# Автоматическая классификация русскоязычных Интернет-текстов по жанрам*

К. В. Лагутина [1], Е.И. Бойчук [2], Н.С. Лагутина [1]

[1] Ярославский государственный университет им. П. Г. Демидова, Ярославль, Россия
[2] Ярославский государственный педагогический университет им. К. Д. Ушинского, Ярославль, Россия

**Аннотация.** Статья посвящена применению современных языковых моделей на основе BERT и трех типов лингвистических характеристик текста для автоматического определения жанра, а также сравнительному анализу данных моделей с точки зрения компьютерной и классической лингвистики. Собран корпус из русскоязычных Интернет-текстов восьми жанров: посты ВКонтакте, комментарии, статьи с портала Хабр, описания компаний, новости, научные статьи, реклама, отзывы на фильмы с сайта Кинопоиск. Каждый текст представлен в виде вектора числовых характеристик с помощью каждой из выбранных моделей: пяти вариаций BERT и лингвистических характеристик уровней символов, структуры и ритма.

**Ключевые слова:** стилометрия, обработка естественного языка, ритмические характеристики, жанры, классификация текстов, BERT.

**DOI** 10.14357/20718594230410 **EDN** BAKBAF

## Введение

Жанр текста является одним из ключевых структурных элементов современной лингвистики и позволяет установить связь между строением текста и его целями и функциями [1]. Такие характеристики используются для решения классических задач информационного поиска, машинного перевода, автоматического аннотирования, классификации документов [2], а также для сложного семантического анализа текста в сфере политики и коммуникации [3], образования [4], в изучении социума и культуры [5].

В современной компьютерной лингвистике нет устоявшегося набора классов жанров. В каждой работе такой набор определяется предметной областью и целями исследования. Однако методы решения в целом соответствуют современным способам классификации текстов. Они включают в себя выбор параметров для моделирования текста и представления его в виде вектора числовых характеристик, алгоритмы классификации и кластеризации многомерных векторов, нейронные сети, в том числе методы глубокого обучения [6].

## 1. Обзор литературы

В работе [7] авторы рассматривали классификацию текстов по трем жанрам: экспрессивные, апеллятивные и информативные. В статье представлен сравнительный анализ нескольких типов признаков: стилометрические, используемые при атрибуции авторства, лингвистические, n-граммы символов, n-граммы частей речи, наиболее часто встречающиеся слова, и различных классификаторов (наивный байесовский метод, метод опорных векторов, логистическая регрессия, k-ближайших

* Работа поддержана стипендией Президента Российской Федерации для молодых ученых и аспирантов, осуществляющих перспективные научные исследования и разработки по приоритетным направлениям модернизации российской экономики: № СП-2109.2021.5.

соседей, случайный лес в сочетании с методами ансамблевого обучения). Лучшие результаты показал ансамбль Random Subspace Random Forest с четырьмя типами признаков (признаки, используемые при атрибуции авторства, n-граммы символов, n-граммы частей речи, наиболее часто встречающиеся слова). Оценкой качества явилась F-мера и составила 94,43 %. Для экспериментов был вручную аннотирован корпус из 3759 отзывов о книгах и фотоаппаратах.

Результаты кластеризации корпуса классических арабских текстов KSUCCA с использованием стилометрических признаков рассматривался в статье [8]. Автор выделил шесть жанров текста (научный, религиозный, литературный, биографический, социологический, лингвистический) и сравнивал четыре меры расстояния между числовыми характеристическими векторами. Эксперименты показали, что кластеризация на основе стилометрии хорошо согласуется с жанрами, определяемыми человеком. Проведенная затем классификация с теми же мерами близости векторов показала лучшее значение метрики качества accuracy – 85 %.

Батраева, Нарцев и Лезгян [9] применили сверточные нейронные сети для анализа текстов с точки зрения определения их жанровой принадлежности. Авторы выбрали пять классов: история, детективы, детская литература, поэзия и песни, фантастика и фэнтези. Числовыми характеристиками текста послужили векторные представления слов на основе модели word2vec. Для классификации исследователи предложили архитектуру сверточной нейронной сети. Эксперименты проводились на корпусе русскоязычных текстов Максима Мошкова, содержащем более 25 000 книг. Точность на тестовой выборке составила 73,12 %. Кроме того, ученые оценили, что вероятность принадлежности произведения определенному жанру вполне коррелирует с литературным пониманием этого текста.

Одним из самых современных и эффективных подходов к автоматической обработке текстов является языковая модель BERT. Le Mens и др. [10] применили эту модель к классификации аннотаций книг по двум жанрам: мистика и романтика. Значение F-меры оказалось 74 % для мистики и 71 % для романтики. Кроме того, авторы определяли корреляцию между результатами классификации BERT и человека. Она оказалась около 90 %, что позволило сделать вывод о том, что модель хорошо отражает человеческие суждения о жанрах. Дополнительно, ученые обратили внимание на слишком малое количество работ с BERT в этой области.

Большой интерес проявляют исследователи к определению жанров Интернет-текстов. В работе [11] представлен подход к классификации больших веб-корпусов по жанрам с помощью функциональных параметров текста (Functional Text Dimension, FTD). Автор исследования выделил 20 видов Интернет-текстов и предложил рассматривать отдельный текст с точки зрения сходства с жанрами-прототипами. Предложенный набор категорий был предназначен для повышения качества аннотирования текстов. Столько же (20) жанров интернет-страниц выделили авторы статьи [12]. Они использовали параметры текста на основе n-грамм символов и эмбеддингов, сгенерированных нейронной сетью GATE v8.1. Исследователи поставили четыре задачи машинного обучения: классификация с одной меткой, где каждая веб-страница аннотируется одним жанром; классификация с несколькими метками где, каждая веб-страница аннотируется несколькими жанрами; иерархическая классификация с одной меткой, где жанры организованы в иерархию; иерархическая классификация с несколькими метками. При решении этих задач использовались алгоритмы машинного обучения для классификации и кластеризации. Лучшая микро F-мера оказалась 95 %.

