MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 28, NO. 3, 2021
journal homepage: www.mais-journal.ru

THEORY OF DATA

Text Classification by Genre Based on Rhythm Features

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

DOI: 10.18255/1818-1015-2021-3-280-291

[1]P. G. Demidov Yaroslavl State University, 14 Sovetskaya str., Yaroslavl 150003, Russia.
[2]Yaroslavl State Pedagogical University named after K. D.Ushinsky, 108/1 Respublikanskaya str., Yaroslavl 150000, Russia.

MSC2020: 68T50
Research article
Full text in Russian

Received August 20, 2021
After revision August 30, 2021
Accepted September 1, 2021

The article is devoted to the analysis of the rhythm of texts of different genres: fiction novels, advertisements, scientific articles, reviews, tweets, and political articles. The authors identified lexico-grammatical figures in the texts: anaphora, epiphora, diacope, aposiopesis, etc., that are markers of the text rhythm. On their basis, statistical features were calculated that describe quantitatively and structurally these rhythm features.
The resulting text model was visualized for statistical analysis using boxplots and heat maps that showed differences in the rhythm of texts of different genres. The boxplots showed that almost all genres differ from each other in terms of the overall density of rhythm features. Heatmaps showed different rhythm patterns across genres. Further, the rhythm features were successfully used to classify texts into six genres. The classification was carried out in two ways: a binary classification for each genre in order to separate a particular genre from the rest genres, and a multi-class classification of the text corpus into six genres at once. Two text corpora in English and Russian were used for the experiments. Each corpus contains 100 fiction novels, scientific articles, advertisements and tweets, 50 reviews and political articles, i.e. a total of 500 texts. The high quality of the classification with neural networks showed that rhythm features are a good marker for most genres, especially fiction. The experiments were carried out using the ProseRhythmDetector software tool for Russian and English languages. Text corpora contains 300 texts for each language.

Keywords: stylometry; natural language processing; rhythm features; genres; text classification

INFORMATION ABOUT THE AUTHORS

Ksenia Vladimirovna Lagutina correspondence author	orcid.org/0000-0002-1742-3240. E-mail: lagutinakv@mail.ru postgraduate student.
Nadezhda Stanislavovna Lagutina	orcid.org/0000-0002-6137-8643. E-mail: lagutinans@rambler.ru PhD, associate professor.
Elena Igorevna Boychuk	orcid.org/0000-0001-6600-2971. E-mail: elena-boychouk@rambler.ru PhD, associate professor.

Funding: The reported study was funded by RFBR, project number 19-07-00243.

For citation: K. V. Lagutina, N. S. Lagutina, and E. I. Boychuk, “Text Classification by Genre Based on Rhythm Features”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 280-291, 2021.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 28, № 3, 2021
сайт журнала: www.mais-journal.ru

THEORY OF DATA

Классификация текстов по жанрам на основе ритмических характеристик

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

DOI: 10.18255/1818-1015-2021-3-280-291

[1]Ярославский государственный университет им. П. Г. Демидова, ул. Советская, д. 14, г. Ярославль, 150003 Россия.
[2]Ярославский государственный педагогический университет им. К. Д. Ушинского, ул. Республиканская, д. 108/1, г. Ярославль, 150000 Россия.

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 20 августа 2021 г.
После доработки 30 августа 2021 г.
Принята к публикации 1 сентября 2021 г.

Статья посвящена анализу ритма текстов различных жанров: художественных романов, рекламы, научных статей, отзывов, твитов и политических статей. Авторы выделили в текстах лексико-грамматические средства: анафору, эпифору, диакопу, апозиопезу и т. п., которые являются маркерами ритма текста. На их основе были подсчитаны статистические характеристики, описывающие количественно и структурно данные ритмические средства.
Полученная модель текста была визуализирована для статистического анализа с помощью диаграмм размаха и тепловых карт, которые показали отличия в ритме текстов различных жанров. Диаграммы размаха показали, что практически все жанры отличаются друг от друга по общей плотности ритмических характеристик. Тепловые карты показали различную структуру ритма у жанров.
Далее ритмические характеристики успешно использовались для классификации текстов по шести жанрам. Высокое качество классификации показало, что ритмические характеристики являются хорошим маркером для большинства жанров, в особенности для художественной литературы. Эксперименты проводились с помощью программного инструмента ProseRhythmDetector для русского и английского языков. Корпуса текстов содержат по 300 текстов для каждого языка.

