**КЛАССИФИКАЦИЯ РУССКИХ НАРОДНЫХ СКАЗОК С ИСПОЛЬЗОВАНИЕМ МОДЕЛИ нейронных сетей BERT**

**В. Соловьев1, М. Солнышкина1, А. Тен2, Н.Прокопьев3**

1 Казанский федеральный университет, Казань, Россия

2 Нобилис Тим, Казань, Россия

3 Институт прикладной семиотики АН РТ,

**Абстракт**

**Введение:** Автоматическое профилирование и классификация жанров имеют решающее значение для оценки доступности текстов для различных категорий респондентов и поэтому уже их актуальность весьма значима в образовании, веб-аналитических инструментариях, сентимент-анализе и машинном переводе. Сказки представляют собой один из наиболее сложных и ценных объектов для изучения благодаря своей неоднородности и широкому спектру неявных идиосинкразий. Однако традиционные методы классификации, включая стилометрические и параметрические алгоритмы, не только трудоемки и занимают много времени, но и непродуктивны для выявления классификационных дискриминантов. Исследования в этой области крайне немногочисленны, их результаты продолжают оставаться весьма дискуссионными.

**Цель:** Представленное исследование призвано заполнить данную исследовательскую нишу и предлагает алгоритм, позволяющий осуществить такмономию русских сказок на основе заданных параметров. Мы представляем новую модель классификации русских сказок на основе нейронной сети BERT, тестируем гипотезу о потенциале BERT для классификации текстов на русском языке и валидируем ее на репрезентативном корпусе из 743 русских сказок.

**Метод:** Предварительно обученный на коллекции из трех классов документов трансформер BERT был настроен для конкретной задачи таксономии. Алгоритм включает токенизацию, векторное представление единиц текста (embeddings) как ключевых компонентов обработки текста в BERT, оценку стандартных эталонов, используемых для обучения классификационных моделей, анализ сложных (контаминированных) случаев, возможных ошибок, поиск и применение способов, повышающих точность классификационных моделей. Оценка эффективности моделей проводится на основе функции потерь, точности предсказания.

**Результаты:** Мы оценили потенциал BERT для классификации русских текстов, способность повысить производительность и качество существующих моделей NLP. Наши эксперименты с вариантами сети cointegrated/rubert-tiny, ai forever/ruBert-base и DeepPavlov/rubert-base-cased-sentence на различных задачах классификации показывают, что наши модели не только соответствуют уровню современных алгоритмов и моделей, с наилучшей точностью 95,9% для сети cointegrated/rubert-tiny, но и она превосходит модели ai forever/ruBert-base и DeepPavlov/rubert-base-cased-sentence с большим отрывом. Таким образом, достигнутая разработанной нашей моделью точность классификации настолько высока, что может конкурировать с экспертной классификацией.

**Заключение:** Полученные результаты подчеркивают важность тонкой настройки моделей классификации. BERT демонстрирует большой потенциал для совершенствования технологий NLP и повышения качества автоматического анализа текстов, а также открывает новые возможности для исследований и применения данной модели в широком спектре областей, включая идентификацию и таксономию релевантных по содержанию текстов, способствуя принятию адекватных решений. Разработанный и проверенный алгоритм может быть масштабирован для классификации такого сложного и неоднозначного дискурса, как художественная литература, что улучшает наше понимание специфических категорий текстов. Для целей такого рода требуются значительно большие массивы данных.

**Ключевые слова:**

машинное обучение, модель Bert, сказки, классификация текста, нейронные сети

**ВВЕДЕНИЕ**

Обработка естественного языка (NLP) представляет собой важную область исследований, которая играет ключевую роль в развитии искусственного интеллекта. Анализ и генерация текста компьютерами имеют широкий спектр применений, включая поиск информации, анализ тональности текстов, машинный перевод и многое другое. Однако до недавнего времени методы обработки естественного языка не использовались для оценки контекста и сложных внутритекстовых взаимосвязей. Последнее справедливо как в отношении имплицитных, так и эксплицитных дискурсивных связей, и исследователи признают, что даже гибридные подходы, сочетающие глубокое обучение и традиционные методы, с трудом справляются с задачами, которые в значительной степени предполагают понимание способов связи между сущностями (Santoro et al., 2018).

Нейросетевые модели, особенно основанные на архитектуре Transformer (см. Gerasimenko и др., 2022), значительно улучшили результаты NLP с момента создания и развития первой модели на основе BERT. BERT, Bidirectional Encoder Representations from Transformers, представленная Google, выделяется среди других благодаря своей концептуальной простоте и эмпирической мощности. Созданные и разработанные для предварительного обучения глубоких двунаправленных представлений модели BERT настраиваются только с одним дополнительным выходным слоем также как и более современные модели (см. Devlin et al., 2018). Спектр применения BERT чрезвычайно широк, включая сентимент-анализ, выявление фальшивых новостей, системы ответов на вопросы, классификацию документов и текстов, извлечение информации и т. д. (Rasmy et al., 2021, Atagün et al., 2021, Wang et al., 2020, Jwa et al., 2019, Sun et al., 2019).

Для применения модели BERT в различных задачах осуществляется обучение базовой языковой модели на большом объеме обучающих данных для решения задачи маскированного языкового моделирования (MLM). Данная задача заключается в восстановлении пропущенных (маскированных) слов в тексте, и модель, обученная на такой задаче, учится генерации слов в тексте с учетом контекста (Fu et al., 2022).

Одна из причин, по которой BERT, как предварительно обученная языковая модель, широко используется в настоящее время, заключается в ее способности обучать контекстуализированные представления слов из больших неаннотированных корпусов и восстанавливать маскированные фрагменты (Lai et al., 2020). Успех этих моделей часто объясняется их способностью улавливать сложные синтаксические и семантические характеристики слов (Peters et al., 2018).

В настоящее время BERT рассматривается как золотой стандарт обработки текста. Модели на основе BERT заметно различаются по количеству нейронов и параметрам. Например, сointegrated/rubert-tiny − это небольшая модель с 11,8 млн. параметров, входящая в состав известной библиотеки HuggingFace's Transformers (github.com/huggingface/transformers). Описание преимуществ cointegrated/rubert-tiny перед другими 10 моделями на основе BERT осуществлено в (Bolshakov et al., 2023), где утверждается, что BERT демонстрирует хороший баланс точности и скорости расчетов при обработке предложений. Модель настоятельно рекомендуется для быстрого расчета небольших наборов данных.

Мы предполагаем, что (1) классификация пересекающихся таких классов текстов, какими являются русские сказки, является когнитивно сложной задачей и (2) ее автоматизированная классификация может быть выполнена при помощи модели BERT с ее расширенными возможностями категоризации.

До настоящего времени задачи классификации текстов выполнялись на больших наборах данных исключительно для высокоресурсных языков, таких как английский (Tangherlini & Chen, 2024), французский (см. Martin et al., 2019, Bayer et al., 2021), немецкий (Chan et al., 2020, Labusch et al., 2019, Leitner et al., 2020). Автоматизация классификации русских сказок, насколько нам известно, представляет собой исследовательскую проблему и до настоящего времени не производилась.

В связи с вышеизложенным цель данной работы – продемонстрировать потенциал BERT в задаче классификации русских народных сказок и проверить его на репрезентативном корпусе из 743 русских сказок.