Задача автоматического определения жанров возникает в разных предметных областях. Авторы работы [13] прогнозировали жанры по обложкам книг. Классификация осуществлялась по тексту и по изображению. Результаты исследования показали, что классификаторы на основе текста превосходят классификаторы на основе изображений.

Da Cunha и Montané [14] провели лингвистический анализ корпуса испаноязычных текстов пяти текстовых жанров, написанных непрофессионалами и адресованных государственным органам (утверждение, сопроводительное письмо, письмо с жалобой, притязание, приложение). Для 100 текстов, по 20 каждого жанра, был проведен полуавтоматический расчет характеристик на основе структуры текста (разделов, заголовков), частей речи, дискурса текста (выявление противопоставления, причины,

следствия и т. п.). Статистический анализ показал, что различия между жанрами проявляются по всем числовым параметрам. В работе [14] обсуждалась необходимость использования числовых характеристик текста, зависящих от предметной области или поставленной задачи.

Горбич и Живодеров [15] рассмотрели и отобрали набор числовых параметров текста (индексов) для классификации научных и научно-популярных текстов. Авторы подчеркивают, что предложенные индексы являются формально-статистическими и применимы для автоматической обработки: частоты n-грамм гласных, индексы энтропии, сжимаемости текста. Точность классификации с использованием линейной регрессии достигла 86 %.

Авторы исследования [16] предложили модель текста в виде графа, которая связывает предложения на основе их семантического сходства. Ученые показали, что с использованием только этой модели без дополнений и оптимизации книги по философии и исследовательские работы распознавались с точностью 92,5 %.

Таким образом, классификация текстов по жанрам является актуальной задачей современной компьютерной лингвистики, где мало исследованы аспекты, связанные с разными типами характеристик текста, включая языковую модель BERT.

Целью работы является классификация Интернет-текстов по жанрам при помощи языковых моделей BERT, характеристик уровней слов, символов и ритма и анализ результатов классификации с лингвистической точки зрения. Рассматривается восемь жанров текстов: посты ВКонтакте, комментарии, статьи с портала Хабр, описания компаний, новости, научные статьи, реклама, отзывы на фильмы с сайта Кинопоиск. Данные жанры представляют собой специфику Интернет-дискурса, который имеет определенные структурные и языковые особенности. В нашем исследовании мы рассмотрим сравнительный аспект изучения текстов с точки зрения таких параметров, как характеристика уровня символов, которая включает в себя частоты букв и знаков препинания, среднюю длину предложения в символах и словах, среднюю длину слова в символах, а также ритмические показатели анализируемых текстов.

С нашей точки зрения, данные параметры не менее точно и полно отображают специфику текстов Интернет-дискурса, а некоторые показатели способствуют разграничению анализируемых жанров.

## 2. Метод

Корпус русскоязычных Интернет-текстов был собран авторами самостоятельно на основе опубликованных текстов и других корпусов. Он содержит 16000 текстов восьми жанров, по 2000 текстов на жанр. Каждый текст соответствует только одному жанру. Жанры включают в себя посты ВКонтакте, комментарии из набора с сайта Kaggle, статьи с портала Хабр, описания компаний, новости lenta.ru, научные статьи из журналов «Грамота», «Диалог» и «Кардиология», рекламу, отзывы на фильмы с сайта Кинопоиск.

Тексты отзывов и посты в социальной сети Вконтакте были собраны через API соответствующих сайтов. Выбирались фильмы и группы, посвященные разнообразным тематикам. Из ВКонтакте выбирались не посты о новостях, которые часто являются дублями статей онлайн-медиа, а оригинальные авторские тексты, созданные именно для социальной сети. Из всех собранных данных были случайным образом выбраны по 2000 текстов каждого жанра, чтобы все категории для последующей классификации имели одинаковый объем. Тексты моделировались двумя способами: как векторы из классических и лингвистических характеристик уровней символов, ритма и структуры, а также как эмбеддинги, построенные на основе языковых моделей BERT для русского языка.

Характеристики уровня символов включают в себя частоты букв и знаков препинания, среднюю длину предложения в символах и словах, среднюю длину слова в символах. Частоты букв брались относительные: количество появлений каждой буквы делилось на общее количество букв в тексте. Аналогично вычислялись частоты для знаков препинания, остальные символы не учитывались.

Характеристики уровня ритма вычислялись при помощи алгоритмов, разработанных в предыдущих исследованиях авторов [17, 18]. Они основаны на наборе ритмических средств: анафоре, эпифоре, симплоке, анадиплозисе, эпаналепсисе, многосоюзии, диакопе, эпизевксисе, хиазме, апозиопезе, повторяющихся

вопросительных и восклицательных предложениях, аллитерации и ассонансе. Аллитерация и ассонанс здесь ищутся как любые повторения гласных и согласных звуков в одном или паре соседних предложений, что является достаточно широкой трактовкой данных терминов. Однако это приводит к тому, что каждый текст содержит хотя бы фонетические ритмические средства, следовательно, для него можно вычислять и ритмические характеристики.

Ритмические характеристики включают в себя: количество появлений в тексте конкретного средства, деленное на количество предложений; доли существительных, прилагательных, глаголов, наречий, имен собственных, местоимений, соединительных союзов, подчинительных союзов, междометий и предлогов среди слов, составляющих средства; максимальное и среднее расстояния между первым и последним повторяющимся в средстве словом. Расстояние измеряется количеством слов.