Ключевые слова: стилометрия; обработка естественного языка; ритмические характеристики; жанры; классификация текстов

ИНФОРМАЦИЯ ОБ АВТОРАХ

Ксения Владимировна Лагутина
автор для корреспонденции
orcid.org/0000-0002-1742-3240. E-mail: lagutinakv@mail.ru
аспирант.

Надежда Станиславовна Лагутина
orcid.org/0000-0002-6137-8643. E-mail: lagutinans@rambler.ru
канд. физ.-мат. наук, доцент.

Елена Игоревна Бойчук
orcid.org/0000-0001-6600-2971. E-mail: elena-boychouk@rambler.ru
доктор фил. наук, доцент.

Финансирование: Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00243.

Для цитирования: K. V. Lagutina, N. S. Lagutina, and E. I. Boychuk, “Text Classification by Genre Based on Rhythm Features”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 280-291, 2021.

Введение

Развитие методов автоматической обработки естественного языка позволяет исследователям ставить и решать сложные задачи на уровне дискурса, характеризующего смысловую организацию текстов. Идентификация жанра текста, как его функциональной характеристики, позволяет сделать акцент на иерархической природе текста в противоположность моделированию языка в виде плоской последовательности или неупорядоченного множества слов или букв [1].

Определение жанра текста является важной задачей как языкознания, так и создания корпусов текстов, без которых невозможно решение задач компьютерной лингвистики [2]. Этот факт отмечают и российские учёные [3]. Классификация текстов по жанру используется в исследованиях классической литературы и литературного языка [4], оказывается актуальной для анализа и извлечения информации из интернет-ресурсов [5], играет существенную роль для качественного машинного перевода текстов [6].

Чаще всего исследователи рассматривают функциональные стили текста, соответствующие научному, художественному, официально-деловому и публицистическому жанрам. Кроме того, учёные решают более специфические задачи по определению жанров художественной литературы [4, 7] или web-страниц [5]. В этих задачах методы решения можно условно отнести к одному из двух основных подходов: статистическому анализу стилометрических характеристик текста [8] и классификации на основе машинного обучения [7]. Однако в обоих случаях учёные подчёркивают, что самая важная часть работы связана с отбором релевантных параметров текста и исследованием роли различных типов характеристик для автоматической классификации по жанрам с конечной целью выявления наиболее эффективных по качеству.

В своих работах авторы данного исследования предложили комплекс высокоуровневых параметров художественного текста, основанных на фигурах речи образуемых повторением слов и словосочетаний [9, 10]. Эти параметры описывают ритм текста, который позволяет выявить уникальный авторский стиль и успешно классифицировать тексты по времени и авторам. Поэтому в данной работе была поставлена задача проанализировать влияние ритмических характеристик текстов, как нового типа стилометрических параметров, на определение жанра. Для этого выполняется статистический анализ ритмических характеристик и классификация текстов по жанрам: художественные романы, научные статьи, политические статьи, рекламные статьи, отзывы, твиты.

Обзор смежных работ

Наиболее распространённый подход к классификации текстов, в том числе по жанрам, основан на подборе и адаптации методов машинного обучения. В статье [11] проведён сравнительный анализ пяти различных базовых алгоритмов классификации (наивный байесовский классификатор, метод опорных векторов, логистическая регрессия, метод k-ближайших соседей и алгоритм «случайный лес») в сочетании с методами ансамблевого обучения (такими как Boosting, Bagging и Random Subspace). На основе эмпирического анализа представлена схема классификации ансамблей, которая объединяет Random Subspace и случайного леса с четырьмя типами признаков (признаки, используемые в атрибуции авторства, n-граммы символов, n-граммы части речи и частота редких слов). Для корпуса текстов LFA наивысшая средняя прогностическая эффективность, полученная по предложенной схеме, составляет 94,43 %. Однако немаловажную часть этой работы занимает сравнение и подбор подходящих характеристик текста.

