УДК 82+004.8 DOI: 10.31249/chel/2025.01.12

Максименко П.И.[1]

ЖАНРОВАЯ КЛАССИФИКАЦИЯ ЛИТЕРАТУРНЫХ ТЕКСТОВ С ПРИМЕНЕНИЕМ НЕЙРОСЕТЕВЫХ МЕТОДОВ

(НА МАТЕРИАЛЕ РУССКОЯЗЫЧНОЙ ЭЛЕКТРОННОЙ БАЗЫ ФАНФИКШН)©

НИУ «Высшая школа экономики»,
Россия, Санкт-Петербург, p.maksimenko@hse.ru

Аннотация. Статья посвящена дообучению и применению большой языковой модели типа BERT для решения задачи классификации литературных текстов по жанрам. В качестве источника материала для создания обучающей выборки используются фанфикшн-произведения, входящие в русскоязычную электронную базу фанфикшн, которая насчитывает более 160 тыс. текстов. Набор данных для обучения нейросетевого алгоритма содержит фанфикшн-тексты, каждый из которых имеет одну из восьми жанровых меток или более. В статье представлены результаты апробации и оценки эффективности трех версий жанрового классификатора (многометочная, многоклассовая и бинарная) как на исходных данных, так и на тестовой выборке художественной литературы. В ходе исследования также были сопоставлены показатели качества классификации для текстов разных жанров.

Ключевые слова: фанфикшн; жанровая классификация; массовая литература; BERT.

Получена: 16.09.2024 Принята к печати: 20.10.2024

[1] Публикация подготовлена в результате проведения исследования по проекту «Текст как Big Data: методы и модели работы с большими текстовыми данными» в рамках Программы фундаментальных исследований НИУ ВШЭ в 2024 г.

Maksimenko P.I.[1]

Genre classification of literary texts through neural network methods (based on the russian-language fanfiction electronic database)©

National Research University Higher School of Economics,
Russia, *Saint-Petersburg*, *p.maksimenko@hse.ru*

Abstract. The article focuses on the fine-tuning and application of a large language model based on BERT for genre classification of literary texts. The training data for the neural network algorithm were created based on fanfiction works from a Russian-language fanfiction electronic database containing more than 160,000 texts. The training dataset for the neural network algorithm contains fanfiction texts, each of which is marked with one of eight genre labels or more. The article presents the results of testing and evaluation of the effectiveness of three genre classifier versions (multi-label, multi-class, and binary) on both the original dataset and a test sample of literary fiction. The research also includes the comparison of classification quality values for texts of different genres.

Keywords: fanfiction; genre classification; mass literature; BERT.

Received: 16.09.2024 Accepted: 20.10.2024

Введение

Задача определения жанра литературного произведения как одной из его основополагающих характеристик была и остается актуальной по сей день. В настоящее время ввиду внушительных объемов текстов все чаще для решения этой задачи применяют компьютерные методы обработки естественного языка, в частности, глубокие нейронные сети. Вычислительные методы, позволяющие исследовать большие текстовые массивы, широко используются для изучения художественной литературы. Этот подход, предложенный Ф. Моретти, называют «дальним чтением» (*distant reading*) [Moretti, 2013]. Примерами использования компьютерных методов для классификации художественных текстов служат исследования С. Гупты и др., А. Гояла и П. Вуппулури, В.Б. Барахнина и др., П.Л. Николаева, К.В. Лагутиной [Gupta, Agarwal, Jain, 2019;

[1] The paper was prepared within the framework of the Basic Research Program at HSE University in 2024 (project “Text as big data: methods and models for working with large textual data”).

Goyal, Vuppuluri, 2022; Автоматизированная классификация …, 2017; Николаев, 2022; Lagutina, 2022]. Алгоритмы автоматической жанровой классификации представляют интерес не только с исследовательской, но и прикладной точки зрения: они могут быть применены для классификации текстов в электронных библиотеках, базах данных.