Характеристики уровня структуры включают в себя n-граммы частей речи, n = 1,2,3,4. Каждый текст представляется как последовательность из частей речи из списка универсальных зависимостей (Universal Dependencies). Далее ищутся 40 самые часто встречающихся n-грамм для каждого n, и частота появлений каждой популярной n-граммы делится на общее количество появлений топ-40 n-грамм для данного n.

Второй способ смоделировать текст — это применить языковую модель, основанную на BERT, мультиязычную или адаптированную для русского языка [19]. Авторы сравнивают пять моделей, включающих две RuBert-модели от DeepPavlov: rubert-base-cased и rubert-base-cased-conversational; две маленькие RuBert-модели от cointegrated: rubert-tiny и rubert-tiny2, а также мультиязычный BERT [20].

У BERT-моделей имеется ограничение на длину входного текста–не более 512 токенов. Из текстов, которые оказались большего размера, были взяты только начальные части. Как покажут эксперименты в следующем разделе, такой объем фрагментов достаточен для определения жанра.

Векторные модели использовались для мультиклассификации текстов на восемь жанров. Корпус текстов был разделен на три выборки: тренировочную, валидационную и тестовую в пропорции 60/20/20 %. Для сравнения были выбраны два нейросетевых классификатора. Первый—это двунаправленная LSTM, рекуррентная нейронная сеть со слоем двунаправленной долгой краткосрочной памяти с 64 блоками и полносвязным выходным слоем. Второй — это перцептрон с одним скрытым слоем, количество нейронов которого на 4 больше числа входных параметров. Функцией активации в выходном слое обеих нейросетей являлась Softmax, оптимизатором — Adam, размер батча составлял 20. Данные гиперпараметры были подобраны на основе экспериментов с валидационной выборкой.

Для обучения применялась технология ранней остановки: если в течение пяти эпох точность на валидационной выборке не изменялась более чем на 0,001, то обучение останавливалось. Таким образом для LSTM обучение производилось за 18-23 эпохи, а для языковых моделей и перцептрона — за 8-12 эпох.

Пятикратная кросс-валидация на всей выборке показала, что выбранные классификаторы с данными гиперпараметрами достаточно стабильны: среднее отклонение для метрик качества не превышало 1,5 (в диапазоне от 0 до 100). Оценка качества классификации на тестовой выборке выполнялась с помощью трех стандартных мер: точность, полнота и F-мера.

## 3. Эксперименты и анализ их результатов

Интернет-тексты классифицировались на восемь жанров при помощи каждого уровня лингвистических характеристик, комбинаций двух или трех уровней и при помощи BERT-моделей. Результаты мультиклассификации представлены в Табл. 1 и 2, каждая посвящена одному классификатору. В первом столбце указывается набор лингвистических характеристик или тип эмбеддингов, для которых проходила классификация. В остальных трех столбцах представлены значения стандартных метрик качества. Полужирным шрифтом выделены лучшие значения F-меры среди наборов лингвистических характеристик и языковых моделей.

Сравнение моделей показывает, что эмбеддинги лучше справляются с разделением текстов на жанры: качество доходит до 99 % F-меры. Лучшие метрики качества демонстрируют модели DeepPavlov/rubert-base-cased и DeepPavlov/rubert-base-cased-conversational: 98.3 и 99.2 % F-меры соответственно. Это значит, что они делают очень маленькое число ошибок

Табл. 1. Качество классификации Интернет-текстов на восемь жанров при помощи лингвистических характеристик и эмбеддингов, классификатор — персептрон

| **Модель** | **Точность** | **Полнота** | **F-мера** |
|---|---|---|---|
| Уровень символов | 89.7 | 89.7 | 89.7 |
| Уровень ритма | 75.5 | 76.1 | 75.8 |
| Уровень структуры | 77.0 | 77.2 | 77.1 |
| Уровни символов и структуры | 91.4 | 91.3 | **91.4** |
| Уровни символов и ритма | 91.1 | 91.1 | **91.1** |
| Уровни структуры и ритма | 81.6 | 81.1 | 81.3 |
| Уровни символов, структуры и ритма | 92.5 | 92.4 | **92.4** |
| cointegrated/rubert-tiny | 91.0 | 92.4 | 91.2 |
| cointegrated/rubert-tiny2 | 93.8 | 95.4 | 94.3 |
| DeepPavlov/rubert-base-cased | 98.5 | 98.2 | **98.3** |
| DeepPavlov/rubert-base-cased-conversational | 99.3 | 99.1 | **99.2** |
| bert-base-multilingual-cased | 93.5 | 94.5 | 93.9 |

Табл. 2. Качество классификации Интернет-текстов на восемь жанров при помощи лингвистических характеристик, классификатор — LSTM

| **Модель** | **Точность** | **Полнота** | **F-мера** |
|---|---|---|---|
| Уровень символов | 87.1 | 87.0 | 87.0 |
| Уровень ритма | 72.8 | 73.7 | 73.3 |
| Уровень структуры | 74.0 | 74.8 | 74.4 |
| Уровни символов и структуры | 89.0 | 89.2 | **89.1** |
| Уровни символов и ритма | 90.3 | 90.4 | **90.4** |
| Уровни структуры и ритма | 79.5 | 80.1 | 79.8 |
| Уровни символов, структуры и ритма | 90.5 | 90.5 | **90.5** |

при определении жанра. Остальные эмбеддинги показывают меньшее качество, но все же оно оказывается выше 91 %. Причина в том, что rubert-tiny и rubert-tiny2 изначально меньшие по размеру, чем остальные языковые модели, а bert-base-multilingual-cased мультиязычна.