На анализе подходящих методов машинного обучения базируются многие исследователи национальных языков. Учёные провели классификацию арабских текстов по одному из четырех жанров: реклама, новости, личные и научные документы [12]. Они использовали те же классификаторы, что и в исследовании [11]. Качество определения жанра оказалось очень высоким: F-мера более 90 %. Авторы отметили, что подход к вычислению характеристик текста на уровне «мешок слов»

дал низкую эффективность, поэтому они использовали более сложные параметры стиля текста. Интересно, что предсказание текстов личного и научного жанра более точно, чем прогнозирование рекламы и новостей. Самый лучший результат определения жанра дали деревья решений, но они были построены на основе статистического анализа конкретного корпуса текстов, поэтому высокое качество ожидаемо, но не носит обобщающий характер.

Более универсальный подход к анализу текстов с точки зрения определения их жанровой принадлежности основан на применении сверточных нейронных сетей [13]. Авторами разработана архитектура сверточной нейронной сети с использованием векторного представления слов на основе модели word2vec. Эффективность работы построенной модели проверена для пяти жанров: история, детективы, детская литература, поэзия, фантастика. Точность классификации составила 78,64 %. В качестве обучающих данных был выбран корпус русскоязычных текстов Максима Мошкова.

Эксперименты по определению стилей и жанров поэтических текстов (ода, элегия, баллада, эпиграмма и т. д.) с использованием корпуса текстов лицейской лирики А. С. Пушкина описаны в работе [14]. Авторы выбирали наиболее точный алгоритм классификации с использованием известных приемов ансамблирования базовых алгоритмов, таких как взвешенное голосование, бустинг и стекинг. В качестве характеристических признаков текстов использовались униграммы, биграммы и триграммы слов. Было установлено, что даже с помощью простых классификаторов на основе этих лексических признаков можно получить хороший результат решения задачи. Лучшая точность оказалась у многослойного персептрона 99 % при использовании в качестве характеристик текстов триграмм слов. Авторы отметили важность применения такого анализа поэзии для эксперта-лингвиста.

Хотя стандартные алгоритмы классификации показывают высокое качество определения жанров текста, все авторы описанных выше работ в большей или меньшей степени обращают внимание на выбор характеристик текста. Обобщая их результаты можно обратить внимание на то, что использование более сложных стилометрических параметров, таких как n-граммы слов, даёт лучшее качество решения задачи. К такому же выводу приходят авторы работы [5], которые проанализировали вклад различных типов характеристик в решение проблемы определения жанра в компьютерной лингвистике. Аналогичная задача решается в исследовании [8]. В нём отмечается, что ключевую роль играют синтаксические особенности текста, влияние которых, различается в зависимости от жанра.

Исследователи русского языка, обращаясь к проблеме определения жанра текста, выявляют сложные лексические особенности стиля, отличающие разные жанры. В работе [15] выдвигается гипотеза о том, что коэффициенты соотношения частот семантически противопоставленных предлогов русского языка могут указывать на стилевую принадлежность текстов. Материалом для экспериментов послужили корпусы текстов разных функциональных стилей и разной тематики: общий, художественный, публицистический, нехудожественный, устный из Национального корпуса русского языка (НКРЯ), корпусов Araneum Russicum Russicum и Araneum Russicum Externum, корпуса текстов из социальных сетей Facebook и Twitter, корпуса художественных текстов с сайта Либрусек. Эксперименты подтвердили значимость ряда коэффициентов в диагностике стиля и типа текстов. Так же была получена информация о семантическом наполнении предложных конструкций, которая важна для определения стилевых и жанровых характеристик текстов.

Другие учёные предлагают вместе со стилометрическими параметрами использовать дополнительные статистические числовые характеристики. Авторы работы [16] классифицируют по жанрам фрагменты научных и научно-популярных текстов академика Александра Евгеньевича Ферсмана, выдающегося ученого и популяризатора науки. Они используют характеристики уровня символов, индексы на основе частоты гласных букв, частот биграмм и триграмм символов, индексы

энтропии и сжимаемости текстов. Однако выбранный для экспериментов корпус очень мал, всего 44 фрагмента, что скорее всего обусловлено трудоёмкостью сбора и качественной разметки таких ресурсов.