**ОБЗОР ЛИТЕРАТУРЫ**

Сказки являются уникальным жанром литературы со свойственными им структурой и стилем. Исследователи отмечают, что сказки часто содержат повторяющиеся мотивы, архетипы и сюжеты, что делает их интересным объектом для автоматической классификации и анализа. Классификации сказок многочисленны и основаны на различных признаках: «ведущий конфликт», мотив, главные герои и т.д. Общепринятый индекс ATU или Aarne-Thompson-Uther Index (Aarne, 1910, Uther, 2004) делит сказки на 5 разделов: (1) Сказки о животных, (2) Обычные народные сказки, включая сказки, религиозные сказки, реалистические сказки или новеллы, сказки о глупом людоеде, великане или дьяволе, (3) Анекдоты и шутки, (4) Формульные сказки, (5) Неклассифицированные сказки. Предлагаемый Томпсоном алгоритм определения типа сказки, хотя и был опубликован в 1928 году, то есть уже после выпуска первого каталога AT в 1910 году, не содержит главного классификационного принципа. В дальнейшем сказки классифицировались по сюжетам, персонажам, мотивам и т.д., но во всех случаях каталоги содержат множество исключений, пересечений и наложений выделенных классов. Даже общепринятая классификация сказок А. Аарне после пересмотра ее Н. П. Андреевым была сокращена до трех: сказки о животных, волшебные сказки и бытовые или реалистические сказки (см. Tudorovskaya, 1961). Тем не менее, в предисловии к своему «Указателю» Н. П. Андреев отмечает, что принятая классификация имеет ряд недостатков, так как деление всегда условно и неоднозначно, а применяемые принципы деления разнообразны (Андреев, 1929).

Пытаясь преодолеть возникшие трудности, исследователи выделяют так называемое «твердое ядро» и «мягкую оболочку» жанра сказки. Если первая включает в себя «классические сказки о животных» или «волшебные сказки», то вторая состоит из сказок, которые могут быть классифицированы по-разному, исходя из одного выбранного параметра. Кроме того, повествование, то есть сюжет, может меняться от жанра к жанру, приобретая черты самых разных повествований, встречающихся на его пути. Все это говорит о том, что классификация сказок − интересный, хотя и крайне трудоемкий объект для автоматической классификации и анализа (Pompeu, 2019). Вероятно, именно поэтому классификационные исследования сказочных текстов относительно редки, хотя в последнее время их число растет (Tangherlini & Chen, 2024).

**Анализ классификации текстов**

Классификация текстов − одна из классических задач вычислительной лингвистики, имеющая важное практическое применение, к области которых относятся рекомендательные системы, классифицирующие тексты на предмет специфических интересов пользователя и т. д. (Куприянов и др., 2023, Солнышкина и др., 2024, Reusens et al., 2024). Еще в 1997 году Б. Кесслер, Г. Нумерг и Х. Шютце предложили классифицировать «жанры как совокупности структур, коррелирующих с различными поверхностными признаками, и утверждали, что распознавание жанра на основе поверхностных признаков столь же успешно, как и на основе более глубоких структурных свойств» (Kessler et al., 1997: p. 32). Samothrakis & Fasli (2015) применили методы машинного обучения для классификации художественной литературы из коллекции Project Gutenberg по шести жанрам: «научная фантастика», «ужасы», «вестерн», «фэнтези», «криминальная фантастика», «тайна». Алгоритм включал извлечение релевантной информации с помощью Natural Language Toolkit и измерение эмоционального содержания в каждом предложении с помощью Wordnet-Affect. Акцент в исследовании был сделан на анализе эмотивной лексики: авторы пришли к выводу, что наиболее отличительным признаком, дискриминирующим вышеупомянутые жанры, являются номинации эмоции страха.

Три года спустя Worsham & Kalita (2018) применили набор различных нейросетевых моделей и классификаторов для определения шести жанров: научной фантастики, приключений, исторической фантастики, любовных романов, детективов и загадок, а также вестернов. Авторы также использовали несколько стратегий для компенсации экстремальной длины документов в наборе данных и показали, что при обучении набора данных коллекции Project Gutenberg на BOW-форме XGBoost оказался «наиболее оптимизированным и был отмечен как лучший градиентный бустинг» (Worsham & Kalita, 2018: p. 1969).

В настоящее время для извлечения данных, управления и структурирования неструктурированных данных используются различные методы машинного обучения (Parida et al., 2021) и нейронные сети глубокого обучения. BERT ознаменовал собой новый уровень исследований и продемонстрировал значительное улучшение по сравнению с предыдущими моделями на различных задачах NLP, включая классификацию текстов. В 2020 году ученые признали, что наибольшую популярность для решения задач классификации получили конволюционные нейронные сети (CNN) и рекуррентные нейронные сети (RNN), по праву признанные наиболее эффективными (Батраева и др., 2020). В подробных обзорах применения нейронных сетей для решения задач классификации, опубликованных (Minaee et al., 2021) и (Reusens et al., 2024), были сделаны революционные выводы. Например, Reusens et al. (2024) утверждают, что BiLSTM является лучшим методом, который значительно превосходит все остальные методы, кроме LR TF-IDF и RoBERTa, с доверительным уровнем 95%.

Английский язык всегда рассматривался как наиболее хорошо изученный и высокоресурсный язык. Используя опыт алгоритмов, полученный в исследованиях с текстами на английском языке, исследователи, работающие с другими языками, осуществляют классификации жанров на таких малоресурсных языках, как русский, арабский (El-Halees, 2017), иврит (Devlin, et al. 2018, Liebeskind et al., 2023) и даже неалфавитные языки, например, китайский (Jin et al., 2020), корейский (Liu et al., 2022) и японский (Lippert et al., 2022). Что касается выбора коллекций текстов, то, как показывают исследования, наиболее изученными являются новости, в том числе фейковые. Диапазон классов включает в себя тематику, эмоции, полярность и даже определение сарказма. Хотя существует множество исследований других типов текстов и дискурсов, например, работа Баррос, Л., Родригес, П. и Ортигоса, А. (2013) посвящена автоматической классификации испанской поэзии Франсиско де Кеведо с использованием категоризации эмоционального содержания и настроения.

За последние несколько лет (2019-2024 гг.) произошел заметный прогресс в классификации русских текстов, который в значительной степени был обусловлен применением методов глубокого обучения и моделей на основе трансформаторов (Solovyev et al., 2023, Tomin et al., 2023). В настоящее время BERT широко применяется в многочисленных приложениях на основе русских наборов данных, таких как художественная литература (детективы, детская литература, поэзия, фэнтези и научная фантастика), академический дискурс (история, естественные науки, медицина и здоровье, культура), бизнес, новости, исследования и политический дискурс, реклама, твиты, отзывы и т. д. Коллекции текстов, инструменты и алгоритмы, используемые для экспериментов с классификацией русских текстов, сильно различаются. Например, эксперименты, проведенные А. Р. Дубовик (2017) на текстах четырех функциональных стилей − научного, художественного, делового и медиа − с помощью стилометрических методов оказались чрезвычайно успешными: показатель F1 варьировался от 0,7 в медиатекстах до 1,0 в деловых. И.А. Батраева, А.Д. Нарцев и А.С. Лезгян (2020) применили сверточные нейронные сети (CNN) к коллекции из пяти жанров, а именно, истории, детективы, детская литература, поэзия и научная фантастика, достигнув 73,12 % точности классификации для всех 5 классов. Лагутина К.В. и соавторы (2021) сообщают, что применение «ритмических паттернов» для разделения на классы научных статей, рекламы, твитов, романов, рецензий и политических статей привело к самой высокой точности (F1=98%) для художественной литературы. Два года спустя та же группа исследователей, используя аналогичный алгоритм, выполнила еще более амбициозную задачу, классифицировав романы, статьи, рецензии, посты ВКонтакте и новости OpenCorpora с еще более высокой точностью (F1=99%) (Lagutina, 2023). Более сложная задача – таксономия десяти жанров, включая фантастику, фэнтези, детективы, прозу, историю, информационные технологии, естественные науки, исторические науки, медицина и здоровье, кулинария, культура, искусство (Николаев, 2022). Наилучшая точность результатов (F1=71,11%) была получена всего после трех эпох обучения нейронной сети.