Таким образом, эксперименты показывают, что большая специализированная для русского языка модель обеспечивает существенно лучшие результаты, чем другие языковые модели. Лингвистические характеристики демонстрируют меньшее качество классификации: не более 92.4 % F-меры для случая с комбинацией всех характеристик. Тем не менее оно находится на достаточно высоком уровне. Наилучший вклад в качество результата вносит уровень символов: 89.7 % F-меры — это самый высокий результат для одного уровня характеристик, а комбинации других уровней с символьным дают значимый прирост качества до 91-92 % F-меры. Характеристики уровня ритма показывают 75.8 % F-меры, это наименьший результат, но при этом их количество тоже меньшее, чем характеристик других уровней.

Усложнение классификатора с персептрона на LSTM не дает улучшения результата. Для BERT-моделей значения метрик практически идентичны прежним, а для лингвистических характеристик оказываются меньше на 2-3 % (Табл.2), что может быть объяснено погрешностью.

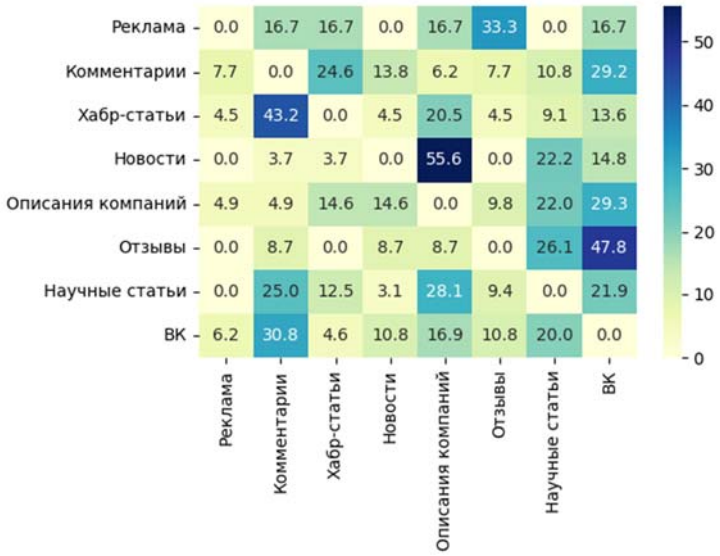


Рис. 1. Ошибки классификации Интернет-текстов на восемь жанров при помощи трех уровней лингвистических характеристик и классификатора LSTM

В целом можно сделать вывод, что для решения задачи первую роль играет модель текста, а нейросетевой классификатор можно брать достаточно простой.

Для классификации с помощью комбинации трех уровней лингвистических характеристик была построена матрица ошибок (Рис. 1). В матрице в каждой строке указываются проценты от суммарного количества ошибок. Наибольший процент ошибок возникает, когда новости принимаются за описания компаний (55.6 %) и когда отзывы принимаются за посты ВК (47.8 %).

Реклама классифицируется лучше всего: ошибки с ее текстами совершаются более-менее равномерно, и за нее принимаются наименьшие доли других жанров. Больше всего тексты ошибочно принимаются за комментарии, описания компаний и посты ВК. В этих столбцах наибольшие доли ошибок.

Для того, чтобы подробнее разобраться с качеством классификации при помощи лингвистических характеристик, можно визуализировать их и интерпретировать результаты.

## 4. Обсуждение лингвистических характеристик текстов

Рассмотрим представленные тексты с точки зрения четырех уровней анализа. Что касается частоты встречаемости букв в текстах представленных жанров, то интересным является наблюдение о том, что уровень встречаемости гласных и согласных во всех жанрах примерно одинаковый. Это является свидетельством того, что для русского языка в целом эта частотность является постоянной. В языке существует определенный набор букв (звуков), которые являются наиболее частотными, что и подтверждают данные, полученные на материале русского языка. Так, для анализируемых текстов характерны такие гласные, как а, о, е, и. Среди согласных наиболее частотными являются т, д, ж, з, н.

Примечательным является то, что именно реклама и комментарии сосредотачивают в себе наибольшее количество повторений практически всех букв алфавита. Исключение составляют буквы ё, ж, ф, ш, щ, ъ, э, ю для рекламы (они отсутствуют совсем), а для комментариев – ё, ф, ъ. Вероятно, неупотребление данных букв в рекламе связано с эвфонией, благозвучием, которое необходимо при аудиальном восприятии текста рекламы. Соотношение количества букв в анализируемых текстах отражено на Рис. 2.

Знаки препинания и различные символы (!,#, %, $, &, (, ), ; @ _ ‘ № -) также представлены в разной степени для разных анализируемых текстов, в частности, знак восклицания является наиболее частотным в принципе и характерен для рекламных текстов. Наиболее распространенными являются тире, запятая, раскрытие и закрытие скобок, для всех практически жанров, за исключением рекламы, комментариев и статей с портала Хабр.

Рис. 2. Частотность употребления букв в текстах по жанрам

Средняя длина предложения в словах является наибольшей для научных статей с портала Хабр и из научных журналов. Она достигает 15 слов в предложении для статей с портала Хабр и 20 для научных статей из журналов. Наименьшими по объему являются предложения в рекламных текстах и характерный объем предложения не более 5 слов. Этот факт очевиден и объясняется более сложными структурами предложения в научных текстах, наличием сложносочиненных и сложноподчиненных предложений, в отличие от рекламы, которая часто отличается безличными конструкциями.