В статье [17] проводились эксперименты на материале русскоязычных корпусов текстов, принадлежащих четырём функциональным стилям: научному, художественному, официально-деловому и публицистическому. В качестве характеристик текста использовались частоты морфологических параметров, частоты pos-тегов, некоторые биграммы, длины слов и предложений. Кроме того, рассчитывался комбинированный параметр β, отражающий соотношение динамичности и статичности текстов коллекции [18]. На основе статистического анализа была определена система правил для классификации текстов по жанрам и получены высокие результаты. F–мера составила 99 % для художественного и делового стилей, 83 % для научного, 70 % для публицистического. Результаты исследования аналогичны прогнозированию жанра текста на основе дерева решений [12], однако объём использованного корпуса текстов значительно меньше, по 65 текстов для каждого стиля.

Таким образом, анализ использования сложных стилометрических характеристик текста для определения его жанра является перспективной и актуальной задачей. Определение степени влияния различных типов параметров позволит строить эффективные системы извлечения информации из текстов на естественном языке и проведения лингвистических исследований.

Ритмические характеристики

Числовые ритмические характеристики основываются на ритмических средствах, непосредственно появляющихся в тексте. В данном исследовании используются лексико-грамматические средства: анафора, эпифора, симплока, анадиплозис, эпаналепсис, многосоюзие, диакопа, эпизевксис, хиазм, апозиопеза, повторяющиеся вопросительные и восклицательные предложения. Определения ритмических средств и алгоритмы их поиска приведены в предыдущих работах авторов [9, 10]. Апозиопеза и повторяющиеся вопросительные и восклицательные предложения основываются на появлениях знаков препинания. Остальные ритмические средства состоят из повторяющихся слов или словосочетаний.

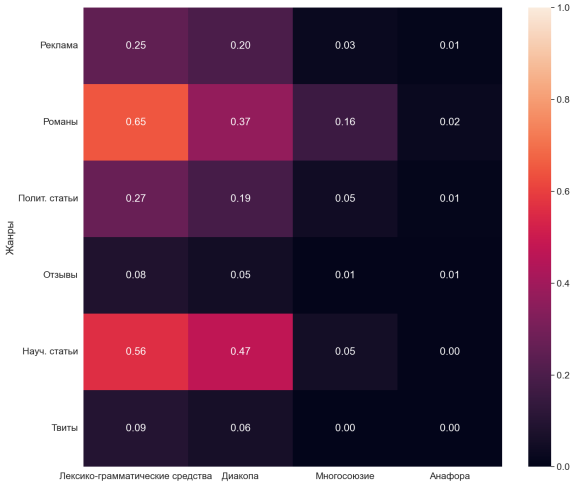
Для лексико-грамматических средств были вычислены следующие стилометрические характеристики:

1. количество появлений в тексте конкретного средства, делённое на количество предложений;
2. количество появлений в тексте всех средств, делённое на количество предложений;
3. доля уникальных слов среди всех, составляющих средства, т. е. тех, которые повторяются только один раз;
4. доли существительных, прилагательных, глаголов и наречий среди слов, составляющих средства.

Данные ритмические характеристики описывают как относительную частоту появления лексико-грамматических средств, т. е. их плотность, так и статистику для структуры лексических средств.

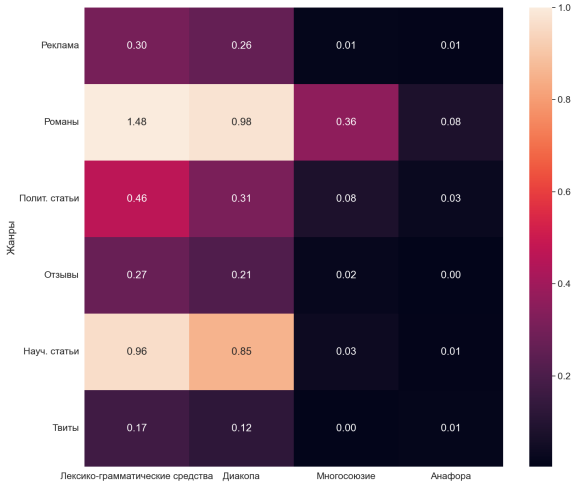
Статистический анализ ритмических характеристик