**Классификация сказок**

Благодаря доступным наборам данных и технологическому прогрессу современные исследователи решают чрезвычайно амбициозные задачи по классификации сказок. Одними из первых в этой области были (Nguyen et al., 2012, 2013), которые обучили классификационные модели, а именно SVM (2012) и метод Learning to Rank (2013), для голландских сказок. Авторы сообщили, что средний балл F1 для классификации сказок составил 0,62, и указали на высокое влияние n-грамм персонажей. Хотя реализованные модели продемонстрировали весьма скромный успех, их примеру последовали другие. В 2013 году Д.Нгуен, Д.Тришнигг, Т.Медер и М. Теун разработали классификатор сказок с использованием Learning to Rank и запросов BM25. В качестве признаков в исследовании использовались лексическое и сюжетное сходство, меры информационного поиска, а также субъектно-глагольные и объектные триплеты. Результаты показали наиболее высокий уровень средней точности взаимного ранжирования − 0,82. В том же 2013 году Ф.Карсдорп и А.Ван ден Бош опубликовали работу «Identifying Motifs in Folktales using Topic Models», в которой утверждали, что Labeled LDA и Big Document Model создают представления, которые относительно хорошо соответствуют системе классификации мотивов, построенной вручную и используемой в исследованиях народных сказок. Шесть лет спустя, в 2019 году, Д. П. Помпеу, основываясь на иерархической сети внимания (HAN), успешно оценил межъязыковой нейросетевой подход на самой большой коллекции данных − английской коллекции народных сказок. В 2022 году Р. А. Остроу сообщает об уникальной классификационной модели с общим результатом F1 = 0,77, способной классифицировать героев сказок на пропповские архетипы, отслеживая их вероятностную связь с языковым выражением и даже частеречными характеристиками. Исследователь утверждает, что схема классификации позволяет расширить классификацию сказок на типы, выделенные В. Проппом (1984). Таким образом, следует признать, что существующий уровень классификационных моделей для сказок не позволяют разработать надежную и точную таксономию, которая успешно используется в других задачах компьютерной лингвистики. Кроме того, в исследованиях, проводимых в этой области, используются различные филологические классификации сказок, но отсутствуют объединяющие их теоретические основы. Что касается русских сказок, то, насколько нам известно, они никогда не использовались для типологической или жанровой классификации с использованием методов искусственного интеллекта. Все вышесказанное открывает большие перспективы для выхода за рамки традиционных подходов к изучению сказки как жанра.

**МЕТОДЫ И ДАННЫЕ**

**Данные**

Источником набора данных послужил сайт [Nukadeti.ru](https://nukadeti.ru/skazki/) и http://www.rodon.org/other/rnsoj.htm («Народные русские сказки» из сборника А. Н. Афанасьева, Москва, издательство «Правда», 1982 г.), который предоставляет наиболее объемную коллекцию русских народных сказок. Для обучения и оценки модели были выбраны три основных типа сказок: волшебные, бытовые и сказки о животных. Эти типы сказок различаются сюжетами, темами, стилем, что делает их подходящими для задачи классификации. Общая статистика коллекции исследования приведена в таблице 1.

**Таблица 1**

*Корпус исследования*

|  |  |  |  |
| --- | --- | --- | --- |
| Сказки | Слова | Предложения | Кол-во сказок |
| Бытовые | 10 766 | 1 179 | 203 |
| О животных | 10 754 | 1 018 | 342 |
| [Волшебные](https://huggingface.co/DeepPavlov/rubert-base-cased-sentence) | 9 371 | 874 | 198 |
| Всего | 30 891 | 3 071 | 743 |

**Метод**

Метод обучения нейронной сети, реализованный в данном исследовании, является стандартным и включает в себя (1) обучение сети на классифицированных и размеченных экспертами текстах; (2) реорганизацию параметров сети в результате нескольких этапов обучения и (3) оценку точности и эффективности на проверенных наборах данных.

Обучение BERT предполагает настройку гиперпараметров: размера мини-наборов данных, количества эпох, скорости обучения и т. д. Функция потерь (Loss) рассматривается как важный параметр, который измеряет эффективность модели в предсказании целевых значений по сравнению с истинными. Функция потерь вычисляет ошибку модели и используется для обновления параметров модели в процессе обучения с помощью градиентного спуска или других алгоритмов оптимизации.

Для обучения модели данные были разбиты на два набора: обучающий (df\_train) и валидационный (df\_val) в соотношении 80/20, где 80 % данных используется для обучения, а 20 % − для валидации. Последнее позволяет проверить качество обобщающих функций модели на основе данных, которые не использовались в обучении.

Ниже мы приводим описание параметров, процедур обучения и тестирования.

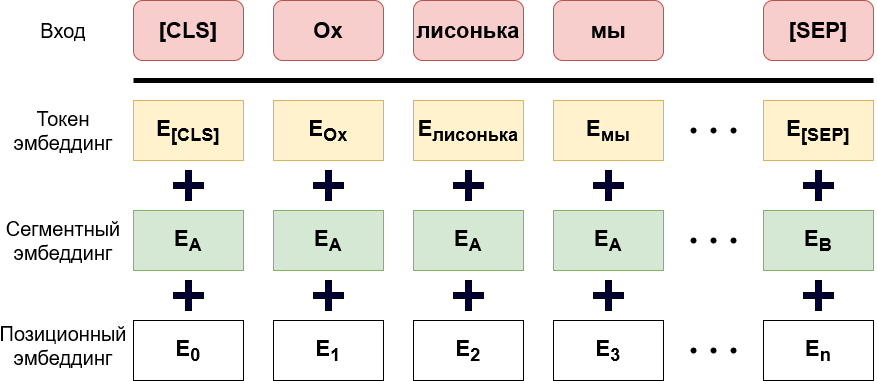
BERT обучается одновременно на двух задачах: генерации пропущенного токена и предсказания следующего предложения. На вход BERT подаются токенизированные пары предложений, с замаскированными токенами. Благодаря технике MLM, сеть обучается глубокому двунаправленному представлению языка, учитывает контекст предложения. Сама задача предсказания следующего предложения как задача бинарной классификации формулируется следующим образом: является ли второе предложение продолжением первого. Благодаря технике MLM сеть можно обучить различать наличие связи между предложениями в тексте.

Хотя BERT, вообще говоря, двунаправленный трансформер, в данной статье используется только энкодерная часть (кодировщик входа). Основная идея трансформеров заключается в применении механизма внимания, который позволяет модели взвешивать значимость различных частей входного текста для каждого обрабатываемого токена.

Общая архитектура BERT проиллюстрирована на рис. 1 с фрагментом предложения, поданного на вход.

**Рисунок 1**

*Архитектура BERT*



В архитектуре BERT используется несколько типов векторных представлений (эмбеддингов), которые преобразуют текстовые данные в числовые векторы.

1. **Векторное представление словоформ (Token Embedding)**

Каждое слово или часть слова представляется как уникальный вектор, что является стандартной практикой в современных моделях NLP. Модели на основе BERT функционируют аналогично и подразумевают следующее:

- Модель токенизации: BERT реализует токенизацию с помощью алгоритма WordPiece, который разбивает слова на части. Такой подход помогает модели эффективно работать с редкими словами и морфологическими вариантами. Например, слово «unbelievable» может быть разбито на «un», «believ» и «able», каждое из которых имеет свой собственный вектор (эмбеддинг).

- Векторы словоформ: каждое полное слово или часть слова, получает числовое представление − вектор фиксированной длины (например, 768 для BERT base и 1024 для BERT large). Этот вектор содержит семантическую информацию, помогающую модели понять значение и контекст слов.

- Характеристики: Вектор словоформы позволяет модели понимать взаимосвязи между словами, даже если они не следуют друг за другом в предложении. Это очень важно для трансформеров, работающих в автономном режиме. Вектор словоформы позволяет модели определять значение слов и их частей, выстраивая семантические связи.

Таким образом, вектор словоформ предоставляет модели информацию о значении и семантическом контексте отдельных слов в предложении.

