https://doi.org/10.17323/jle.2024.24030

КЛАССИФИКАЦИЯ РУССКИХ НАРОДНЫХ СКАЗОК С ИСПОЛЬЗОВАНИЕМ МОДЕЛИ НЕЙРОННЫХ СЕТЕЙ BERT

Соловьев Валерий Дмитриевич [1], **Солнышкина Марина Ивановна** [1], **Тен Андрей** [2], **Прокопьев Николай Аркадиевич** [3]

[1] Казанский федеральный университет, Казань, Россия

[2] Нобилис Тим, Казань, Россия

[3] Институт прикладной семиотики АН РТ

АННОТАЦИЯ

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: Предложить алгоритм, позволяющий осуществить таксономию русских сказок на основе заданных параметров.

Методология: 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. 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: Протестированные модели не только соответствуют уровню современных алгоритмов и моделей, но и превосходят их. Наилучшая точность, достигнутая для сети cointegrated/rubert-tiny, составляет 95,9 %, что значительно превышает результаты моделей ai forever/ruBert-base и DeepPavlov/rubert-base-cased-sentence. Таким образом, точность классификации, обеспеченная нашей моделью, 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. Кроме того, открывает новые возможности для исследований и применения данной модели в различных областях, 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

Для цитирования: Соловьев, В.Д., Солнышкина, М.И., Тен, А., & Прокопьев, Н.А. (2024). Модель классификации на основе BERT: пример применения к русским сказкам. *Journal of Language and Education, 10*(4), 100-114. https://doi.org/10.17323/jle.2024.24030

Correspondence:
Соловьев Валерий Дмитриевич,
maki.solovyev@mail.ru

Получена: 21 ноября 2024

Принята: 16 декабря 2024

Опубликована: 30 декабря 2024

ВВЕДЕНИЕ

Обработка естественного языка (NLP) представляет собой важную область исследований, которая играет ключевую роль в развитии искусственного интеллекта. Анализ и генерация текста компьютерами находят применение в различных областях, включая поиск информации, анализ тональности текстов, машинный перевод и др. 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, исследователи признают, что при использовании гибридных подходов, сочетающих глубокое обучение и традиционные методы, возникают трудности с задачами, которые в значительной степени предполагают понимание способов связи между сущностями (Santoro et al., 2018).

Нейросетевые модели, особенно основанные на архитектуре Transformer (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 (Devlin et al., 2018). Область применения BERT чрезвычайно широка и включает 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 к решению различных задач необходимо обучение базовой языковой модели на большом объеме данных, предназначенных для выполнения задачи маскированного языкового моделирования (MLM). Эта задача заключается в восстановлении пропущенных (маскированных) слов в тексте, 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 её 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, cointegrated/rubert-tiny — это небольшая модель с 11,8 млн параметров, входящая в состав известной библиотеки HuggingFace's Transformers (github.com/huggingface/transformers). В работе Bolshakov et al. (2023) приведено описание преимуществ cointegrated/rubert-tiny перед другими 10 моделями на основе BERT и 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 её 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 (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.