Для статистического анализа ритмических характеристик были собраны корпуса текстов в шести жанрах на русском и английском языках. Каждый корпус содержит по 100 художественных романов, научных статей, рекламных текстов и твитов, по 50 отзывов и политических статей, т. е. суммарно 500 текстов. Романы были взяты из корпуса из предыдущих исследований авторов [10]. Научные статьи были собраны из журналов Грамота, Диалог, International journal of digital evidence и Philosophical Transactions of the Royal Society of London. Рекламные тексты были взяты с сайтов auto.ru, detmir.ru и smartmedicalbuyer.com. Отзывы были собраны с сайта tripadvisor.com. Политические статьи представляют собой текстовые расшифровки речей президентов и министров России и США.



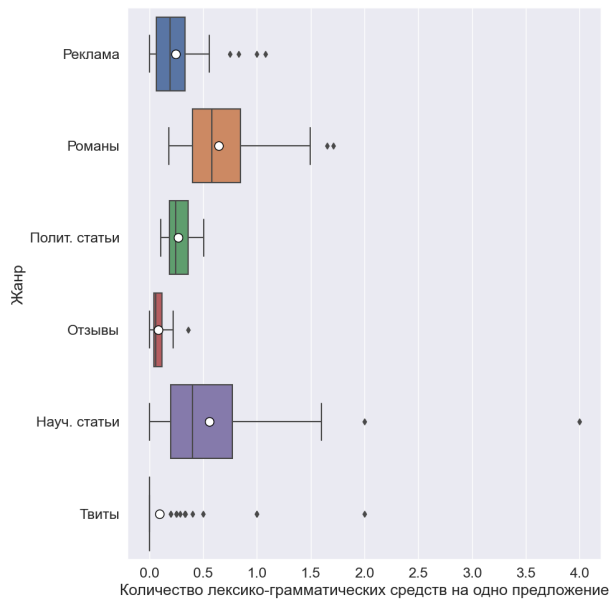
a)

Fig. 1. Heat map of mean values for genres for frequent features a) in Russian-language texts, b) in English-language texts



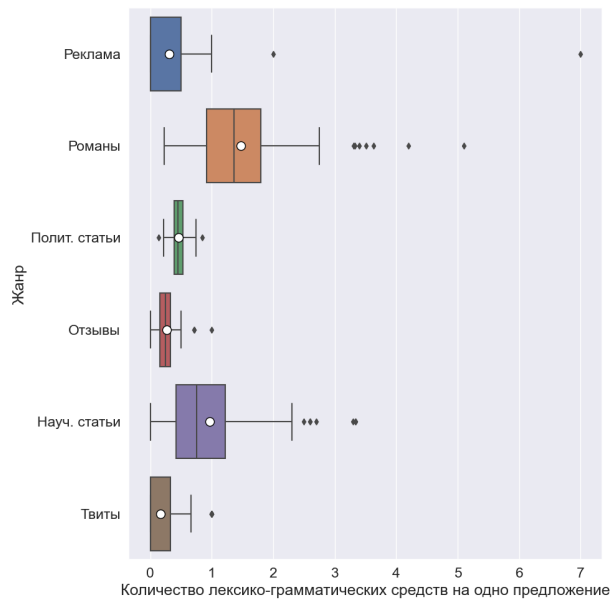
b)

Рис. 1. Тепловая карта средних значений по жанрам для часто встречающихся средств а) в русскоязычных текстах, b) в англоязычных текстах



a)

Fig. 2. Boxplot for lexico-grammatical features a) in Russian-language texts, b) in English-language texts



b)

Рис. 2. Диаграмма размаха лексико-грамматических средств а) в русскоязычных текстах, b) в англоязычных текстах

Ритмические характеристики вычисляются независимо для каждого текста и визуализируются с помощью тепловых карт и диаграмм размаха.

Тепловые карты представлены на рис. 1. Строки соответствуют жанрам, столбцы — самым часто встречающимся ритмическим средствам. В ячейках указано, сколько раз в среднем данное средство появляется в одном предложении в текстах данного жанра — плотность ритмического средства.