1. **Вектор сегмента (Segment Embedding)**

При решении задач, требующих понимания двух предложений или частей текста, BERT способен различать семантику слов в разном окружении. Работа модели BERT предполагает следующее:

- Вектор сегмента: BERT обучается на задачах, требующих дифференциации контекстов двух предложений. Данная функция аналогична функции Natural Language Inference (NLI), предполагающей способность идентификации противоречий, отсутствия связей или причинно-следственных связей между двумя предложениями.

- Кодирование сегментов: каждая словоформа получает специальный вектор (эмбеддинг) сегмента, указывающий, с каким предложением она соотносится:

- Сегмент A содержит словоформы из первого предложения;

- Сегмент B включает словоформы из второго предложения (если оно есть).

- Ввод одного предложения: если текст содержит только одно предложение, все словоформы относятся к одному вектору сегмента, то есть все они принадлежат одному предложению. Это не мешает модели понимать смысл и структуру, так как векторы сегментов обеспечивают решение задач при обращении к нескольким частям текста.

Векторы сегментов позволяют модели понимать не только текстовые особенности слов, но и контекстуальные особенности в двухчастных задачах, таких как ответы на вопросы и задачи на умозаключение.

1. **Позиционный вектор (Position Embedding)**

Работа трансформаторов, включая BERT, построена на самонаблюдении, предполагающем, что модель может просматривать все словоформы одновременно, но не знает их местоположения. Для учета порядка слов к словоформам добавляются позиционные векторы.

- Отсутствие информации о порядке слов: трансформаторы не могут самостоятельно распознавать последовательность словоформ, поскольку видят все слова одновременно и не содержат информации о последовательности словоформ. Это отличает BERT от RNN, которые обрабатывают информацию последовательно с учетом порядка словоформ.

- Позиционные векторы: для того чтобы помочь модели различать позиции словоформ, каждой словоформе присваивается позиционный вектор. Каждая позиция уникальна и соответствует первой, второй и т. д. позициям словоформ в предложении. Эти векторы помогают модели понять относительное положение слов, которое необходимо для точного отражения структуры и последовательности слов в тексте.

- Математическая формула: BERT создает позиционные векторы с помощью синусоидальных функций различной амплитуды, которые позволяют модели определять позиции как на малых, так и на больших расстояниях. Каждая позиция словоформы имеет уникальный вектор, основанный на данных синусоидальных функциях.

Позиционные векторы предоставляют модели информацию о положении каждой словоформы, что важно для сохранения структуры и последовательности слов в тексте, особенно в длинных предложениях.

1. **Информация на входе**

Информация на входе модели BERT – это сумма всех трех векторов:

вектор на входе = вектор словоформы + вектор сегмента + позиционный вектор.

Каждая словоформа представлена в модели как сумма его векторного представления словоформы, позиции и сегмента, которые вместе образуют вектор фиксированной длины (обычно 768 или 1024, в зависимости от конфигурации модели). Таким образом, входной вектор для каждой словоформы содержит информацию не только о том, каким является сама словоформа (Token Embedding), но и о его позиции в предложении (Position Embedding) и о том, к какому сегменту он принадлежит (Segment Embedding). Для задачи классификации данных важно правильно подготовить датасет. Данные размещаются таким образом, чтобы каждая текстовая последовательность имела соответствующую метку класса.

Традиционная задача классификации предполагает, что каждый документ принадлежит одному или нескольким классам, то есть меткам. Иногда такую задачу называют задачей классификации с несколькими метками, а в случае с 2 классами — бинарной классификацией.

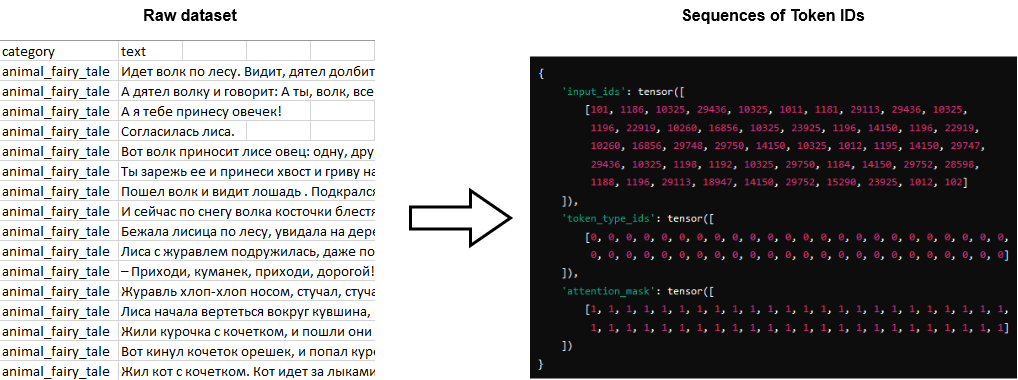
1. **Реализация обучения модели**

В этом разделе опишем и проиллюстрируем реализованную нами процедуру обучения модели.

На первом этапе осуществляется преобразование каждого предложения в идентификатор. Токенизация осуществляется при помощи библиотеки [PyTorch ([pytorch.org/get-started/locally/](https://pytorch.org/get-started/locally/))] (см. Рис. 2).

**Рисунок 2**

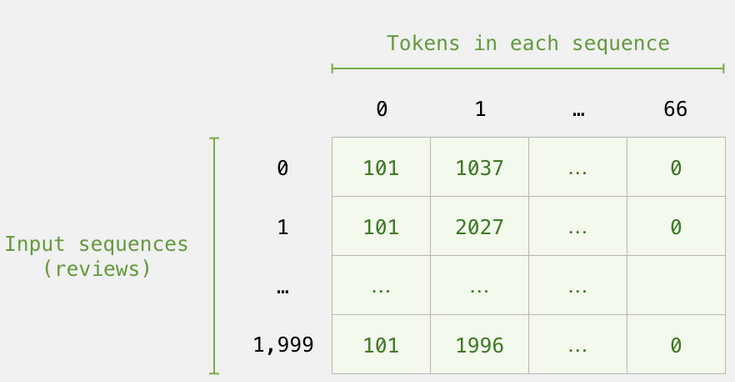
*Токенизация*



Набор данных представляет собой список (или объект Series/DataFrame из pandas – программной библиотеки на языке Python) списков. Прежде чем BERT обработает его на входе, все векторы приводятся к одному размеру путем прибавления к более коротким векторам идентификатора 0 (padding). Рис. 3 иллюстрирует форму представления предложения.

**Рисунок 3**

*Матрица/Тензор для подачи на вход нейросети*

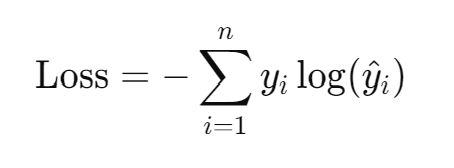


На этапе предобучения модели BERT создается специальный словарь (vocabulary), содержащий большой объем словоформ, размеченных специальным образом и снабженных уникальным идентификатором. Процесс создания словаря включает в себя разбиение слов на части с использованием алгоритма WordPiece. Это позволяет модели эффективно обрабатывать редкие и даже неизвестные слова, осуществляя их «разбиение» на части (не всегда совпадающее с морфемами). Токенизация BERT предполагает классификацию всех словоформ каждой последовательности на инициальные [CLS] и финальные [SEP].

Входные параметры для BERT:

* **Текстовые данные**: сказки на русском языке.
* **Категории сказок**: сказки о животных, бытовые сказки, волшебные сказки.
* **Токены**: текст токенизируется с помощью предобученного токенизатора BERT.
* **Идентификаторы токенов** (input\_ids): числовые представления слов в тексте.
* **Маски внимания** (attention\_mask): указывают, какие токены следует учитывать.
* **Метки категорий**: преобразованы в числовые метки для классификации.