Цель исследования: продемонстрировать потенциал BERT для задачи 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 пять разделов: (1) сказки о животных; (2) обычные народные сказки, включая волшебные сказки, религиозные сказки, реалистические сказки или новеллы, сказки о глупом людоеде, великане или дьяволе; (3) анекдоты и шутки; (4) формульные сказки; (5) неклассифицированные сказки. Несмотря на то что предложенный Томпсоном алгоритм определения типа сказки был опубликован в 1928 г., то есть уже после выпуска первого каталога AT в 1910 г., он 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 её Н. П. Андреевым была сокращена до трех: сказки о животных, волшебные сказки и бытовые или реалистические сказки (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, в том числе в области рекомендательных систем, классифицирующих тексты в соответствии со специфическими интересами пользователей и т. д. (Куприянов и др., 2023; Солнышкина и др., 2024; Reusens et al., 2024). Еще в 1997 г. В своей работе Kessler et al. (1997, p. 32) 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». 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, позволяющим различать вышеупомянутые жанры, являются номинации, связанные с эмоцией 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. Кроме того, авторы 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 неупорядоченных 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 г. 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 текстов на английском языке, ученые, работающие с другими языками, проводят классификации жанров на таких малоресурсных языках, как русский, арабский (El-Halees, 2017), иврит (Devlin et al. 2018; Liebeskind et al., 2023), а также на неалфавитных языках, например, китайском (Jin et al., 2020), корейском (Liu et al., 2022) и японском (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 Barros et al. (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, в 2017 г. А. Р. Дубовик провела эксперименты на текстах четырех функциональных стилей: научного, художественного, официально-делового и публицистического, используя стилометрические методы. Результаты оказались чрезвычайно успешными: показатель F1 варьировался от 0,7 в медиатекстах, до 1,0 в текстах официального-делового стиля. И. А. Батраева, А. Д. Нарцев и А. С. Лезгян (2020) применили свёрточные нейронные сети (CNN) для анализа коллекции из пяти жанров: историй, детективов, детской литературы, поэзии и научной фантастики. В результате была достигнута точность классификации 73,12 % для всех пяти 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 %) в категории 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 — это таксономия таких жанров, как фантастика, фэнтези, детективы, проза, история, 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), они обучили классификационные модели, а именно SVM (2012) и метод Learning to Rank (2013), для голландских сказок. Авторы сообщили о среднем балле F1 0,62 для классификации сказок и указали на высокое влияние n-грамм персонажей. Несмотря на то что реализованные модели продемонстрировали весьма скромный успех, их примеру последовали другие. В 2013 г. Д. Нгуен, Д. Тришнигг, Т. Медер и М. Теун разработали классификатор сказок, используя Learning to Rank и запросы BM25. В качестве признаков в исследовании выступали 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 г. 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 достаточно хорошо соответствуют системе классификации мотивов, построенной вручную и используемой в исследованиях народных сказок. Шесть лет спустя, в 2019 г., 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 г. 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 с частеречными характеристиками. Исследователь утверждает, что схема классификации обеспечивает более широкую классификацию сказок по типам, определенным В. Проппом (1984). Таким образом, следует признать, что существующий уровень классификационных моделей для сказок не позволяет разработать надежную и точную таксономию, аналогичную достигнутой для других задач в области 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.

МЕТОДОЛОГИЯ

Данные

Источником набора данных послужил сайт Nukadeti.ru и http://www.rodon.org/other/rnsoj.htm («Народные русские сказки» из сборника А. Н. Афанасьева, Москва, издательство «Правда», 1982 г.), который представляет 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.

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, включая генерацию пропущенного токена и предсказание 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. Техника MLM позволяет сети обучаться различать наличие связи между предложениями в тексте.

Несмотря на то что BERT является двунаправленным трансформером, 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 Рисунке 1 с фрагментом предложения, поданного на вход.

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

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).

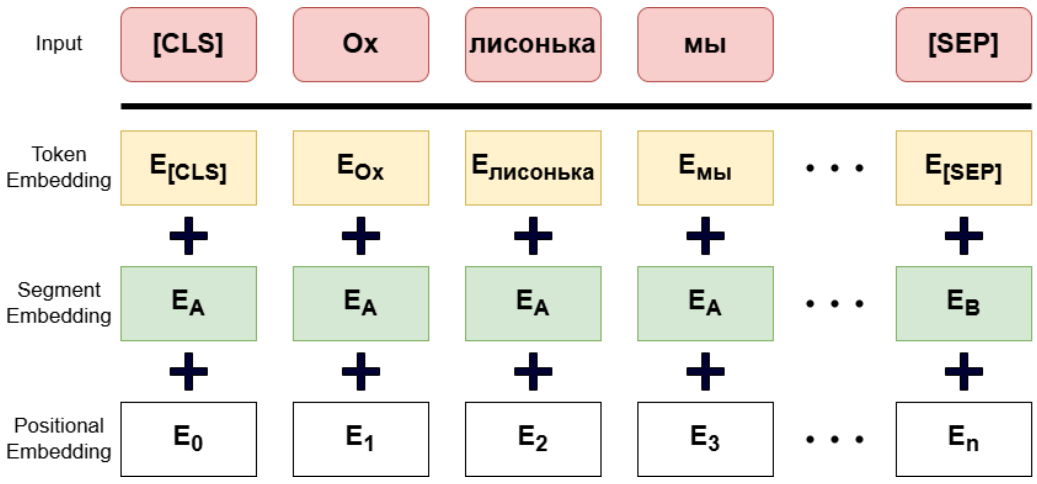
Таблица 1

Корпус исследования

Сказки	Слова	Предложения	Кол-во сказок
Бытовые	10 766	1 179	203
О животных	10 754	1 018	342
Волшебные	9 371	874	198
Всего	30 891	3 071	743

Рисунок 1

Архитектура BERT



- 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 помогает 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.