В текстах лексико-грамматические средства появляются чаще всего в романах и научных статьях, реже всего — в рекламе, отзывах и твитах. Самые популярные лексико-грамматические средства — диакопа, многосоюзие и анафора. Романы содержат больше разнообразных ритмических средств, чем тексты других жанров, в научных статьях обычно появляется диакопа и очень редко — другие средства. Твиты практически не содержат ритмических средств.

Отзывы на английском языке содержат в среднем в несколько больше маркеров ритма, чем на русском, и по плотности ритмических средств близки к рекламе. В остальном жанры на разных языках похожи по тенденциям ритма с поправкой на то, что англоязычные тексты содержат в среднем больше ритмических средств, чем русскоязычные тексты в тех же жанрах.

Более подробно распределение плотности лексико-грамматических средств представлено на диаграммах размаха 2. Прямоугольник показывает границы первого и третьего квартилей распределения, чёрная вертикальная линия внутри него — медианное значение, белый круг — среднее значение. Чёрная горизонтальная линия показывает граничные значения распределения, чёрные ромбы — выбросы.

Диаграммы показывают, что отзывы и политические статьи достаточно однородны по количеству ритмических средств. Среди русскоязычных текстов наиболее разнообразны по плотности ритма твиты и научные статьи, среди англоязычных — реклама и романы.

В целом графики и диаграммы показывают, что каждый жанр имеет свои особенности ритма, и по совокупности характеристик жанры отличаются друг от друга. Это означает, что ритмические характеристики могут быть хорошими маркерами жанра, что должна подтвердить классификация текстов.

Классификация по жанрам

Описанный корпус текстов был классифицирован по жанрам на основе числовых ритмических характеристик. Для классификации были взяты все ритмические характеристики, кроме количества появлений в тексте всех средств, делённого на количество предложений, поскольку она является суммой других характеристик.

Классификация проводилась двумя способами:

- бинарная классификация для каждого жанра, когда тексты классифицировались на принадлежащие и не принадлежащие конкретному жанру;
- мультиклассовая классификация на шесть жанров: художественные романы, научные статьи, политические статьи, рекламные статьи, отзывы, твиты.

Для обоих способов применялись одни и те же классификаторы:

- классификатор AdaBoost — мета-алгоритм машинного обучения, который объединяет результаты 50 классификаторов-деревьев решений, корректирующих неправильно классифицированные тексты;
- двунаправленная LSTM — рекуррентная нейронная сеть со слоем двунаправленной долгой краткосрочной памяти (LSTM) с 64 блоками и полносвязным выходным слоем, использующим функцию активации Softmax для мультиклассовой классификации и Sigmoid для бинарной;
- GRU — рекуррентная нейронная сеть со слоем Gated Recurrent Unit (GRU) с 4 блоками и полносвязным выходным слоем, использующим функцию активации Softmax для мультиклассовой классификации и Sigmoid для бинарной.

Table 1. Binary text classification by genres for Russian language

Таблица 1. Бинарная классификация текстов по жанрам для русского языка

Жанр	Классификатор	Точность	Стд. откл.	Полнота	Стд. откл.	F-мера	Стд. откл.
Реклама	GRU	44.8	1.3	50.0	0.0	47.3	0.4
Романы	GRU	**94.2**	2.8	**92.5**	2.3	**94.1**	1.9
Полит. статьи	GRU	44.6	1.6	50.0	0.0	47.1	0.9
Отзывы	GRU	71.1	18.0	69.6	10.8	66.1	17.4
Научн. статьи	GRU	79.3	2.7	76.1	6.9	77.5	5.1
Твиты	GRU	83.3	3.2	81.0	5.2	81.1	3.3
Реклама	LSTM	**68.5**	21.3	52.7	3.5	52.5	7.4
Романы	LSTM	**96.6**	1.9	**94.6**	2.6	**96.7**	1.9
Полит. статьи	LSTM	85.3	6.7	67.9	11.5	68.0	16.0
Отзывы	LSTM	81.0	11.2	**87.3**	12.1	**86.9**	6.3
Научн. статьи	LSTM	80.5	2.8	76.1	4.6	78.5	2.2
Твиты	LSTM	85.3	7.3	83.3	7.1	86.2	2.4
Реклама	AdaBoost	58.1	7.2	**56.5**	5.4	**56.7**	6.3
Романы	AdaBoost	**97.0**	1.5	**97.9**	1.0	**97.4**	1.0
Полит. статьи	AdaBoost	**95.7**	4.8	**92.1**	5.6	**93.3**	3.9
Отзывы	AdaBoost	**83.5**	17.1	79.4	15.5	81.2	16.0
Научн. статьи	AdaBoost	**84.6**	1.3	**84.3**	3.3	**84.1**	1.7
Твиты	AdaBoost	**90.9**	2.7	**89.5**	4.1	**89.9**	3.0