**В представленной работе используется к**росс-энтропийная функция потерь (**Cross-Entropy Loss), ш**ироко применяемая в задачах классификации, особенно в нейронных сетях. Для многоклассовой классификации используется формула:

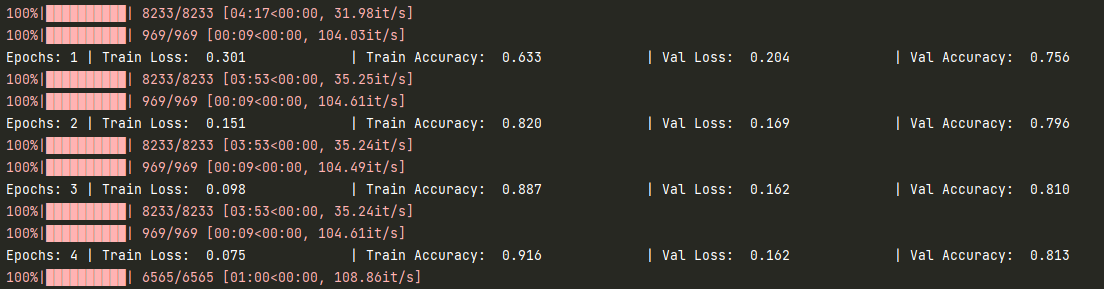


В процессе обучения модель проходит через все пакеты данных в каждой эпохе. Основные шаги включают:

1. **Передача данных в модель:** Данные из обучающего набора передаются в модель для получения предсказаний.
2. **Расчет потерь и обновление весов модели:** Потери вычисляются с использованием функции потерь (CrossEntropyLoss), затем выполняется обратное распространение ошибки и обновление весов модели с помощью оптимизатора.
3. **Печать средней потери и точности на обучающем наборе:** После каждой эпохи рассчитывается и выводится средняя потеря и точность на обучающем наборе данных для мониторинга процесса обучения.
4. **Скриншот фрагмента процесса обучения приведен на рис.** 4.

**Рисунок 4**

*Обучение модели*

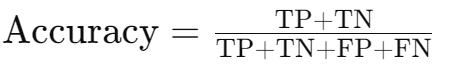


В результате выполнения всех вышеперечисленных шагов формируется эффективная система обучения модели с использованием нейросети BERT для классификации текстов сказок. Этот процесс включает в себя тщательную подготовку данных, настройку оптимизатора и планировщика, а также последовательное обучение и оценку модели для достижения высоких показателей точности и производительности.

**РЕЗУЛЬТАТЫ РАБОТЫ И ОЦЕНКА МОДЕЛИ**

Сравнение моделей проводилось по нескольким общепринятым метрикам: функция потерь (Loss), доля правильных ответов алгоритма (Accuracy), точность (Precision), полнота (Recall). В данной статье сравнивались модели rubert-tiny, ai-forever/ruBert-base, [DeepPavlov/rubert-base-cased-sentence](https://huggingface.co/DeepPavlov/rubert-base-cased-sentence) (таблицы 2-5).

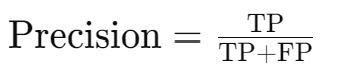
Доля правильных ответов алгоритма оценивает, насколько правильно модель классифицирует все объекты. Это отношение числа правильных предсказаний к общему числу предсказаний:



где:

* **TP** (True Positive): Верно предсказанные положительные классы.
* **TN** (True Negative): Верно предсказанные отрицательные классы.
* **FP** (False Positive): Неверно предсказанные положительные классы.
* **FN** (False Negative): Неверно предсказанные отрицательные классы.

Точность измеряет, сколько из всех предсказанных моделью положительных классов действительно являются положительными по следующей формуле:



**Таблица 2**

*Результаты для различных моделей при числе эпох 3*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Имя модели** | **Loss (функция потерь)** | **Accuracy (д**оля правильных ответов алгоритма**)** | **Precision (точность)** | **Recall**  **(полнота)** | **Число эпох** |
| cointegrated/rubert-tiny | 0.098 | 0.810 | 0.820 | 0.815 | 3 |
| ai-forever/ruBert-base | 0.114 | 0.804 | 0.810 | 0.805 | 3 |
| [DeepPavlov/rubert-base-cased-sentence](https://huggingface.co/DeepPavlov/rubert-base-cased-sentence) | 0.189 | 0.727 | 0.735 | 0.730 | 3 |

**Таблица 3**

*Результаты для различных моделей при числе эпох 4*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Имя модели** | **Loss (функция потерь)** | **Accuracy (д**оля правильных ответов алгоритма**)** | **Precision (точность)** | **Recall**  **(полнота)** | **Число эпох** |
| cointegrated/rubert-tiny | 0.075 | 0.813 | 0.810 | 0.815 | 4 |
| ai-forever/ruBert-base | 0.114 | 0.804 | 0.800 | 0.805 | 4 |
| [DeepPavlov/rubert-base-cased-sentence](https://huggingface.co/DeepPavlov/rubert-base-cased-sentence) | 0.189 | 0.727 | 0.730 | 0.725 | 4 |

**Таблица 4**

*Результаты для различных моделей при числе эпох 5*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Имя модели** | **Loss (функция потерь)** | **Accuracy (д**оля правильных ответов алгоритма**)** | **Precision (точность)** | **Recall**  **(полнота)** | **Число эпох** |
| cointegrated/rubert-tiny | 0.054 | 0.875 | 0.870 | 0.870 | 5 |
| ai-forever/ruBert-base | 0.114 | 0.804 | 0.800 | 0.805 | 5 |
| [DeepPavlov/rubert-base-cased-sentence](https://huggingface.co/DeepPavlov/rubert-base-cased-sentence) | 0.189 | 0.727 | 0.730 | 0.725 | 5 |

Из результатов видно, что оптимальным числом эпох для рассматриваемых моделей является 5 (см. таб. 2-5), так как при увеличении количества эпох до 6, доля правильных ответов на валидационном наборе начинает снижаться (см. таб. 5).

**Таблица 5**

*Сопоставление результатов обучения по эпохам 5 и 6*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Число эпох** | Train Loss | Train Accuracy | Validation Loss | Validation Accuracy |
| 5 | 0.008 | 0.989 | 0.054 | 0.875 |
| 6 | 0.008 | 0.983 | 0.054 | 0.863 |

Наилучший результат показала модель cointegrated/rubert-tiny с точностью 0.875 и минимальными потерями.

Однако модель получилась неидеальной: около 12% данных были классифицированы неверно. Выгрузим все ошибочные примеры из валидационной выборки в таблицу 6.

**Таблица 6**

*Примеры ошибочной классификации*

|  |  |  |
| --- | --- | --- |
| № | Текст | Категория |
| 729 | В старое время жил да был мужичек. У мужичка была пчела. | Сказка о животных |
| 594 | Видит, что овцы разбрелись по полю, давай их ловить да глаза выдирать. Всех переловил, всем глаза выдолбил, собрал стадо в одну кучу и сидит радехонек. | Сказка о животных |
| 497 | Да смотри, большого возу не накладывай, а вперед на меня не надейся: сегодня дай да завтра дай, а потом | Сказка о животных |
| 407 | Не спал, все барскую загадку отгадывал. Раздумает, так мало ли чего на свете не бывает, а и то в ум придет: “Может это и бывает, только я не…” | Сказка о животных |

Например, из сказки “Как Дьякона медом угощали”, модель классифицировала предложение “В старое время жил да был мужичок. У мужичка была пчела” как класс «сказка о животных», хотя согласно тестовому набору данных оно относится к классу «бытовая сказка», а предложение “Говорит им прохожий: «Вы бы, добрые молодцы, чем нукать да дёргать, слезли бы с телеги. Вот лошадь и въедет на гору!».” из сказки “Семь Агафонов бестолковых”, классификатор отнес к классу «волшебная сказка».

Мы предлагаем решить эту проблему, изменив процесс обучения, поскольку в протестированной версии модель воспринимает и обрабатывает ключевые слова и локальный, но не глобальный контекст. Дополнительной причиной может быть проблема дисбаланса классов или проблема перекрытия (контаминации) классов в исходном наборе данных. Последняя относится к случаям, когда (в нашем случае) сказки из разных классов имеют аналогичные признаки. Эта проблема считается одной из наиболее сложных: «one of the toughest problems in machine learning and data mining communities» (см. Xiong et al., 2010: p. 491). В ситуациях, когда классификация текстов затруднена, рекомендуется увеличить размер входных данных (например, для класса реалистичных сказок) и переобучить модель. Предлагаемый минимальный размер − один абзац.