Одним из наиболее эффективных инструментов для создания цифрового представления языка остается нейросетевая модель BERT, созданная на основе архитектуры Transformer в 2019 г., к тому же она предоставляет возможность дообучения (*fine-tuning*) на целевых данных для решения задачи классификации [BERT …, 2019]. В рамках настоящего исследования применяется основанная на многоязычной модели BERT дистиллированная модель rubert-tiny2 от разработчика Cointegrated, имеющая 12 млн параметров, 3 скрытых слоя и 12 голов внимания[1]. Эта версия модели обладает такими преимуществами, как высокая скорость обучения и сопоставимая с базовыми моделями производительность при ограниченных вычислительных мощностях, а также имеет расширенную с 512 до 2048 токенов максимальную длину входной последовательности, что позволяет обрабатывать более объемные тексты, в том числе литературные.

Отметим, что в рамках исследования решается задача автоматической классификации, которая состоит в отнесении каждого документа из коллекции к определенному классу с заранее известными параметрами [Автоматическая обработка текстов …, 2017, с. 11]. Более того, для выявления оптимального алгоритма сопоставляются три различных типа классификации: *бинарная* (объекты распределяются по двум непересекающимся классам), *многоклассовая* (multi-class, множественные классы обучающих данных не пересекаются) и *многометочная* (multi-label, допускает наличие нескольких пересекающихся меток в рамках одного объекта) [Епрев, 2010, с. 66; Multi-label classification …, 2020].

Материал исследования

В качестве материала исследования выступает фанфикшн – один из самых популярных и быстро развивающихся видов сетевой массовой литературы [Coppa, 2006]. Его применение обусловлено

[1] Rubert-tiny2 / HuggingFace. – URL: https://huggingface.co/cointegrated/rubert-tiny2

обширностью и разнообразием существующих текстов, наличием жанровой разметки и доступностью извлечения. Источником обучающих данных в настоящем исследовании является русскоязычная электронная база фанфикшн-текстов, включающая в себя 160 986 текстов общим объемом 623 057 864 токенов и метаданные о них, которые были автоматически извлечены с самого популярного специализированного фанфикшн-ресурса на русском языке «Книга фанфиков»[1] [Максименко, 2023]. Среди метаданных выделяется категория жанров, которая послужила критерием распределения текстов на классы.

Фанфикшн отличается от других литературных форм в первую очередь вторичностью по отношению к оригиналу (канону), а также тем, что создается в рамках интерпретативных фанатских сообществ (фандомов) [Попова, 2006; Copra, 2006]. Можно утверждать, что фанфикшн существует параллельно конвенциональному литературному миру, однако в то же время обладает многими свойствами текстов массовой литературы, в частности, жанры фанфикшн во многом опираются на традиционные литературные жанры, расширяют и переосмысляют их [Самутина, 2013, с. 151; Антипина, 2011]. Жанр в рамках феномена фанфикшн можно определить как «типичную модель построения фанатских текстов, отражающую общие черты для группы конкретных произведений»: основные события, сюжетные элементы, эмоциональную составляющую [Коробко, 2015, с. 156].

Для дообучения модели жанровой классификации были выбраны фан-тексты, содержащие хотя бы одну из следующих восьми меток: драма, фэнтези, мистика, экшн, фантастика, ужасы, детектив, приключения. Данные жанровые метки были выбраны ввиду высокой частотности среди фанфикшн-произведений и близости к жанрам, составляющим «жанровое ядро» массовой литературы [Купина, Литовская, Николина, 2009, с. 107]. Описание тематических тегов во многом соответствует схожим жанрам массовой литературы по содержательным и сюжетным элементам, в то время как структурная составляющая текста никак не ограничивается.

[1] С 9 июля 2024 г. доступ к ресурсу «Книга фанфиков» http://ficbook.net ограничен на территории Российской Федерации по требованию Роскомнадзора по статье 15.1 Федерального закона от 27.07.2006 № 149-ФЗ «Об информации, информационных технологиях и о защите информации».

Подготовка обучающих датасетов

Для последующего дообучения модели на основе электронной базы было сформировано три датасета из фанфикшн-текстов: для многометочной, многоклассовой и бинарной классификации (табл. 1).

Таблица 1.