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, который указывает, 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.

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.

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), а также 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 двумя классами — бинарной классификацией.

Реализация обучения модели

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

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

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, разбивая их 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 кросс-энтропийная функция потерь (Cross-Entropy Loss), широко применяемая в задачах классификации, особенно в нейронных сетях. Для многоклассовой классификации используется формула:

$$\text{Loss} = -\sum_{i=1}^{n} y_i \log(\hat{y}_i)$$

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

- Передача данных в модель. Данные из обучающего набора передаются в модель для получения предсказаний.
- 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.
- 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.
- Скриншот фрагмента процесса обучения приведен на Рисунке 4.

В результате выполнения всех вышеперечисленных шагов формируется эффективная система обучения модели с использованием нейросети BERT для классификации текстов

Рисунок 2

Токенизация

Raw dataset

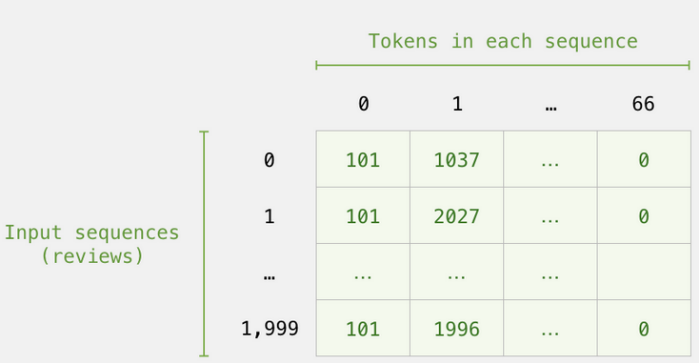
category	text
animal_fairy_tale	Идет волк по лесу. Видит, дятел долбит
animal_fairy_tale	А дятел волку и говорит: А ты, волк, все
animal_fairy_tale	А я тебе принесу овечек!
animal_fairy_tale	Согласилась лиса.
animal_fairy_tale	Вот волк приносит лисе овец: одну, дру
animal_fairy_tale	Ты зарежь ее и принеси хвост и гриву на
animal_fairy_tale	Пошел волк и видит лошадь . Подкрался
animal_fairy_tale	И сейчас по снегу волка косточки блестя
animal_fairy_tale	Бежала лисица по лесу, увидала на дере
animal_fairy_tale	Лиса с журавлем подружилась, даже по
animal_fairy_tale	– Приходи, куманек, приходи, дорогой!
animal_fairy_tale	Журавль хлоп-хлоп носом, стучал, стуча
animal_fairy_tale	Лиса начала вертеться вокруг кувшина,
animal_fairy_tale	Жили курочка с кочетком, и пошли они
animal_fairy_tale	Вот кинул кочеток орешек, и попал кур
animal_fairy_tale	Жил кот с кочетком. Кот идет за лыками

Sequences of Token IDs

```
{
    'input_ids': tensor([
        [101, 1186, 10325, 29436, 10325, 1011, 1181, 29113, 29436, 10325,
         1196, 22919, 10260, 16856, 10325, 23925, 1196, 14150, 1196, 22919,
         10260, 16856, 29748, 29750, 14150, 10325, 1012, 1195, 14150, 29747,
         29436, 10325, 1198, 1192, 10325, 29750, 1184, 14150, 29752, 28598,
         1188, 1196, 29113, 18947, 14150, 29752, 15290, 23925, 1012, 102]
    ]),
    'token_type_ids': tensor([
        [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
    ]),
    'attention_mask': tensor([
        [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
         1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
    ])
}
```

Рисунок 3

Матрица/Тензор для подачи на вход нейросети



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

Сравнение моделей проводилось по нескольким общепринятым метрикам: функции потерь (Loss), доле правильных ответов алгоритма (Accuracy), точности (Precision), полноте (Recall). В данной статье сравнивались модели rubert-tiny, ai-forever/ruBert-base, DeepPavlov/rubert-base-cased-sentence (Таблицы 2–5).