Как было отмечено выше, мы выбрали 5 эпох, но тензорная структура входных данных по бытовым сказкам стала в 4 - 5 раз больше. Результаты для лучшей cointegrated/rubert-tiny модели при увеличении размера входных данных до одного абзаца значительно улучшили Accuracy, Precision и Recall (см. табл. 7).

**Таблица 7**

*Результаты rubert-tiny при размере данных – 1 абзац и при числе эпох 5*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Имя модели** | **Loss (функция потерь)** | **Accuracy (д**оля правильных ответов алгоритма**)** | **Precision (точность)** | **Recall**  **(полнота)** | **Число эпох** |
| cointegrated/rubert-tiny | 0.034 | 0.959 | 0.915 | 0.920 | 5 |

Разумеется, классификация не на 100% точная. В Таблице 8 в качестве примера для 4 сказок приведены вероятности отнесения их к тому или иному классу.

**Таблица 8**

*Вероятности принадлежности произведений к 3 классам*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Название жанра | Произведение | Вероятность принадлежности произведения к каждому классу | | |
| Бытовые | О животных | Волшебные |
| Бытовые | “Каша из топора” А. Афанасьев | 0.9281 | 0.0349 | 0.0368 |
| Бытовые | “Солдатская шинель” С. Сапцов | 0.8563 | 0.0174 | 0.1260 |
| О животных | “Ворона и рак” К. Ушинский | 0.0913 | 0.7042 | 0.2045 |
| Волшебные | “Гуси-Лебеди” А. Толстой | 0.0667 | 0.0932 | 0.8398 |

**Обсуждение**

Классификация как один из основных научных применяемых повсеместно методов требует особой осторожности, если в качестве материала используются произведения искусства. Это обусловлено, прежде всего, их природой − способностью отражать весь мир и заключать в себе мириады идей. Последнее делает произведения искусства трудно классифицируемыми.

Цель данного исследования − продемонстрировать потенциал нейронных сетей последнего поколения для решения вышеупомянутой проблемы. А русские сказки представляют собой модельную задачу, поскольку полученную классификацию легко сверить с существующим индексированным каталогом, составленным профессиональными филологами. Исследование показало, что классификационная модель на основе BERT демонстрирует высокую точность классификации сказок по трем основным категориям. Ниже мы приводим наше мнение о полученных результатах и перспективах исследования.

Мы достигли значительно более высокой точности: полученная в исследовании точность в 95 % на 13% выше, чем соответствующий параметр в работе Нгуен Д., Тришнигг Д., Медер Т. и Теун М. (2012, 2013). Хотя сам по себе этот факт не является прорывом, он свидетельствует об устойчивости и конкурентоспособности нашего алгоритма. Классификации других типов текстов, как указывалось, производились с различной степенью успеха, зависящим в первую очередь, от классифицируемых групп. Например, К.В.Лагутина, Н.С.Лагутина, Е.И.Бойчук (2021), осуществляя таксономию столь различных классов текстов, какими являются научные статьи, объявления, твиты, романы, рецензии и политические статьи, достигли 98% точности. В связи с этим классификация сказок как одного жанра на подклассы является гораздо более сложной задачей, и 98 % точности на данный момент рассматривается как недостижимая и невыполнимая. Помимо основного результата − достаточно высокого процента точности классификации, достигнутого нейросетью на русских сказках, − мы получили ряд вспомогательных результатов, потенциально полезных для дальнейших исследований. А именно: (а) при сравнении трех модификаций BERT наилучшие результаты показал вариант cointegrated/rubert-tiny; (б) оптимальное количество эпох обучения оказалось равным 5; (в) увеличение объема входных данных до одного абзаца, приводит к повышению точности. Все вышеперечисленное можно рассматривать как обязательные условия разработанного алгоритма.

Обсуждая неудачи экспериментальной классификации текстов, исследователи указывают на ряд причин. Первая из них обычно связана либо с недостаточной репрезентативностью, недостаточностью или несбалансированностью обучающей коллекции, категорий или подкатегорий исследуемых текстов (Pompeu, 2019). Аналогичным образом, в нашем случае производительность модели имеет тенденцию к увеличению количества образцов для каждой категории в коллекции, что предполагает, что результаты могут улучшиться при увеличении объема обучающих данных. Еще одной причиной неточностей при классификации сказок следует признать упомянутую выше «проблему перекрытия классов» (Xiong et al., 2010), когда составляющие подклассов внутри класса обладают очень похожими характеристиками. Последнее справедливо и в отношении сказок, «так как трудно определить, какой из признаков [в сказке] является главным, то задача сводится к отнесению одной и той же сказки к двум или нескольким классам (группам)» (Андреев, 1929: 7). То, что нам удалось сделать в данном исследовании, – это постановка новой проблемы и получение нового уровня точности классификации.

Перспективы предложенного классификатора сказок следующие:

1. Планируется продолжить исследование на материале коллекции сказок мира. Сказки представляют большой интерес для лингвистов, историков, культурологов и антропологов, поскольку открывают перспективы открытий в области когнитивных наук.

2. Поскольку в современной исследовательской парадигме сказки рассматриваются как жанр, манифестирующий и транслирующий культурные ценности и как таковой способный ориентироваться на различные аудитории, планируется реализовать разработанные алгоритмы и создать профайлер сказок с функцией классификации сказок по целевым возрастным и культурным группам. Подобный профайлер сказок даст возможность проводить стилометрический и многофакторный анализ сказок для конкретных возрастных групп, что позволит находить и открывать сходства и различия в культурах народов и способах хранения информации разными этносами.

3. Поскольку существует ряд жанров, манифестирующих схожие со сказками черты: басни, мифы, фэнтези и т.д., эксперименты с нейронной сетью, обученной на трех вышеперечисленных типах сказок, представляют для авторов особый исследовательский интерес.

**Ограничения**

Стандартным ограничением при использовании нейронных сетей является набор данных, а точнее, его количество и качество. Относительно небольшая коллекция сказок, использованная в представленном исследовании, вероятно, повлияла на точность классификации. Другая проблема − неоднозначность параметров классификации, принимаемых (или игнорируемых) экспертами, но вызывающих фундаментальные вопросы: какая из предложенных классификаций является «правильной» (если таковая имеется), какую из них следует использовать для обучения нейронной сети и можно ли любую классификацию нейронной сети назвать «правильной». Полученные нами результаты не являются абсолютными, хотя и положительны по отношению к выбранной классификации.

**ЗАКЛЮЧЕНИЕ**

Представленное исследование подчеркивает значительную целесообразность автоматической классификации сказок и подтверждает необходимость дальнейшего изучения модели классификации на основе BERT. BERT представляет собой значительное достижение в области обработки естественного языка благодаря своей способности обеспечивать глубокий анализ и контекст процесса. Исследование высветило и подчеркнуло значительную классификационную способность и эффективность BERT при разработке таксономии русских народных сказок. Предварительно обученный на репрезентативном корпусе и настроенный под конкретные задачи BERT способен с высокой степенью точности классифицировать тексты, выявляя тонкие взаимосвязи и контекстуальные особенности, характерные для русских народных сказок. В частности, такие модели, как cointegrated/rubert-tiny, ai-forever/ruBert-base и DeepPavlov/rubert-base-cased-sentence, продемонстрировали высокий уровень точности, причем наилучшая точность составила 95,9% для модели cointegrated/rubert-tiny.