Описание трех датасетов для разных типов классификации

Тип классификации	Многометочная	Многоклассовая	Бинарная
Количество классов	8	8	2
Количество текстов	42 145	25 426	14 702
Драма	12 806	4171	–
Фэнтези	8453	3649	7822
Мистика	8541	2577	–
Экшн	7199	2740	–
Фантастика	7113	3182	–
Ужасы	7030	2732	–
Детектив	7511	3143	6880
Приключения	7031	3232	–
Общий объем датасета (в токенах)	154 576 751	76 645 441	65 039 114
Средняя длина текста (в токенах)	3668	3014	4424

В датасете для многометочной классификации каждый из текстов имеет не менее одной из восьми указанных жанровых меток, для многоклассовой классификации каждый текст отмечен лишь одним тегом из перечисленных. Для бинарной классификации были выбраны два жанра, реже всего пересекающиеся в рамках фан-работ и наиболее разнящиеся содержательно: детектив и фэнтези. При формировании датасетов были удалены выбросы и нерелевантные документы: был произведен отбор текстов по длине (от 200 до 10000 токенов), также из выборки были исключены тексты, одновременно относящиеся более чем к шести жанрам и/или включающие отрывки разных жанров, и стихотворные работы.

В наибольшей выборке общее количество документов составило 42 145 суммарным объемом около 154,5 млн токенов, на каждый класс приходится более семи тыс. текстов. Объем датасета для многоклассовой классификации вдвое меньше – около 76,6 млн токенов, в него входит 25 426 текстов. Для бинарной классификации были отобраны детективные и фэнтези-тексты, которые относятся только к одному из двух классифицируемых жанров, но могут быть отмечены любыми другими жанровыми тегами. Общий объем третьего датасета составил 14 702 текста, 65 039 114 токенов. По приведенным данным заметно, что тексты жанров «детектив» и «фэнтези» значительно длиннее в сравнении с общими средними значениями.

Тексты каждого из датасетов были предобработаны с применением инструментов пакета NLTK[1] следующим образом: была проведена токенизация, удалены знаки препинания и стоп-слова, тексты были приведены к нижнему регистру.

Дообучение модели жанровой классификации

Дообучение модели rubert-tiny2 для жанровой классификации проводилось на языке программирования Python в среде Google Colaboratory (Colab)[2] с применением инструментов от разработчиков HuggingFace[3], библиотек PyTorch[4] и Scikit-Learn[5], предназначенных для машинного и глубокого обучения. Все тексты были токенизированы и преобразованы в эмбеддинги (векторное представление) с помощью встроенного токенизатора rubert-tiny2 для подачи на вход модели. Метки классов были также преобразованы в цифровой формат.

Для дообучения всех типов классификаторов использовались следующие значения гиперпараметров, подобранные экспериментальным путем (поиск по сетке):

– обучающая выборка 80% данных, валидационная 10%, тестовая 10%;

– максимальная длина входной последовательности 2048;

[1] NLTK 3.8.1 Documentation – URL: https://www.nltk.org/

[2] Google Colab – URL: https://colab.google/

[3] HuggingFace Documentation – URL: https://huggingface.co/docs

[4] PyTorch 2.3 Documentation – URL: https://pytorch.org/docs/stable/index.html

[5] Scikit-Learn 1.4.2. Documentation – URL: https://scikit-learn.org/stable/index.html

- размер батча 8;
- количество эпох обучения 4;
- коэффициент скорости обучения (learning rate) 3e-5;
- оптимизатор AdamW[8];
- доля исключения случайных нейронов (dropout) 0,1%.

По результатам четырех эпох обучения лучшая версия каждой из моделей была определена по наибольшему значению F1-меры на валидационной выборке.

Оценка эффективности дообученных моделей

Оценка качества классификации производилась на тестовой выборке с использованием метрик точности, полноты и F1-меры. Значения метрик оценки качества для модели многометочной классификации представлены в табл. 2 (здесь и далее полужирным шрифтом выделены наибольшие значения показателей по жанрам).

Таблица 2.

Значения метрик оценки качества многометочного классификатора

Класс	Точность	Полнота	F1-мера
Драма	0,64	0,43	0,52
Фэнтези	0,58	**0,53**	0,55
Мистика	0,53	0,38	0,45
Экшн	0,59	0,34	0,44
Фантастика	**0,76**	0,51	**0,61**
Ужасы	0,64	**0,53**	0,58
Детектив	**0,74**	**0,6**	**0,66**
Приключения	0,64	0,4	0,49
Микро-усреднение (micro-average)	0,64	0,46	0,53
Макро-усреднение (macro-average)	0,64	0,46	0,54
Взвешенное усреднение (weighted average)	0,64	0,46	0,53
Выборочное усреднение (samples average)	0,6	0,51	0,53

Качество модели многометочной классификации составляет в среднем 53%, вместе с тем эффективность определения разных классов различается. Так, точнее всего модель определяет тексты жанров «фантастика» и «детектив», в то же время высокие значения показателя полноты относятся к классам «ужасы» и «фэнтези». Эти четыре жанра можно определить как более качественно классифицируемые (F-мера выше среднего значения). Значительно хуже распознаются моделью жанры «экшн», «мистика», «приключения» и «драма».