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

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}},$$

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:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}.$$

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

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.

Рисунок 4

Обучение модели

```
100%|██████████| 8233/8233 [04:17<00:00, 31.98it/s]
100%|██████████| 969/969 [00:09<00:00, 104.03it/s]
Epochs: 1 | Train Loss:  0.301            | Train Accuracy:  0.633           | Val Loss:  0.204           | Val Accuracy:  0.756
100%|██████████| 8233/8233 [03:53<00:00, 35.25it/s]
100%|██████████| 969/969 [00:09<00:00, 104.61it/s]
Epochs: 2 | Train Loss:  0.151            | Train Accuracy:  0.820           | Val Loss:  0.169           | Val Accuracy:  0.796
100%|██████████| 8233/8233 [03:53<00:00, 35.24it/s]
100%|██████████| 969/969 [00:09<00:00, 104.49it/s]
Epochs: 3 | Train Loss:  0.098            | Train Accuracy:  0.887           | Val Loss:  0.162           | Val Accuracy:  0.810
100%|██████████| 8233/8233 [03:53<00:00, 35.24it/s]
100%|██████████| 969/969 [00:09<00:00, 104.61it/s]
Epochs: 4 | Train Loss:  0.075            | Train Accuracy:  0.916           | Val Loss:  0.162           | Val Accuracy:  0.813
100%|██████████| 6565/6565 [01:00<00:00, 108.86it/s]
```

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 рассматривается как «одна из самых сложных проблем в области машинного обучения и интеллектуального анализа данных» (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 пять 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 (Таблица 7).

Разумеется, классификация не на 100 % точная. В Таблице 8 в качестве примера для четырех сказок представлена веро-

Таблица 2

Результаты для различных моделей при числе эпох 3

Имя модели	Loss (функция потерь)	Accuracy (доля 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	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 (доля 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	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 (доля 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	0.189	0.727	0.730	0.725	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

ятность того, что сказки будут отнесены к тому или иному классу.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Классификация как один из основных, широко используемых научных 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 классификацию произведений искусства сложной задачей. 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 — 95 %, что на 13 % выше, чем соответствующий параметр в работе Nguyen et al. (2012, 2013). Сам по себе этот факт не является прорывом, но он свидетельствует об устойчивости и конкурентоспособности нашего алгоритма. Классификации других типов текстов, как уже упоминалось, давали различные результаты, которые зависели, в первую очередь, от классифицируемых групп. Например, Лагутинаи соавт. (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 значительно более сложной задачей, и достижение 98 % точности на данный момент считается практически невозможным. 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.

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

Таблица 7

Результаты rubert-tiny при размере данных — 1 абзац при числе эпох 5

Имя модели	Loss (функция потерь)	Accuracy (доля правильных ответов алгоритма)	Precision (точность)	Recall (полнота)	Число эпох
cointegrated/rubert-tiny	0,034	0,959	0,915	0,920	5

Таблица 8

Вероятности принадлежности произведений к трем 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

Обсуждая неудачи экспериментальной классификации текстов, исследователи указывают на ряд причин. Первая из них обычно связана либо с недостаточной репрезентативностью, нехваткой или дисбалансом обучающей коллекции, категорий или подкатегорий исследуемых текстов (Pompeu, 2019). Таким образом, в нашем случае производительность модели имеет тенденцию к увеличению количества образцов для каждой категории в коллекции, что предполагает улучшение результатов 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, p. 7). То, что нам удалось сделать в данном исследовании, — это постановка новой проблемы и достижение нового уровня точности классификации.

Перспективы предложенного классификатора сказок:

(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, а также 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 выявить сходства и различия в культурах народов и способах хранения информации разными этносами.
(3.) Поскольку существует ряд жанров, обладающих схожими со сказками чертами, таких как басни, мифы, фэнтези и т. д., эксперименты с нейронной сетью, обученной на трех вышеперечисленных типах сказок, представляют для авторов особый исследовательский интерес.