Классификационные возможности BERT открывают широкие перспективы для дальнейших исследований и приложений, однако, несмотря на достигнутый прогресс, остаются открытые вопросы и направления для будущих исследований, включая улучшение качества токенизации и векторов, а также адаптацию модели к различным языкам и специфическим задачам. В целом BERT демонстрирует огромный потенциал для совершенствования технологий NLP и создания более сложных и интеллектуальных систем NLP. Это мощный инструмент, который может значительно улучшить качество автоматизированного анализа текстов и предложить новые возможности для исследований и применения в самых разных областях.

Проблемы классификации сказок обусловлены множеством факторов, включая тематическое сходство объектов классификации, разнородность их составляющих и отсутствие общепринятой жанровой классификации. Еще два момента: это нечеткие границы сказки как понятия и ее способность включаться в более крупные жанры, например, такие произведения как «Мастер и Маргарита», «Понедельник начинается в субботу», «Властелин колец». Дальнейшие исследования с использованием все более мощных систем искусственного интеллекта могут привести к лучшему пониманию и концептуализации художественной литературы. Наши выводы свидетельствуют как о проблемах, так и о перспективах в этой области.

**Благодарность**

Исследование выполнено при поддержке гранта РНФ 24-28-01355 «Жанрово-дискурсивные характеристики текста как функция лексического диапазона».

**Декларация конфликта интересов**

Отсутствует.

**Вклад авторов**

Валерий Соловьев: концептуальный анализ; исследование; методология; администрирование проекта.

Марина Солнышкина: формальный анализ; написание – оригинал; привлечение финансирования.

Андрей Тен: ресурсы; программное обеспечение; визуализация.

Николай Прокопьев: написание – корректура и редактирование.

**Литература**

Aarne, A. (1910). *Verzeichnis der Märchentypen.* [List of Fairy Tale Types]. Folklore Fellows' Communications. 3 Helsinki: [Suomalaisen Tiedeakatemian Toimituksia,](https://de.wikisource.org/wiki/Verzeichnis_der_M%C3%A4rchentypen) (in German).

[Андреев Н. П.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B4%D1%80%D0%B5%D0%B5%D0%B2,_%D0%9D%D0%B8%D0%BA%D0%BE%D0%BB%D0%B0%D0%B9_%D0%9F%D0%B5%D1%82%D1%80%D0%BE%D0%B2%D0%B8%D1%87_(%D0%BB%D0%B8%D1%82%D0%B5%D1%80%D0%B0%D1%82%D1%83%D1%80%D0%BE%D0%B2%D0%B5%D0%B4)) (1929). [Указатель сказочных сюжетов по системе Аарне](https://commons.wikimedia.org/wiki/File:%D0%90%D0%BD%D0%B4%D1%80%D0%B5%D0%B5%D0%B2_%D0%9D.%D0%9F._-_%D0%A3%D0%BA%D0%B0%D0%B7%D0%B0%D1%82%D0%B5%D0%BB%D1%8C_%D1%81%D0%BA%D0%B0%D0%B7%D0%BE%D1%87%D0%BD%D1%8B%D1%85_%D1%81%D1%8E%D0%B6%D0%B5%D1%82%D0%BE%D0%B2_%D0%BF%D0%BE_%D1%81%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B5_%D0%90%D0%B0%D1%80%D0%BD%D0%B5.pdf). Л.: Изд. Госуд. Русского Геогр. Общества.

Atagün, E., Hartoka, B. & Albayrak A. (2021). Topic Modeling Using LDA and BERT Techniques: Teknofest Example. In *6th International Conference on Computer Science and Engineering (UBMK).* 660–664*.* doi.org/10.1109/UBMK52708.2021.9558988

Barros, L., Rodriguez, P., & Ortigosa, A. (2013). Automatic Classification of Literature Pieces by Emotion Detection: A Study on Quevedo's Poetry. In *Humaine Association Conference on Affective Computing and Intelligent Interaction,* 141–146*.*

Батраева, И. А., Нарцев, А. Д., Лезгян, А. С. (2020). Использование анализа семантической близости слов при решении задачи определения жанровой принадлежности текстов методами глубокого обучения. *Вестник Томского государственного университета. Управление, вычислительная техника и информатика*, 50, 14-22. URL: http://vital.lib.tsu.ru/vital/access/manager/Repository/vtls:000709081

Bayer, M., Kaufhold, M.-A., & Reuter, Ch. (2021). A survey on data augmentation for text classification. *arXiv preprint.* arXiv:2107.03158

Bolshakov, V., Kolobov, R., Borisov, E., Mikhaylovskiy, N., & Mukhtarova, G. (2023). Scaled Down Lean BERT-like Language Models for Anaphora Resolution and Beyond. In *Компьютерная лингвистика и интеллектуальные технологии,* 1032-1040.

Chan, B., Schweter, S., & Möller, T. (2020). [German’s Next Language Model](https://aclanthology.org/2020.coling-main.598). In *Proceedings of the 28th International Conference on Computational Linguistics*, 6788–6796, Barcelona (Online): International Committee on Computational Linguistics.

**Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.**  *arXiv preprint.* arXiv:1810.04805.

Дубовик, A.Р. (2017). Автоматическое определение стилистической принадлежности текстов по их статистическим параметрам. *Компьютерная лингвистика и вычислительные онтологии*, 1*,* 29–45.

Fu, Z., Zhou W., Xu J., Zhou H., & Li L. (2022). [Contextual Representation Learning beyond Masked Language Modeling](https://aclanthology.org/2022.acl-long.193). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 2701-2714. Dublin: Association for Computational Linguistics.

El-Halees, A. M. (2017). Arabic Text Genre Classification. In *Journal of Engineering Research and Technology 4*(3), 105–109.

Gerasimenko, N.A., Chernyavsky, A.S. & Nikiforova, M.A. (2022) ruSciBERT: A Transformer Language Model for Obtaining Semantic Embeddings of Scientific Texts in Russian. *Dokl. Math. 106* (Suppl. 1), 95–96. doi.org/10.1134/S1064562422060072

Jin, Q., Xue, X., Peng, W., Cai, W., Zhang, Y., Zhang, L. (2020). Tblc-rattention: A deep neural network model for recognizing the emotional tendency of Chinese medical comment. *IEEE Access 8*, 96811–96828

Jwa, H. D. Oh, K. Park, J. M. Kang, & H. Lim (2019). exBAKE: Automatic fake news detection model based on Bidirectional Encoder Representations from Transformers (BERT), *Applied Sciences* (Switzerland), *9*(19), 4062. 10.3390/app9194062

Karsdorp, F. & Bosch, Van den A. (2013). Identifying Motifs in Folktales using Topic Models. In *Proceedings of BENELEARN 2013*, 41–49. Neijmegen: Radboud University.

Kelodjoue, E., Goulian, J., & Schwab Didier (2022). [Performance of two French BERT models for French language on verbatim transcripts and online posts](https://aclanthology.org/2022.icnlsp-1.10). In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022),* 88–94. Trento: Association for Computational Linguistics.

Kessler B., Numberg G. & Schütze H. (1997). Automatic detection of text genre. In *Proceedings of the eighth conference on European chapters of the Association for Computational Linguistics,* 32–38.

Куприянов Р. В., Солнышкина М. И., Лехницкая П. А. (2023). Параметрическая таксономия учебных текстов. *Вестник Волгоградского государственного университета*, Серия 2, Языкознание, 22(6), 80–94. DOI: https://doi.org/10.15688/jvolsu2.2023.6.6

Labusch, K., Kulturbesitz, P., Neudecker, C., & Zellhofer, D. (2019). Bert for named entity recognition in contemporary and historical German. *In Proceedings of the 15th conference on natural language processing,* 9–11.

Lagutina, K. V., Lagutina, N. S., & Boychuk, E. I. (2021). Text classification by genre based on rhythm features. *Modeling and analysis of information systems*, *28*(3), 280–291.

Lagutina, K. V. (2023). Genre Classification of Russian Texts Based on Modern Embeddings and Rhythm. *Automatic Control and Computer Sciences, 57*(7), 817–827.