Проведенные исследования относительно частотности ритмических средств в художественных текстах показали, что наиболее активным средством является диакопа. Это по-разному может объясняться с лингвистической точки зрения, в частности относительной свободой повторяющихся элементов внутри самого средства, нефиксированный порядок повторяющихся элементов. Кроме того, употребление диакопы в других языках, например, английском, французском или испанском, обусловлено наиболее частотным употреблением однокоренных слов рядом, словообразовательной деривацией. В анализируемых русских текстах представленных жанров диакопа также является наиболее распространенной. Однако есть исключение: это рекламные тексты, в которых диакопа практически отсутствует. Уровень употребления диакопы колеблется для разных жанров, в частности, он является наименьшим для отзывов на фильмы с сайта Кинопоиск и наибольшим для статей с платформы Хабр (89.45 %) по отношению к другим ритмическим средствам, далее уступают по частотности научные статьи из журналов (84.80 %) и новостные тексты (80.22 %). Что касается рекламных текстов, то наиболее частотным средством создания ритма является повторение вопросительных предложений (85.35 %). Соотношение ритмических средств можно видеть на Рис. 3. Менее активной, но также значимой является апозиопеза (7.84 %). Таким образом, для рекламных текстов с ритмической точки зрения наиболее рабочими являются структуры с эмотивной значимостью, которые выделяются графически пунктуационными знаками.

Что касается частей речи, которые представляют те или иные ритмические средства, то естественным является преобладание существительных для всех жанров, однако для рекламы в наибольшей степени (0,55). Наименьшее количество употреблено в обзорах и комментариях (0,35). Далее наиболее активными являются прилагательные, что особенно характерно для рекламы (0,24) и обзора компаний (0,22). На третьем месте по частотности– глагол в особенности в комментариях и новостях (0,2).

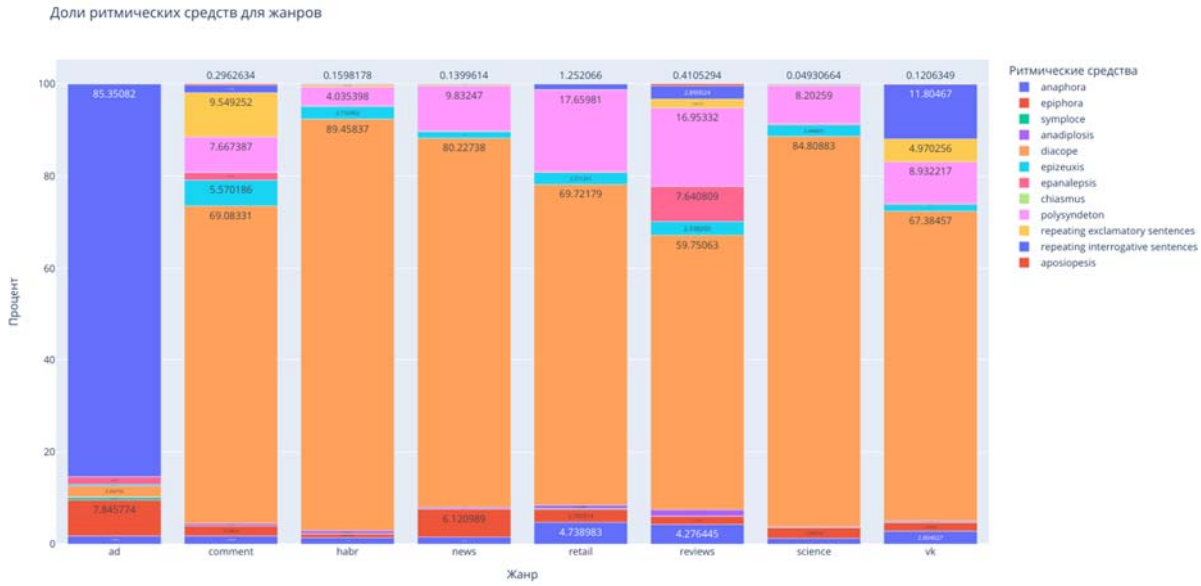


Рис. 3. Соотношение ритмических средств в текстах анализируемых жанров

Характеристики уровня структуры представляют собой отслеживание сочетаемости частей речи в анализируемых текстах. Исследование показало, что наиболее активным является сочетание имен собственных и оно характерно в наибольшей степени для рекламных текстов (0,19). Основным для всех анализируемых текстов является сочетание прилагательного с существительным, однако частотность также варьируется в зависимости от жанра. Наиболее активны данные сочетания в описаниях компаний (0,12), далее по убывающей – научные тексты из журналов (0,10), статьи с платформы Хабр (0,088), отзывы на фильмы (0,080), реклама (0,079), новости (0,076), посты в ВК (0,075). Следующая категория сочетаний, которая является такой же частотной – это существительное + существительное. Здесь также в зависимости от жанра есть некоторые различия по частотности. На первом месте – описания компаний (0,099), далее–научные статьи в журналах (0,097), новости (0, 084), научные статьи с платформы Хабр (0,081). Наименьшее количество таких сочетаний у рекламных текстов. Следует отметить, что рекламные тексты отличаются также наименьшим разнообразием сочетаний различных частей речи. Соотношение сочетаний частей речи можно видеть на Рис. 4.

Следует отметить некоторые особенности, которые касаются употребления наиболее распространенного ритмического средств – диакопы. В связи с тем, что оно является наиболее активным в языке ввиду некоторых причин, которые в первую очередь обусловлены подвижностью повторяющихся элементов диакопы по отношению к основному элементу, довольно странным выглядит ее очень редкое употребление или неупотребление в научных текстах с платформы Хабр, в рекламе и комментариях. Примерно равное соотношение диакопы в новостных текстах и обзорах фильмов, а также соотносятся по количеству описания компаний и комментарии в ВК. Вероятно, низкая частотность повторов таких текстов обусловлена спецификой жанра, а именно: для рекламы как наиболее короткого текста по объему, повтор является неэкономным средством, для научных текстов с платформы Хабр отсутствие повторов обусловлено тематикой повествования и стремлением наиболее кратко и сжато передать информацию. Что касается комментариев, то при их написании скорее всего автор использует короткие синтаксические конструкции, нераспространенные предложения, в которых наибольшую важность имеет эмотивная составляющая.