Ограничения исследования

Стандартным ограничением при использовании нейронных сетей является набор данных, а точнее его объем и качество. Относительно небольшая коллекция сказок, использованная в представленном исследовании, вероятно, повлияла на точность классификации. Еще одной проблемой является 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, тем не менее они положительны по отношению к выбранной классификации.

ЗАКЛЮЧЕНИЕ

Наше 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 выявило 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, обнаруживая 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. Кроме того, существуют нечеткие границы сказки как понятия и её способность интегрироваться в более крупные жанры, как, например, в произведениях «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 ли-

тературы. Наши выводы свидетельствуют как о проблемах, так и о перспективах в этой области.

БЛАГОДАРНОСТИ

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

ДЕКЛАРАЦИЯ КОНФЛИКТА ИНТЕРЕСОВ

Авторы заявлют об отсутствии конфликта интересов.

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). Suomalaisen Tiedeakatemian Toimituksia.

Andreev, N. P. (1929). *Index of fairy-tale plots according to the Aarne System*. Russian Geographical Society.

Atagün, E., Hartoka, B. & Albayrak A. (2021). Topic modeling using LDA and BERT Techniques: Teknofest example.*6th International Conference on Computer Science and Engineering* (pp. 660–664). Akdeniz University Publisher. https://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. *Humaine Association Conference on Affective Computing and Intelligent Interaction* (pp. 141–146). IEEE. https://doi.org/10.1109/ACII.2013.30

Batraeva, I. A., Nartsev, A. D., & Lezgyan, A.S. (2020). Using the analysis of semantic proximity of words in solving the problem of determining the genre of texts within deep learning", *Tomsk State University Journal of Control and Computer Science*, *50*, 14–22. https://doi.org/10.17223/19988605/50/2

Bayer, M., Kaufhold, M.-A., & Reuter, Ch. (2021). *A survey on data augmentation for text classification*. arXiv preprint. arXiv:2107.03158. https://doi.org/10.48550/arXiv.2107.03158

Chan, B., Schweter, S., & Möller, T. (2020). German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6788–6796). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.598

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, https://doi.org/10.48550/arXiv.1810.04805

Dubovik, A.R. (2017). Automatic text style identification in terms of statistical parameters. *Komp'yuternaya lingvistika i vychislitel'nye ontologii*, *1,* 29–45. https://doi.org/10.17586/2541-9781-2017-1-29-45

Fu, Z., Zhou W., Xu J., Zhou H., & Li L. (2022). Contextual representation learning beyond Masked Language Modeling. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (vol. 1: Long Papers, pp. 2701-2714). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.193

El-Halees, A. M. (2017). Arabic text genre classification. *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. *Doklady Mathematics, 106* (Suppl. 1), 95–96. https://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. https://doi.org/10.1109/ACCESS.2020.2994252

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*, *9*(19), 4062. https://doi.org/10.3390/app9194062

Karsdorp, F. & Bosch, Van den A. (2013). Identifying motifs in folktales using topic models. *Proceedings of BENELEARN 2013* (pp. 41–49). Radboud University. https://hdl.handle.net/2066/112943

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

Kessler B., Numberg G. & Schütze H. (1997). Automatic detection of text genre. *Proceedings of the Eighth Conference on European chapters of the Association for Computational Linguistics.* (pp. 32–38). Association for Computational Linguistics. https://doi.org/10.3115/976909.979622

Kupriyanov, R.V., Solnyshkina, M.I. & Lekhnitskaya, P.A. (2023). Parametric taxonomy of educational texts. *Science Journal of VolSU. Linguistics, 22*(6), 80–94. 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. *Proceedings of the 15th Conference on Natural Language Processing* (pp. 9–11). Erlangen.