Перейдем к рассмотрению результатов многоклассовой классификации (табл. 3, рис. 1).

Таблица 3.

Значения метрик оценки качества многоклассового классификатора

Класс	Точность	Полнота	F1-мера
Драма	**0,65**	**0,68**	**0,66**
Фэнтези	0,55	0,63	0,59
Мистика	0,46	0,29	0,36
Экшн	0,45	0,45	0,45
Фантастика	**0,65**	0,65	0,65
Ужасы	0,53	0,63	0,58
Детектив	**0,72**	**0,72**	**0,72**
Приключения	0,52	0,48	0,5
Доля правильных ответов (accuracy)	–	–	0,58
Макро-усреднение (macro-average)	0,57	0,57	0,56
Взвешенное усреднение (weighted average)	0,58	0,58	0,58

В сравнении с многометочным классификатором качество модели повысилось, но не существенно: F-мера составляет 0,58, что на 0,05 больше, чем в предыдущем эксперименте. Среди наиболее точно классифицируемых жанров снова выделяются «детектив» и «фантастика», но в отличие от первой модели значительно эффективнее распознается жанр «драма» (за счет прироста по показателю полноты). Наименее качественно классифицируются жанры «мистика», «экшн» и «приключения», причем мистические тексты верно определяются в среднем реже, чем в случае многометочной классификации.

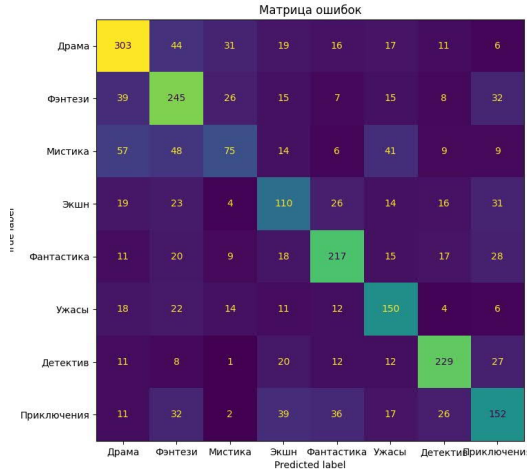

Рис. 1. Матрица ошибок многоклассового классификатора на тестовой выборке

Рассмотрим качество бинарной классификации по жанрам фэнтези и детектива (табл. 4, рис. 2).

Таблица 4.

Значения метрик оценки качества бинарного классификатора

Класс	Точность	Полнота	F1-мера
Фэнтези	0,89	**0,96**	**0,92**
Детектив	**0,95**	0,87	0,9
Доля правильных ответов (accuracy)	–	–	0,92
Макро-усреднение (macro-average)	0,92	0,91	0,91
Взвешенное усреднение (weighted average)	0,92	0,92	0,91

Качество этой версии модели в среднем составляет 91%, что свидетельствует о возможности разграничивать тексты выбранных жанров с минимальной долей ошибок. Отметим, что для жанра «фэнтези» полнота превышает точность на 0,07, в то время как для жанра «детектив» характерно обратное распределение значений – точность выше полноты на 0,08. Усредненное значение F-меры для фэнтезийных текстов на 2% больше, чем для детективных.

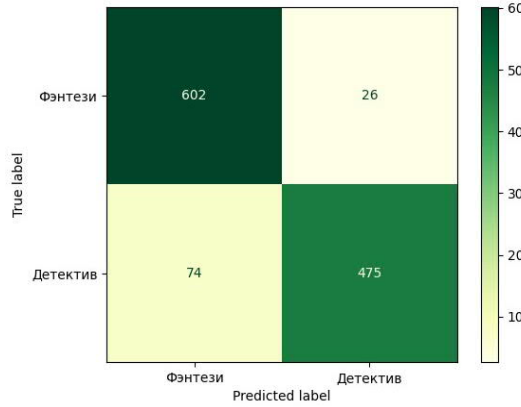

Рис. 2. Матрица ошибок бинарного классификатора на тестовой выборке

Таким образом, модели, дообученные для многометочной и многоклассовой классификации, показывают сопоставимо низкий уровень качества, причем эффективность разнится в зависимости от конкретного жанра. Модель для бинарной классификации достигает качества более 90%. Оценка производительности моделей подтверждает зависимость качества классификации от количества классов: с увеличением числа тегов повышается сложность решения задачи [Moral, Nowaczyk, Pashami, 2022].