Для обучения нейронных сетей LSTM и GRU применяется категориальная кросс-энтропия как функция потерь и алгоритм оптимизации Adam.

Данные классификаторы уже доказали свое качество в решении современных задач обработки ритма текстов [9, 10], поэтому они были выбраны для экспериментов.

Для классификации корпус был разделён случайным образом на обучающую и тестовую выборки в отношении 4:1. Это позволило провести пятикратную кросс-валидацию для анализа стабильности результатов. Оценка качества выполнялась с помощью трёх стандартных мер: точность, полнота и F-мера [19], а также их стандартные отклонения.

Алгоритмы визуализации ритмических характеристик, классификации по жанрам и оценки результатов реализованы в инструменте ProseRhythmDetector, который доступен в Интернете по адресу https://github.com/text-processing/prose-rhythm-detector. Он написан на языке программирования Python и использует библиотеки StanfordNLP 0.2.0, Scikit-Learn 0.23.2 и Keras 2.4.3.

Результаты бинарной классификации представлены в таблицах 1 и 2 для русского и английского языка соответственно. Для точности, полноты и F-меры справа указаны стандартные отклонения при кросс-валидации.

Среди всех жанров лучше всего отделяются от остальных художественные романы с F-мерой более 97 % и политические статьи с F-мерой более 92 %. Отзывы, научные статьи и твиты тоже хорошо классифицируются (F-мера более 76 %). Следует отметить, что научные статьи на английском языке классифицируются лучше, чем на русском, а твиты — наоборот. Русскоязычная реклама отделяется от других жанров хуже остальных.

Стандартные отклонения в большинстве случаев низки: менее 5 %, что говорит о высокой стабильности классификации.

Среди трёх классификаторов лучших значений точности, полноты и F-меры чаще достигает AdaBoost за исключением русскоязычных отзывов, где LSTM превосходит его по F-мере на 5 % и показывает стандартное отклонение ниже на 10 %. Романы классифицируются очень хорошо всеми классификаторами: F-мера более 94 %.

Table 2. Binary text classification by genres for English language

Таблица 2. Бинарная классификация текстов по жанрам для английского языка

Жанр	Классификатор	Точность	Стд. откл.	Полнота	Стд. откл.	F-мера	Стд. откл.
Реклама	GRU	73.4	2.2	67.6	2.4	68.4	4.3
Романы	GRU	**93.4**	1.7	**94.3**	4.1	**94.3**	1.9
Полит. статьи	GRU	82.5	18.8	64.8	9.2	63.8	20.8
Отзывы	GRU	45.4	0.8	51.0	2.0	47.4	0.5
Научн. статьи	GRU	87.1	2.0	83.5	3.1	83.0	3.1
Твиты	GRU	**79.1**	5.8	72.9	3.0	76.8	4.9
Реклама	LSTM	**78.3**	3.3	72.9	5.1	71.8	3.7
Романы	LSTM	**96.5**	2.6	**95.8**	2.5	**95.8**	1.9
Полит. статьи	LSTM	87.7	7.6	80.5	12.1	83.0	6.2
Отзывы	LSTM	68.1	16.5	58.0	6.7	61.5	11.6
Научн. статьи	LSTM	88.2	5.3	85.5	3.5	86.4	3.8
Твиты	LSTM	**82.4**	5.7	73.9	3.8	74.8	5.4
Реклама	AdaBoost	75.6	4.4	**73.3**	4.3	**74.3**	4.4
Романы	AdaBoost	**98.3**	2.0	**98.0**	2.3	**98.1**	2.1
Полит. статьи	AdaBoost	**94.3**	2.7	**92.1**	6.6	**92.7**	3.8
Отзывы	AdaBoost	**79.3**	11.8	**74.2**	8.1	**76.3**	9.3
Научн. статьи	AdaBoost	**91.6**	3.9	**88.3**	3.5	**89.7**	3.6
Твиты	AdaBoost	78.5	4.0	**76.4**	3.1	**77.1**	3.1