Примечательным также является употребление анафоры, характерное исключительно для обзоров фильмов. Это связано прежде всего с тем, что в качестве анафоры выступают повторы местоимений, выполняющих функцию подлежащего. Также следует отметить употребление полисиндетона, характерное только для трех жанров: обзора, описания и научных текстов из журналов.

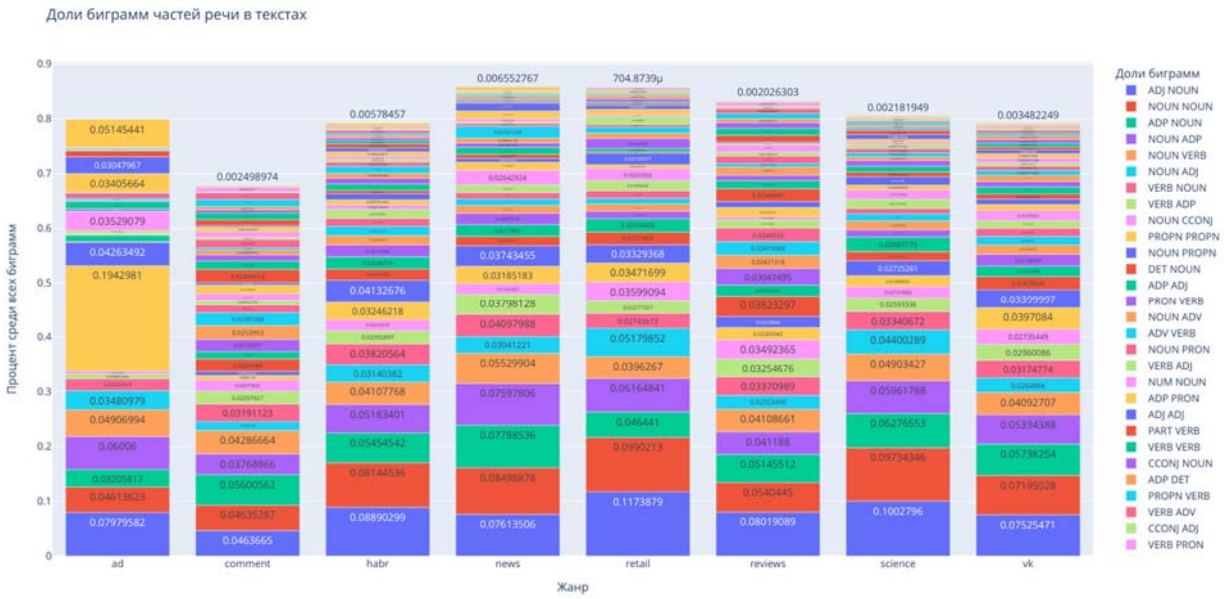


Рис. 4. Сочетаемостные характеристики для анализируемых текстов

## Заключение

Проведено сравнение пяти языковых BERT-моделей и трех уровней лингвистических характеристик для задачи жанровой классификации русскоязычных Интернет-текстов. Мультиклассификация на восемь жанров при помощи нейросетевых алгоритмов показала, что все типы моделей текста успешно справляются с определением жанра: качество достигает не менее 90 % точности, полноты и F-меры для комбинации всех лингвистических характеристик или любой модели на основе эмбеддингов. Языковая модель rubert-base-cased-conversational, которая была предобучена на самом большом корпусе данных для русского языка, продемонстрировала наилучшие значения метрик: 99 % точности, полноты и F-меры.

Результаты, показанные моделями на основе лингвистических характеристик интерпретируются лучше, чем результаты эмбеддингов, с лингвистической точки зрения. Визуализация числовых лингвистических характеристик и их последующий анализ демонстрируют, что у каждого жанра имеются собственные стилевые особенности в частотах букв и знаков препинания, ритмическом составе текстов, сочетаемости частей речи и т. п. Таким образом, у жанров выявляются конкрентные отличительные черты, которые и обеспечивают высокое качество классификации.

Для продолжения данных исследований выглядят перспективными два направления. Во-первых, можно применить лингвистические характеристики для других малоисследованных задач классификации по разнообразным параметрам русскоязычных текстов. Во-вторых, стоит аналогичным образом проанализировать большее число лингвистических характеристик, чтобы глубже исследовать стиль текстов и повысить качество классификации вместе с интерпретируемостью результатов.

## Литература

1. Бахтин М. М., Капанадзе Л. А. Теория речевых жанров в контексте лингвистического градоведения //Социолингвистика: языковой облик современного города 2-е изд., испр. и доп. Учебник и практикум для вузов. М.: Издательство Юрайт, 2022. С. 45–52.
2. Kuzman T., Rupnik P., Ljubešić N. The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild //Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022. P. 1584–1594.
3. Галичкина Е. Н. Типология речевых жанров сетевой компьютерной коммуникации //Известия Волгоградского государственного педагогического университета. 2019. №. 2 (135). С. 97–100.
4. Тарабарина Ю. А. Речевые жанры как профессиональная основа содержания иноязычного обучения студентов градостроительного направления подготовки //Современное педагогическое образование. 2023. №. 1. С. 176–183.
5. Кузнецов А. В., Писанова Т. В. Классификация журналистских жанров в Испании и России: независимые