Апробация модели на тестовой выборке художественной литературы

Дообученные версии модели также были апробированы на материале массовой художественной литературы, в частности на текстах рассказов, отрывках повестей и романов детективного, фантастического, приключенческого и других жанров. Произведения для апробации были выбраны с учетом четкого отнесения к жанру, популярности автора и/или текста и общего объема (поиск осуществлялся по запросу лучших классических или современных произведений того или иного жанра), причем при расхождениях в определении жанра предпочтение отдавалось текстам, подходившим под описание жанровой метки именно в рамках фанфикшн (например, «драма»). Таким образом, была сформирована тестовая

выборка из 40 документов, где на каждый из представленных жанров приходится пять текстов (табл. 5). Тестирование бинарного классификатора на десяти текстах жанров «детектив» и «фэнтези» показало безошибочный результат: все объекты двух классов были определены моделью верно, F1-мера составила 1,0.

В свою очередь, производительность многоклассовой и многометочной модели была существенно ниже. Предсказания двух классификаторов представлены в табл. 5 (прочерком отмечены тексты, для которых модель не предсказала ни одного положительного класса).

Таблица 5.

Предсказания многометочного и многоклассового классификатора на материале художественной литературы

Название	**Автор**	**Жанр**	**Предсказание многометочного классификатора**	**Предсказание многоклассового классификатора**
Серебряный	А.К. Дойл	Детектив	Детектив	Детектив
Тайна голубой вазы	А. Кристи	Детектив	Мистика	Детектив
Три всадника из «Апокалипсиса»	Г.К. Честертон	Детектив	Экшн	Экшн
Убийство на улице Морг	Э. А. По	Детектив	Фантастика	Фантастика
Липучка для мух	Д. Хэммет	Детектив	Детектив	Детектив
Кузнец из Большого Вуттона	Дж. Р.Р. Толкиен	Фэнтези	Фэнтези	Фэнтези
Путешествие Короля	Лорд Дансени	Фэнтези	Фэнтези	Фэнтези
Нечто большее	А. Сапковский	Фэнтези	Ужасы	Ужасы
Дурацкое задание	Дж. Аберкромби	Фэнтези	Фэнтези	Фэнтези
Повелитель Горной долины	Н. Гейман	Фэнтези	Мистика, детектив	Мистика
Голуби из ада	Р.И. Говард	Ужасы	Мистика, ужасы	Ужасы
Клацающие зубы	С. Кинг	Ужасы	Ужасы	Ужасы
Зов Ктулху	Г.Ф. Лавкрафт	Ужасы	Мистика, ужасы	Мистика
Птицы	Д. Дюморье	Ужасы	Мистика, ужасы	Ужасы
Полночный поезд с мясом	К. Баркер	Ужасы	Мистика, детектив	Детектив
Шесть спичек	А. и Б. Стругацкие	Фантастика	Фантастика	Детектив
Робби	А. Азимов	Фантастика	Фантастика	Фантастика
Чудовище	А.В. Вогт	Фантастика	Фантастика	Фантастика

