого ручного расширения с помощью синтетически сгенерированных текстов (Markovify), потенциально снижая репрезентативность данных и ограничивало реалистичность языковых паттернов. В исследовании [6] ключевыми проблемами стали несбалансированность классов, полное отсутствие данных по целям 16 и 17 в используемом датасете, а также субъективность краудсорсинговых аннотаций, сохранявшая риск шума в данных даже после строгой фильтрации по уровню согласия аннотаторов. В обоих случаях качество и полнота исходных данных стали ограничениями.

Основным вызовом стал острый дефицит экспертно размеченных русскоязычных данных в данной предметной области. Для решения этой проблемы была разработана комплексная стратегия, начинающаяся с интеграции двух англоязычных корпусов: OSDG-CD [7] объемом 43 025 текстов, аннотированных международными волонтерами, и SDGi Corpus [8], содержащего 7 350 официальных документов ООН, включая добровольные национальные обзоры.

Процедура предобработки данных включала многоэтапную фильтрацию: удаление текстов короче 256 символов из-за недостаточного контекста, устранение дубликатов и технических артефактов, таких как HTML-теги и спецсимволы, а также коррекцию дисбаланса классов, где наблюдалось значительное доминирование отдельных ЦУР.

Ещё одним этапом стала разработка конвейера автоматизированного перевода с использованием Google Translate API, реализовавшего кэширование результатов и пакетную обработку данных через библиотеку deep_translator с параллелизацией средств ThreadPoolExecutor. Последующая глубокая лингвистическая очистка включала лемматизацию инструментом pymorphy2, удаление стоп-слов с применением NLTK и нормализацию, результатом чего стал корпус из 26 910 русскоязычных текстов.

Для решения задачи классификации сравнивались три архитектуры: Multilingual BERT как базовая мультиязычная модель с поддержкой 104 языков, RuBERT со специализацией на русском языке, обученная на корпусах Wikipedia и социальных медиа, XLM-RoBERTa. Каждая модель была модифицирована добавлением двухслойного классификатора, включающего промежуточный полносвязный слой на 256 нейронов с активацией ReLU и 30%

Dropout, а также выходной слой с 17 нейронами и softmax-активацией, что обеспечило адаптацию к специфике 17 классов ЦУР. Для компенсации дисбаланса классов применялась кросс-энтропия.

Процесс обучения использовал оптимизатор Adam с learning rate 2e-5 и размером батча 16, дополненный регуляризационными мерами: ранней остановкой при отсутствии улучшения F1-меры в течение 5 эпох и динамическим снижением learning rate при достижении плато качества. Эксперименты проводились на GPU NVIDIA RTX 4060Ti со средним временем эпохи 12 минут 43 секунды для объединённой выборки, при этом 20% данных были выделены для валидационной выборки.

Результаты тестирования обученных моделей на различных наборах данных в виде значений метрик предоставлены в таблице 1.

Таблица 1 – Оценка моделей, обученных на разных выборках

Модель	Accuracy	Precision	Recall	F1_score	Выборка
Multilingual	0.683	0.738	0.683	0.694	SDGi Corpus
Multilingual	0.651	0.629	0.582	0.578	OSDG
Multilingual	0.874	0.891	0.874	0.878	SDGi Corpus+OSDG
RuBERT	0.700	0.746	0.700	0.705	SDGi Corpus
RuBERT	0.680	0.726	0.680	0.685	OSDG
RuBERT	0.881	0.892	0.881	0.884	SDGi Corpus+OSDG
XLM	0.848	0.859	0.848	0.850	SDGi Corpus+OSDG

Тестирование на стратифицированной выборке из 2 550 текстов выявило превосходство RuBERT, достигшего F1-меры 0.884 при точности 0.892 и полноте 0.881, превзойдя Multilingual BERT (F1=0.878) и XLM-RoBERTa (F1=0.850).

Результаты подтверждают эффективность специализированной модели RuBERT для русскоязычных текстов, демонстрирующей превосходство над мультиязычными аналогами. Объединение датасетов OSDG-CD и SDGi Corpus с последующим переводом повысило качество классификации на 18%, что подчеркивает потенциал дообучения. Критическим условием для дальнейшего прогресса является создание открытого русскоязычного корпуса с экспертной разметкой по ЦУР.

Третий раздел «Реализация интерфейса для классификации текстов» посвящён разработке веб-приложения с помощью фреймворка Flask на плат-

форме Python для практического применения обученных моделей в контексте Целей устойчивого развития (ЦУР). На базе фреймворка Flask создан интерактивный интерфейс, обеспечивающий загрузку CSV-файлов с новостными статьями, выбор модели классификации (RuBERT, Multilingual BERT или XLM-R) и пакетную обработку данных. Система автоматически определяет текстовые столбцы в загружаемых файлах и реализует GPU-ускорение через TensorFlow 2.12.