пути к единому подходу //Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2020. №. 7 (836). С. 102–112.
6. Li Q., Peng H., Li J., Xia C., Yang R., Sun L., Yu P. & He L. A survey on text classification: From traditional to deep learning //ACM Transactions on Intelligent Systems and Technology (TIST). 2022. V. 13. No 2. P. 1–41.
7. Onan A. An ensemble scheme based on language function analysis and feature engineering for text genre classification //Journal of Information Science. 2018. V. 44. No 1. P. 28–47.
8. Al-Yahya M. Stylometric analysis of classical Arabic texts for genre detection //The Electronic Library. 2018. V. 36. No 5. P. 842–855.
9. Батраева И. А., Нарцев А. Д., Лезгян А. С. Использование анализа семантической близости слов при решении задачи определения жанровой принадлежности текстов методами глубокого обучения //Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2020. №. 50. С. 14–22.
10. Le Mens G., Kovács B., Hannan M. T., & Pros G. Using machine learning to uncover the semantics of concepts: how well do typicality measures extracted from a BERT text classifier match human judgments of genre typicality? //Sociological Science. 2023. V. 10. No 3. P.82–117.
11. Sharoff S. Functional text dimensions for the annotation of web corpora //Corpora. 2018. V. 13. No 1. P. 65–95.
12. Madjarov G., Vidulin V., Dimitrovski I., & Kocev D. Web genre classification with methods for structured output prediction //Information Sciences. 2019. V. 503. P. 551–573.
13. Rasheed A., Umar A. I., Shirazi S. H., Khan Z., & Shahzad M. Cover-based multiple book genre recognition using an improved multimodal network //International Journal on Document Analysis and Recognition (IJDAR). 2023. V. 26. No 1. P. 65–88.
14. Da Cunha I., Montané M. A. A corpus-based analysis of textual genres in the administration domain // Discourse Studies. 2020. V. 22. No 1. P. 3–31.
15. Горбич Л. Г., Живодеров А. А. Использование статистических индексов для различения научных и научно-популярных текстов на примере трудов А. Е. Ферсмана //Программные продукты и системы. 2020. V. 33. No 4. P. 720–725.
16. Corrêa Jr E. A., Marinho V. Q., Amancio D. R. Semantic flow in language networks discriminates texts by genre and publication date //Physica A: Statistical Mechanics and its Applications. 2020. V. 557. P. 124895.
17. Lagutina K., Poletaev A., Lagutina N., Boychuk E., Paramonov I. Automatic extraction of rhythm figures and analysis of their dynamics in prose of 19th-21st centuries. // Proceedings of the 26th Conference of Open Innovations Association FRUCT, Yaroslavl, Russia, 20-24 April 2020. IEEE. P. 247-255.
18. Lagutina K., Lagutina N., Boychuk E., Larionov V., Paramonov I. Authorship verification of literary texts with rhythm features. // Proceedings of the 28th Conference of Open Innovations Association FRUCT, Moscow, Russia, 27-29 January 2021. IEEE. P. 240-251.
19. Kuratov, Y., Arkhipov, M. Adaptation of deep bidirectional multilingual transformers for Russian language. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”, Moscow, May 29—June 1, 2019. P. 333-339.
20. Kenton J. D. M. W. C., Toutanova L. K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. // Proceedings of NAACL-HLT, Minneapolis, Minnesota, June 2 - June 7, 2019. P. 4171-4186.

**Лагутина Ксения Владимировна.** Кандидат технических наук. Старший преподаватель кафедры вычислительных и программных систем. Ярославский государственный университет им. П.Г. Демидова. Области исследований: компьютерная лингвистика, искусственный интеллект, моделирование текстов. E-mail: lagutinakv@mail.ru (ответственный за переписку).

**Бойчук Елена Игоревна.** Доктор филологических наук. Профессор кафедры романских языков. Ярославский государственный педагогический университет им. К.Д. Ушинского. Области исследований: литературоведение, лингвистика текста, компьютерная лингвистика. E-mail: elena-boychouk@rambler.ru

**Лагутина Надежда Станиславовна.** Кандидат физико-математических наук. Доцент кафедры вычислительных и программных систем. Ярославский государственный университет им. П.Г. Демидова. Области исследований: компьютерная лингвистика, искусственный интеллект, моделирование текстов. E-mail: lagutinans@rambler.ru

## Automatic Classification of Russian-Language Internet Texts by Genre

K. V. Lagutina [1], E. I. Boychuk [2], N. S. Lagutina [1]

[1] P.G. Demidov Yaroslavl State University, Yaroslavl, Russia
[2] Yaroslavl State Pedagogical University named after K. D. Ushinsky, Yaroslavl, Russia

**Abstract.** This article is devoted to the use of modern language models based on BERT and models based on three types of text linguistic features for automatic determination of the text genre, as well as a comparative analysis of these models from the points of view of computer and classical linguistics. The authors have collected their own corpus of Russian-language Internet texts in eight genres: VKon-

takte posts, comments, articles from the Habr portal, retail descriptions, news, scientific articles, advertising, movie reviews from the Kinopoisk website. Each text was represented as a vector of numerical features using each of the selected models: five BERT variations and linguistic features of character, structure and rhythm levels. Vectors based on linguistic features were also concatenated for two or three levels to obtain additional text models. Next, the vectors were classified into eight genres using neural network classifiers, a perceptron and LSTM. The results of the classification showed that BERT models achieved a high quality of genre detection: up to 91-99% of precision, recall, and F-measure. The combination of linguistic features made it possible to obtain the F-measure about 90%. An analysis of the classification results and text models from a linguistic point of view revealed the features of individual genres and possible reasons for both high results and classification errors.
**Keywords:** stylometry, natural language processing, rhythm features, genres, text classification, BERT.