Название	Автор	Жанр	Предсказание многометочного классификатора	Предсказание многоклассового классификатора
Кое-что задаром	Р. Шекли	Фантастика	Фантастика	Фантастика
Третья экспедиция	Р. Брэдбери	Фантастика	Фантастика	Фантастика
Мы с моей тенью	Э.Ф. Рассел	Мистика	Ужасы	Ужасы
Не спешу	С. Лукьяненко	Мистика	—	Фантастика
Вий	Н. В. Гоголь	Мистика	Фэнтези	Фэнтези
Призрак двадцатого века	Дж. Хилл	Мистика	Ужасы	Ужасы
Неопытное привидение	Г. Уэллс	Мистика	Мистика	Мистика
Гонки	Дж. Лондон	Приключения	Приключения	Приключения
Коварная коробка	Дж. Даррелл	Приключения	Приключения	Приключения
Повесть об английском докторе и дорожном сундуке	Р.Л. Стивенсон	Приключения	—	Фантастика
«Катти Сарк»	И. Ефремов	Приключения	Экшн, приключения	Приключения
Сто верст по реке	А. Грин	Приключения	Приключения	Приключения
Рыцари убойных новостей из Трейл-Сити	К. Саймак	Экшн	Экшн, детектив	Экшн
Что золото делает с человеком	Л. Ламур	Экшн	Приключения	Приключения
Список ликвидации	Дж. Карр	Экшн	Экшн	Экшн
Те же и Скунс	М. Семёнова, Е. Перехвальская, В. Воскобойников	Экшн	Приключения	Приключения
Крик дьявола	У.Смит	Экшн	Экшн	Экшн
Гордость и предубеждение	Дж. Остен	Драма	Драма, детектив	Детектив
Лучшее во мне	Н. Спаркс	Драма	Фантастика	Приключения
Лорна Дун	Р.Д. Блэкмор	Драма	Фэнтези, приключения	Приключения
История любви	Э. Сигл	Драма	Детектив	Детектив
Валентайн	Э. Уэтмор	Драма	Драма, экшн	Экшн

Многометочная модель наиболее точно классифицировала жанры «фантастика», «ужасы» и «приключения», а многоклассовая модель – «приключения» и «фантастику». Несмотря на высокие показатели качества распознавания детективных текстов на этапе кросс-валидации, многометочная модель верно определила

всего два из пяти произведений этого жанра при апробации. Многоклассовая модель продемонстрировала нулевую точность при классификации жанра «драма» – эта метка не была присвоена ни одному из текстов. Класс «мистика» был предсказан обеими моделями крайне неэффективно.

Можно утверждать, что многометочный классификатор демонстрирует более высокую эффективность на тестовой выборке в сравнении с многоклассовой моделью, так как обладает преимуществом – возможностью отнести текст одновременно к нескольким классам, что повышает вероятность попадания в верную метку. Так, при учете указания на фактический класс вне зависимости от количества предсказанных классов многометочная модель достигает точности в 60%, в то время как многоклассовая модель – 52,5%. Тем не менее важно отметить, что прогнозы моделей имеют высокую согласованность – предсказания не имеют пересечений только в пяти случаях из 40, в двух из которых многометочный классификатор не отнес текст ни к одному из классов. Полученные данные также указывают на то, что метки, предсказанные моделью для одного текста, зачастую оказываются смежными или тематически связанными (мистика и ужасы, мистика и детектив, экшн и приключения, фэнтези и приключения).

На основе приведенных данных можно сделать вывод о целесообразности применения многометочного типа классификации при определении жанра текста компьютерными методами.

Заключение

В рамках исследования на материале фанфикшн-текстов были дообучены три версии жанрового классификатора. По результатам оценки качества самую высокую производительность показала модель бинарной классификации для жанров «фэнтези» и «детектив», в то время как многоклассовая и многометочная классификации по восьми жанрам существенно менее эффективны. Качество предсказаний многоклассового классификатора в среднем незначительно выше, однако, основываясь на данных, полученных в ходе апробации, применение модели с пересекающимися классами представляется релевантным в случае совмещения нескольких жанров в рамках одного текста. К тому же, многометочная модель применима к современным формам сетевой литературы (включая фанфикшн), которым присуща жанровая конвергенция [Лебедева, 2015].

Низкий уровень производительности классификаторов множественного типа можно объяснить разнородностью обучающих данных: фанфикшн-тексты существенно отличаются не только по объему и лексическому наполнению, но и по формату, фандому, рейтингу. Кроме того, нельзя исключать наличие в датасетах произведений низкого качества и работ, для которых неверно определены жанровые метки. Перечисленные причины могут существенно усложнять задачу классификации и, как следствие, оказывать негативное влияние на эффективность модели.

В отношении качества классификации отдельных жанров наблюдаются устойчивые тенденции: более успешно распознаются жанры «детектив», «фантастика», «ужасы», в то время как наибольшее количество ошибок обе версии модели совершают при классификации жанров «мистика», «экшн», «приключения». Жанр «драма» классифицируется многоклассовой моделью значительно эффективнее, чем многометочной.