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. https://doi.org/10.18255/1818-1015-2021-3-280-291

Lagutina, K. V. (2023). Genre classification of Russian texts based on Modern Embeddings and Rhythm. *Automatic Control and Computer Sciences, 57*(7), 817–827. https://doi.org/10.3103/S0146411623070076

Lai, Y. A., Lalwani, G. & Zhang, Y. (2020). context analysis for pre-trained masked language models. *Findings of the Association for Computational Linguistics* (pp. 3789–3804). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.338

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. 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. https://doi.org/10.48550/arXiv.2003.13016

Lippert, Ch., Junger, A., Golam R., Md., Mohammad Ya., Hasan Sh., Md, & Chowdhury, Md. (2022). *Kuzushiji (Japanese Text) classification*. Technical Report. https://doi.org 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* (pp. 486–489). CONFCDS. Publisher. https://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. https://doi.org/10.18653/v1/2020.acl-main.645

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2022). Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys, 54*(3), 1–40. https://doi.org/10.1145/3439726

Nikolaev, P.L. (2022) Classification of books into genres based on text descriptions via deep learning. *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. *Proceedings of the KONVENS 2012* (pp. 378–382). ASAI. http://www.oegai.at/konvens2012/proceedings/56_nguyen12w/

Nguyen, D., Trieschnigg, D., Meder, Th., & Theune, M. (2013) Folktale classification using learning to rank. *Proceedings of the European Conference on Information Retrieval*. *Lecture Notes in Computer Science* (vol. 7814, pp. 195–206). Springer. https://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* (pp. 1–4). IEEE. https://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. https://doi.org/ 10.18653/v1/N18-1202

Pompeu, D. P. (2019). *Interpretable deep learning methods for classifying folktales according to the Aarne-Thompson-Uther Scheme* [Master's Thesis]. Instituto Superior Técnico.

Propp, V. (1984). *The Russian fairy tale*. Izd. LSU.

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. *NPJ Digital Medicine, 4*(1)*, 86.* https://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*. Apress eBooks. https://doi.org/10.1007/978-1-4842-6664-9

Samothrakis, B. S., & Fasli, M. (2015). Emotional sentence annotation helps predict fiction genre. *PloS One, 10*(11), e0141922. https://doi.org/10.1371/journal.pone.0141922

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.* https://doi.org/10.48550/arXiv.1806.01822

Solnyshkina, M.I., Kupriyanov, R.V. & Shoeva, G.N. (2024). Linguistic profiling of text: Adventure story vs. Textbook. In *Scientific Result. Questions of Theoretical and Applied Linguistics*, *10*(1), 115-132. https://doi.org/10.18413/2313-8912-2024-10-1-0-7 (In Rus).

Solovyev, V., Solnyshkina, M., & Tutubalina, E. (2023). Topic modeling for text structure assessment: The case of Russian academic texts. *Journal of Language and Education*, *9*(3), 143-158. https://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. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 1441–1450). Association for Computing Machinery. https://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. https://doi.org/10.1111/oli.12458

Tianqi, Ch. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. 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 *16th International Symposium on Embedded Multicore/Many-core Systems-on-Chip* (pp. 366-370). IEEE. https://doi.org/10.1109/MCSoC60832.2023.00061

Tudorovskaya, E.A. (1961). On classification of Russian folk fairy tales. Specifics of Russian folklore genres. *Specificity of genres of Russian folklore: Theses of the report.* Institute of Russian Literature (Pushkin House).

Uther, H.-J. (2004). The types of international folktales: A classification and bibliography, based on the system of Antti Aarne and Stith Thompson. *Folklore Fellows' Communications* (vol. 3, pp. 284–286). Suomalainen Tiedeakatemia.

Thompson, S. (1928). The types of the folk-tale: A classification and bibliography. *Folklore Fellows' Communications,* (74). Suomalainen Tiedeakatemia.

Thompson, S. (1977). *The folktale*. University of California Press.

Wang, Z., Wu, H. Liu, H. & Cai, Q.-H. (2020). BertPair-networks for sentiment classification. *2020 International Conference on Machine Learning and Cybernetics* (pp. 273–278). IEEE Xplore. https://doi.org/10.1109/ICMLC51923.2020.9469534

Worsham, B, J., & Kalita, J. (2018). Genre identification and the compositional effect of genre in literature. *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1963–1973). Association for Computational Linguistics. https://aclanthology.org/C18-1167

Xiong, H. & Wu, J. & Liu, L. (2010). Classification with ClassOverlapping: A systematic study. *1st International Conference on E-Business Intelligence* (pp. 303–309). Atlantis Press. https://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