Lai, Y. A., Lalwani, G. & Zhang, Y. (2020). [Context Analysis for Pre-trained Masked Language Models](https://aclanthology.org/2020.findings-emnlp.338). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3789–3804, Online. Association for Computational Linguistics.

Liebeskind, Ch., Liebeskind, Sh., & Bouhnik, D. (2023) Machine Translation for Historical Research: A Case Study of Aramaic-Ancient Hebrew Translations. *Journal on Computing and Cultural Heritage, 17*(2), 1–23. [doi.org/10.1145/3627168](https://doi.org/10.1145/3627168)

Leitner, E., Rehm, G., & Moreno-Schneider, J. (2020). A dataset of German legal documents for named entity recognition. *arXiv preprint*. arXiv:2003.13016

Lippert, Ch., Junger, A., Golam R., Md., Mohammad Ya., Hasan Sh., Md & Chowdhury, Md. (2022). Kuzushiji (Japanese Text) Classification. Technical Report. · 10.13140/RG.2.2.22416.07680

Liu, C., Zhao, Y., Cui X. & Zhao, Y. (2022)A comparative research of different granularities in Korean text classification.In *IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, Dalian, 486–489. doi.org/10.1109/AEECA55500.2022.9919047

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., Villemonte de La Clergerie, É., Seddah, D., & Sagot, B. (2019). Camembert: a tasty French language model. *arXiv preprint.* arXiv:1911.03894

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning-based text classification: a comprehensive review. *ACM computing surveys (CSUR)* 54, 1–40.

Николаев П.Л. (2022) Классификация книг по жанрам на основе текстовых описаний посредством глубокого обучения. *International Journal of Open Information Technologies,* *10*(1), 36–40.

Nguyen, D., Trieschnigg, D., Meder, Th., & Theune, M. (2012). Automatic classification of folk narrative genres. In *Proceedings of the Workshop on Language Technology for Historical Text,* Vienna. 378–382. www.oegai.at/konvens2012/proceedings/56\_nguyen12w/

Nguyen, D., Trieschnigg, D., Meder, Th., & Theune, M. (2013) Folktale classification using learning to rank. In *Proceedings of the European Conference on Information Retrieval*.  Lecture Notes in Computer Science, 7814, 195–206. Springer, Berlin, Heidelberg. doi.org/10.1007/978-3-642-36973-5\_17

Ostrow, R. A., (2022). Heroes, Villains, and the In-Between: A Natural Language Processing Approach to Fairy Tales. *Senior Projects Spring,* 275.

Parida, U., Nayak, M., Nayak, A.K., (2021) News text categorization using random forest and naive bayes. In *1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON),* Bhubaneswar,1–4. doi.org/ 10.1109/ODICON50556.2021.9428925

Peters, M., E., Neumann, M., Iyyer, M., Gardner, M., Clark, Ch., Lee, K. & Zettlemoyer, L. (2018). Deep contextualized word representations. *ArXiv*, abs/1802.05365.

Pompeu, D. P. (2019). Interpretable Deep Learning Methods for Classifying Folktales According to the Aarne-Thompson-Uther Scheme. *Master’s Thesis,* Instituto Superior Técnico.

Пропп, В.Я. (1984). Русская сказка. Л.: Изд-во ЛГУ.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. (2021) Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. In *NPJ Digit Med, 4*(1)*, 86.* doi.org/10.1038/s41746-021-00455-y

Reusens, M., Stevens, A., Tonglet, J., De Smedt, J., Verbeke, W., Vanden Broucke, S., & Baesens, B. (2024). Evaluating text classification: A benchmark study. *Expert Systems with Applications, 254.* 124302. 10.1016/j.eswa.2024.124302

Sabharwal, N. & Agrawal, A. (2021). BERT Model Applications: Question Answering System in *Hands-on Question Answering Systems with BERT,* Springer*,* 97–137. doi.org/10.1007/978-1-4842-6664-9

Samothrakis, В. S., & Fasli, M. (2015). Emotional sentence annotation helps predict fiction genre. *PloS one, 10*(11), e0141922

Santoro, A. & Faulkner, R. & Raposo, D. & Rae, J. & Chrzanowski, M. & Weber, Th. & Wierstra, D. & Vinyals, O. & Pascanu, R. & Lillicrap, T. (2018). Relational recurrent neural networks. *arXiv.* doi.org/10.48550/arXiv.1806.01822

Солнышкина, М. И., Куприянов, Р. В., Шоева, Г. Н. (2024). Лингвистическое профилирование текста: приключенческая повесть vs. Учебник. *Научный результат. Вопросы теоретической и прикладной лингвистики,* 10(1), 115-132. DOI: 10.18413/2313-8912-2024-10-1-0-7

Solovyev, V., Solnyshkina, M., & Tutubalina, E. (2023). Topic Modeling for Text Structure Assessment: The case of Russian Academic Texts. In *Journal of Language and Education*, 9(3), 143-158. doi.org/10.17323/jle.2023.16604

Sun, F., Liu, J., Wu, J., Pei, Ch., Lin, X., Ou, W. & Jiang P. (2019). BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19).* Association for Computing Machinery, New York, NY, USA, 1441–1450. doi.org/10.1145/3357384.3357895

Tangherlini, T. & Chen, R. (2024). Travels with BERT: Surfacing the intertextuality in Hans Christian Andersen's travel writing and fairy tales through the network lens of large language model‐based topic modeling. *Orbis Litterarum 79*(6), 519–562. doi.org/10.1111/oli.12458

Tianqi, Ch. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16*, 785–794, New York, NY, USA. ACM. [doi.org/10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)

Tomin, E., Solnyshkina, M., Gafiyatova, E. & Galiakhmetova, A. (2023). Automatic Text Classification as Relevance Measure for Russian School Physics Texts In 2023 *IEEE 16th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*, Singapore, 366-370. doi.org/[10.1109/MCSoC60832.2023.00061](http://dx.doi.org/10.1109/MCSoC60832.2023.00061)

Tudorovskaya, E.A. (1961). On classification of Russian folk fairy tales. Specifics of Russian folklore genres. In *Abstracts of reports. Gorky*, *55-64.*

[Uther, H.-J.](https://en.m.wikipedia.org/wiki/Hans-J%C3%B6rg_Uther) (2004). The types of international folktales: A classification and bibliography, based on the system of Antti Aarne and Stith Thompson. *Folklore Fellows' Communications,* 3 volumes, *284–286.* Helsinki, FI: Suomalainen Tiedeakatemia.

[Thompson, S.](https://en.m.wikipedia.org/wiki/Stith_Thompson) (1928). The Types of the Folk-Tale: A classification and bibliography. Antti Aarne's Verzeichnis der Märchentypen, translated and enlarged. *Folklore Fellows' Communications, 74.* Helsinki, FI: Suomalainen Tiedeakatemia.

[Thompson, S.](https://en.m.wikipedia.org/wiki/Stith_Thompson) (1977). [*The Folktale*](https://books.google.com/books?id=WKN44RtM_loC). Berkeley, CA: University of California Press.

Wang, Z., Wu, H. Liu, H.& Cai, Q.-H. (2020). BertPair-Networks for Sentiment Classification, in *2020 International Conference on Machine Learning and Cybernetics (ICMLC),* 273–278. doi.org/10.1109/ICMLC51923.2020.9469534

Worsham, В, J., & Kalita, J. (2018). Genre identification and the compositional effect of genre in literature. In *Proceedings of the 27th international conference on computational linguistics,* 1963–1973.

Xiong, H. & Wu, J. & Liu, L. (2010). Classification with ClassOverlapping: A Systematic Study.  In *1st International Conference on E-Business Intelligence (ICEBI 2010),* 303–309. Atlantis Press. doi.org/10.2991/icebi.2010.43

Источники данных

Народные русские сказки из сборника А. Н. Афанасьева, Москва, издательство «Правда», 1982 г.

Русские народные сказки, Москва, «Художественная литература», 1965.

Nukadeti.ru

www.rodon.org/other/rnsoj.htm