В качестве перспектив настоящего исследования можно выделить расширение электронной базы фанфикшн и введение более строгих критериев отбора текстов для формирования датасетов, использование техник оптимизации для повышения производительности полученных моделей, применение других существующих методов и алгоритмов автоматической классификации текстов.

Список литературы

Автоматизированная классификация русских поэтических текстов по жанрам и стилям / Барахнин В.Б., Кожемякина О.Ю., Пастушков И.С., Рычкова Е.В. // Вестник Новосибирского государственного ун-та. Серия: Лингвистика и межкультурная коммуникация. – 2017. – Т. 15. – № 3. – С. 13–23.

Автоматическая обработка текстов на естественном языке и анализ данных / Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. – Москва : Изд-во НИУ ВШЭ, 2017. – С. 7–30.

Антипина Ю.В. Жанровые особенности фанатской прозы (на примере фанфикшена по творчеству братьев Стругацких) // Вестник ЧелГУ. – 2011. – № 13. – С. 21–25.

Епрев А.С. Автоматическая классификация текстовых документов // Математические структуры и моделирование. – 2010. – №1(21). – С. 65–81.

Коробко М.А. Жанр в фанфикшн: закономерности использования (на материалах фандомов «Шерлок», «Мерлин», «Сверхъестественное») // Ученые записки Орловского государственного университета. – 2015. – № 6(69). – С. 154–157.

Купина Н.А., *Литовская М.А.*, *Николина Н.А.* Массовая литература сегодня. – Москва: Флинта: Наука, 2009. – 424 с.

Лебедева М.Н. О причинах сюжетной редукции в сверхкратких рассказах // Вестник Тверского государственного университета. Серия: Филология. – 2015. – № 3. – С. 308–312.

Максименко П.И. Русскоязычная электронная база фанфикшн-текстов: принципы создания и анализ метаданных // Информационные технологии в гуманитарных исследованиях: материалы междунар. науч.-практ. конф., Красноярск, 25–28 сентября 2023 г. – Красноярск: СФУ, 2023. – С. 151–159.

Николаев П.Л. Классификация книг по жанрам на основе текстовых описаний посредством глубокого обучения // International Journal of Open Information Technologies. – 2022. – Т. 10. – № 1. – С. 36–40.

Попова С.Н. Произведения «фанфикшн» как особый вид вторичных текстов // Языки в современном мире: материалы V международной конференции. – Москва : КДУ, 2006. – С. 585–589.

Самутина Н. Великие читательницы: фанфикшн как форма литературного опыта // Социологическое обозрение. – 2013. – Т. 12. – № 3. – С. 137–191.

BERT: pre-training of Deep Bidirectional Transformers for language understanding / Devlin J., Chang M., Lee K., Toutanova, K. // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2019. – Vol. 1. – P. 4171–4186.

Coppa F. A Brief history of media fandom // Fan Fiction and Fan Communities in the Age of the Internet / ed. by: K. Busse and K. Hellekson. – Jefferson: McFarland, 2006. – P. 41–60.

Goyal A., *Vuppuluri P.* Statistical and deep learning approaches for literary genre classification // Advances in Data and Information Sciences. – 2022. – Vol. 318. – P. 297–305.

Gupta S., *Agarwal M.*, *Jain S.* Automated genre classification of books using Machine Learning and Natural Language Processing // Proceedings of the 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). – 2019. – P. 269–272.

Lagutina K. Classification of Russian texts by genres based on Modern Embeddings and Rhythm // Modeling and Analysis of Information Systems. – 2022. – Vol. 29. – P. 334–347.

Moral P.D., *Nowaczyk S.*, *Pashami S.* Why is multiclass classification hard? // IEEE Access. – 2022. – Vol. 10. – P. 80448–80462.

Moretti F. Distant reading. – London ; New York : Verso Books, 2013.

Multi-label classification of text documents using deep learning / Mohammed H.H., Dogdu E., Gorur A.K., Choupani R. // 2020 IEEE International Conference on Big Data (Big Data). – 2020. – P. 4681–4689.

References

Barahnin, V.B., Kozhemjakina, O.Ju., Pastushkov, I.S., Rychkova, E.V. (2017). Avtomatizirovannaja klassifikatsija russkih pojeticheskih tekstov po zhanram i stiljam. *Vestn. Novosib. gos. un-ta. Serija: Lingvistika i mezhkul'turnaja kommunikatsija*, *15*(3), 13–23.