**DOI** 10.14357/20718594230410 **EDN** BAKBAF

## References

1. Bahtin, M. M., Kapanadze, L. A. Teoriya rechevyh zhanrov v kontekste lingvisticheskogo gradovedeniya [The theory of speech genres in the context of linguistic urban studies] //Sociolingvistika: yazykovoj oblik sovremennogo goroda 2-e izd., ispr. i dop. Uchebnik i praktikum dlya vuzov. [Sociolinguistics: the linguistic appearance of the modern city, 2nd ed., corr. and add. Textbook and workshop for universities]. Moscow: Izdatel'stvo Yurajt, 2022. P. 45–52.
2. Kuzman T., Rupnik P., Ljubešić N. The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild //Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022. P. 1584–1594.
3. Galichkina E. N. Tipologiya rechevyh zhanrov setevoj komp'yuternoj kommunikacii [Typology of speech genres of network computer communication] //Izvestiya Volgogradskogo gosudarstvennogo pedagogicheskogo universiteta [News of the Volgograd State Pedagogical University]. 2019. No 2 (135). P. 97–100.
4. Tarabarina Y. A. Rechevye zhanry kak professional'naya osnova soderzhaniya inoyazychnogo obucheniya studentov gradostroitel'nogo napravleniya podgotovki [Speech genres as a professional content element of foreign language teaching of urban planning students] //Sovremennoe pedagogicheskoe obrazovanie [Modern Pedagogical Education] 2023. No 1. P. 176–183.
5. Kuznetsov A. V., Pisanov T. V. Klassifikaciya zhurnalistskih zhanrov v Ispanii i Rossii: nezavisimye puti k edinomu podhodu [Classification of journalistic genres in Russia and Spain: independent solutions and consistent approach] //Vestnik Moskovskogo gosudarstvennogo lingvisticheskogo universiteta. Gumanitarnye nauki [Vestnik of Moscow State Linguistic University. Social sciences] 2020. No 7 (836). P. 102–112.
6. Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. & He, L. A survey on text classification: From traditional to deep learning //ACM Transactions on Intelligent Systems and Technology (TIST). 2022. V. 13. No 2. P. 1–41.
7. Onan A. An ensemble scheme based on language function analysis and feature engineering for text genre classification //Journal of Information Science. 2018. V. 44. No 1. P. 28–47.
8. Al-Yahya M. Stylometric analysis of classical Arabic texts for genre detection //The Electronic Library. 2018. V. 36. No 5. P. 842–855.
9. Batraeva I.A., Nartsev A.D., Lezgyan A.S. Ispol'zovanie analiza semanticheskoj blizosti slov pri reshenii zadachi opredeleniya zhanrovoj prinadlezhnosti tekstov metodami glubokogo obucheniya [Using the analysis of semantic proximity of words in solving the problem of determining the genre of texts within deep learning] //Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika [Tomsk State University Journal of Control and Computer Science] 2020. No 50. P. 14–22.
10. Le Mens, G., Kovács, B., Hannan, M. T., & Pros, G. Using machine learning to uncover the semantics of concepts: how well do typicality measures extracted from a BERT text classifier match human judgments of genre typicality? //Sociological Science. 2023. V. 10. No 3. P.82–117.
11. Sharoff S. Functional text dimensions for the annotation of web corpora //Corpora. 2018. V. 13. No 1. P. 65–95.
12. Madjarov, G., Vidulin, V., Dimitrovski, I., & Kocev, D. Web genre classification with methods for structured output prediction //Information Sciences. 2019. V. 503. P. 551–573.
13. Rasheed, A., Umar, A. I., Shirazi, S. H., Khan, Z., & Shahzad, M. Cover-based multiple book genre recognition using an improved multimodal network //International Journal on Document Analysis and Recognition (IJDAR). 2023. V. 26. No 1. P. 65–88.
14. Da Cunha I., Montané M. A. A corpus-based analysis of textual genres in the administration domain // Discourse Studies. 2020. V. 22. No 1. P. 3–31.
15. Gorbich L.G., Zhivoderov A.A. Ispol'zovanie statisticheskih indeksov dlya razlicheniya nauchnyh i nauchno-popularnyh tekstov na primere trudov A. E. Fersmana [Using statistical indexes to distinguish between scientific and popular science texts on the example of the works of A.E. Fersman.] // Programmnye produkty i sistemy. [Software & Systems. 2020.] V. 33. No 4. P. 720–725.
16. Corrêa Jr E. A., Marinho V. Q., Amancio D. R. Semantic flow in language networks discriminates texts by genre and publication date //Physica A: Statistical Mechanics and its Applications. 2020. V. 557. P. 124895.
17. Lagutina, K., Poletaev, A., Lagutina, N., Boychuk, E., Paramonov, I. Automatic extraction of rhythm figures and analysis of their dynamics in prose of 19th-21st centuries. // Proceedings of the 26th Conference of Open Innovations Association FRUCT, Yaroslavl, Russia, 20-24 April 2020. IEEE. P. 247-255.

18. Lagutina, K., Lagutina, N., Boychuk, E., Larionov, V., Paramonov, I. Authorship verification of literary texts with rhythm features. // Proceedings of the 28th Conference of Open Innovations Association FRUCT, Moscow, Russia, 27-29 January 2021. IEEE. P. 240-251.
19. Kuratov, Y., Arkhipov, M. Adaptation of deep bidirectional multilingual transformers for Russian language. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019", Moscow, May 29—June 1, 2019. P. 333-339.
20. Kenton J. D. M. W. C., Toutanova L. K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. // Proceedings of NAACL-HLT, Minneapolis, Minnesota, June 2 - June 7, 2019. P. 4171-4186.

**Lagutina Ksenia V.** Candidate of technical sciences. Senior lecturer, Department of Computing and Program Systems, P.G. Demidov Yaroslavl State University. Research areas: computational linguistics, artificial intelligence, text modeling. E-mail: lagutinakv@mail.ru

**Boychuk Elena I.** Doctor of philological sciences, ɜrofessor, professor of the Department of Romance Languages, Yaroslavl State Pedagogical University named after K.D.Ushinsky. Research areas: literary studies, text linguistics, computational linguistics. E-mail: elena-boychouk@rambler.ru

**Lagutina Nadezhda S.** Candidate of physical and mathematical sciences, docent. Docent of the Department of Computing and Program Systems, P.G. Demidov Yaroslavl State University. Research areas: computational linguistics, artificial intelligence, text modeling. E-mail: lagutinans@rambler.ru