Функционал включает возможность загрузить CSV-файл с текстом для его последующей классификации, выбор предобученной BERT-модели, и визуализацию результатов классификации по ЦУР: для каждой статьи отображается номер соответствующей цели (от 1 до 17), уровень уверенности модели в процентах. Результаты доступны для экспорта в CSV-формате, обеспечивая интеграцию с внешними аналитическими инструментами мониторинга ЦУР.

При тестировании на 699 текстов из новостных источников различных регионов России, представленых в социальной сети ВКонтакте, выявлено доминирование ЦУР 8 («Достойная работа» – 18.5%) и ЦУР 11 («Устойчивые города» – 18.6%), что помогает аналитикам идентифицировать медиатренды.

Токенизация адаптирована под русскоязычные контексты.

Приложение с открытым исходным кодом доступно на платформе GitHub, что может позволить исследователям использовать его для отслеживания прогресса по ЦУР в русскоязычном медиапространстве.

Разработанный инструмент решает практическую задачу автоматизации мониторинга ЦУР в новостных потоках, предоставляя аналитикам эффективный механизм для выявления тенденций.

ЗАКЛЮЧЕНИЕ

В рамках выполнения бакалаврской работы был разработано приложение для классификации новостных статей по тематике целей устойчивого развития с использованием BERT-моделей.

Были отобраны и адаптированы англоязычные датасеты (OSDG-CD, SDGi Corpus), переведённые на русский язык. На их основе построены модели Multilingual BERT, RuBERT и XLM-RoBERTa для классификации текстов по 17 целям устойчивого развития (ЦУР). Объединение выборок позволило увеличить точность модели для русскоязычных текстов. Наилучший результат показала модель RuBERT (F1 = 0.884), что подтверждает преимущество специализированных русскоязычных архитектур. Для практического применения разработан web-интерфейс, обеспечивающий пакетную обработку данных, выбор моделей и экспорт результатов, что делает решение удобным инструментом для аналитиков.

Ключевым результатом работы стало подтверждение гипотезы о применимости дообучения для задач классификации ЦУР, даже при ограниченном объёме русскоязычных данных. Однако выявлены ограничения:

— обучение BERT-моделей требует значительных вычислительных мощностей и эффективной параллелизации процессов, особенно при работе с большими объёмами данных;
— дефицит экспертно размеченных русскоязычных датасетов, что снижает точность моделей по сравнению с англоязычными аналогами;
— дисбаланс классов в обучающих выборках, требующий дополнительной аугментации и взвешивания меток.

Перспективными направлениями для будущих исследований являются:

— создание открытого русскоязычного корпуса текстов, аннотированных экспертами в области ЦУР;
— разработка мультимодальных моделей, учитывающих не только текст, но и структурные элементы документов (графики, таблицы);
— интеграция методов активного обучения для сокращения затрат на разметку данных.

Практическая значимость работы заключается в автоматизации анализа медиатекстов, что ускоряет мониторинг прогресса в достижении ЦУР и поддержку принятия решений в государственном и корпоративном секторах.

Реализованное решение демонстрирует потенциал BERT-архитектур для обработки русскоязычных данных и может служить основой для дальнейшего развития инструментов NLP в контексте устойчивого развития.

Основные источники информации:

1. Kotsiantis, S. B., Zaharakis, I. D., Pintelas, P. E. Machine learning: a review of classification and combining techniques //Artificial Intelligence Review. – 2006. – Т. 26. – С. 159-190.
2. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text Classification Algorithms: A Survey. Information, 10(4), 150. https://doi.org/10.3390/info10040150
3. Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). A Survey on Text Classification Algorithms: From Text to Predictions. Information, 13(2), 83. https://doi.org/10.3390/info13020083
4. Цели устойчивого развития: Повестка дня на период до 2030 года // ООН. URL: https://www.un.org/sustainabledevelopment/ru (дата обращения: 15.05.2025).
5. Guisiano, J. Automatic classification of multilabel texts related to Sustainable Development Goals (SDGs) / J. Guisiano, R. Chiky // Proceedings of the TECHENV EGC2021. – Montpellier, France, Jan. 2021. – URL: https://hal.science/hal-03154261v1 (дата обращения: 16.04.2025).
6. Angin, M. A RoBERTa Approach for Automated Processing of Sustainability Reports / M. Angin, B. Tasdemir, C. A. Yilmaz, G. Demiralp, M. Atay, P. Angin, G. Dikmener // Sustainability . – 2022. – Vol. 14, № 23. – P. 16139. – DOI: 10.3390/su142316139 .
7. OSDG Community Dataset (OSDG-CD) // Zenodo URL: https://zenodo.org/records/11441197 (дата обращения: 13.04.2025).
8. SDGi Corpus // Hugging Face URL: https://huggingface.co/datasets/UNDP/sdgi-corpus (дата обращения: 14.04.2025).