Bolshakova, E.I., Vorontsov, K.V., Efremova, N.Je., Klyshinskij, Je.S., Lukashevich, N.V., Sapin, A.S. (2017). In: *Avtomaticheskaja obrabotka tekstov na estestvennom jazyke i analiz dannyh*, pp. 7–30. Moscow: Izd-vo NIU VshJe.

Antipina, Ju.V. (2011). Zhanrovye osobennosti fanatskoj prozy (na primere fanfikshena po tvorchestvu brat'ev Strugackih). *Vestnik ChelGU*, *13*, 21–25.

Eprev, A.S. (2010). Avtomaticheskaja klassifikatsija tekstovyh dokumentov. *Matematicheskie struktury i modelirovanie*, *1*(21), 65–81.

Korobko, M.A. (2015). Zhanr v fanfikshn: zakonomernosti ispolzovanija (na materialah fandomov "Sherlok", "Merlin", "Sverhestestvennoe"). *Uchenye zapiski Orlovskogo gosudarstvennogo universiteta*, *6*(69), 154–157.

Kupina, N.A., Litovskaja, M.A., Nikolina, N.A. (2009). *Massovaja literatura segodnja*. Moscow: Flinta, Nauka.

Lebedeva, M.N. (2015). O prichinah sjuzhetnoj reduktsii v sverhkratkih rasskazah. *Vestnik Tverskogo gosudarstvennogo universiteta. Serija*: *Filologija*, *3*, 308–312.

Maksimenko, P.I. (2023). Russkojazychnaja jelektronnaja baza fanfikshn-tekstov: printsipy sozdanija i analiz metadannyh. In: *Informacionnye tehnologii v gumanitarnyh issledovanijah*, pp. 151–159. Krasnoyarsk: SFU.

Nikolaev, P.L. (2022). Klassifikatsija knig po zhanram na osnove tekstovyh opisanij posredstvom glubokogo obuchenija. *International Journal of Open Information Technologies*, *10*(1), 36–40.

Popova, S.N. (2006). Proizvedenija "fanfikshn" kak osobyj vid vtorichnyh tekstov. In *Jazyki v sovremennom mire*, pp. 585–589. Moscow: KDU.

Samutina, N. (2013). Velikie chitatelnitsy: fanfikshn kak forma literaturnogo opyta. *Sotsiologicheskoe obozrenie*, *12*(3), 137–191.

Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019). BERT: pre-training of Deep Bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, *1*, 4171–4186.

Coppa, F. (2006). A brief history of media fandom. In Busse, K., Hellekson, K. (eds.) *Fan Fiction and Fan Communities in the Age of the Internet*, pp. 41–60. Jefferson: McFarland.

Goyal, A., Vuppuluri, P. (2022). Statistical and Deep Learning Approaches for literary genre classification. *Advances in Data and Information Sciences*, *318*, 297–305.

Gupta, S., Agarwal, M., Jain, S. (2019). Automated genre classification of books using machine learning and natural language processing. In *Proceedings of the 2019 9th International Conference on Cloud Computing, Data Science & Engineering* (*Confluence*), pp. 269–272.

Lagutina, K. (2022). Classification of Russian texts by genres based on modern embeddings and rhythm. *Modeling and Analysis of Information Systems*, *29*, 334–347.

Moral, P.D., Nowaczyk, S., Pashami, S. (2022). Why is multiclass classification hard? *IEEE Access*, *10*, 80448–80462.

Moretti, F. (2013). *Distant reading*. London; New York: Verso Books.

Mohammed, H.H., Dogdu, E., Gorur, A.K., Choupani, R. (2020). Multi-label classification of text documents using deep learning. In *2020 IEEE International Conference on Big Data* (*Big Data*), pp. 4681–4689.

Об авторе

Максименко Полина Игоревна – стажер-исследователь Лаборатории языковой конвергенции, Национальный исследовательский университет «Высшая школа экономики», Россия, Санкт-Петербург, p.maksimenko@hse.ru

About the author

Maksimenko Polina Igorevna – Intern Researcher, Language Convergence Laboratory, National Research University Higher School of Economics, Russia, Saint-Petersburg, p.maksimenko@hse.ru