Table 3. Multi-class text classification by genres for Russian language

Таблица 3. Мультиклассовая классификация текстов по жанрам для русского языка

Классификатор	Точность	Стд. откл.	Полнота	Стд. откл.	F-мера	Стд. откл.
GRU	71.5	11.2	69.7	4.0	65.5	2.9
LSTM	**77.1**	5.6	**77.5**	5.5	**77.5**	6.8
AdaBoost	43.3	14.0	43.3	9.0	36.7	12.1

Таблицы 3 и 4 демонстрируют результаты для мультиклассовой классификации. Здесь уже классификатор на основе нейросети LSTM существенно превосходит остальные: он достигает F-меры более 74 %, тогда как AdaBoost не показывает и 40 % F-меры.

В целом результаты у мультиклассовой классификации ниже, чем в лучших случаях у бинарной, а стандартные отклонения такие же низкие. Тем не менее точность, полнота и F-мера значительно высоки: более 72 %. Мультиклассовая классификация выполняется более эффективно для русского языка: для него точность, полнота и F-мера составляют 77 %.

Для того, чтобы обнаружить и проанализировать ошибки классификации, для мультиклассовой классификации алгоритмом LSTM были собраны неверные предсказания алгоритма. Они агрегированы в таблицах 5 и 6 как пример ошибок из одного раунда классификации при делении корпусов текстов случайным образом на обучающую и тестовую выборки в отношении 4:1.

Строки таблицы соответствуют исходным жанрам текстов, а столбцы — жанрам, предсказанным неверно. В ячейках указывается, сколько текстов исходного жанра было ошибочно причислено к

Table 4. Multi-class text classification by genres for English language

Таблица 4. Мультиклассовая классификация текстов по жанрам для английского языка

Классификатор	Точность	Стд. откл.	Полнота	Стд. откл.	F-мера	Стд. откл.
GRU	71.0	4.6	71.7	3.4	68.9	5.1
LSTM	**72.4**	2.7	**75.3**	2.2	**74.1**	3.7
AdaBoost	42.7	8.0	46.6	2.7	39.5	3.5

Table 5. Errors of the multi-class text classification by genres for Russian language

Таблица 5. Ошибки мультиклассовой классификации текстов по жанрам на русском языке

Исходный жанр	Реклама	Романы	Полит. статьи	Отзывы	Научн. статьи	Твиты
Реклама	-	0	5	0	4	3
Романы	0	-	0	0	0	0
Полит. статьи	0	0	-	0	1	0
Отзывы	0	0	0	-	0	2
Научн. статьи	0	0	3	0	-	3
Твиты	1	0	0	0	2	-

Table 6. Errors of the multi-class text classification by genres for English language

Таблица 6. Ошибки мультиклассовой классификации текстов по жанрам на английском языке

Исходный жанр	Реклама	Романы	Полит. статьи	Отзывы	Научн. статьи	Твиты
Реклама	-	0	0	2	0	4
Романы	0	-	2	0	0	0
Полит. статьи	0	1	-	0	0	0
Отзывы	3	0	0	-	2	0
Научн. статьи	1	0	1	1	-	0
Твиты	7	0	0	0	2	-

жанру, указанному в столбце. Например, пять рекламных текстов на русском языке были приняты за политические статьи.

Из результатов агрегирования ошибок можно сделать вывод, что классификатор нередко причисляет рекламу к любому жанру, кроме художественных романов. Твиты и реклама часто смешиваются между собой. Ошибки в остальных жанрах достаточно случайны и, вероятно, вызываются особенностями конкретных текстов. Например, два англоязычных романа, которые были ошибочно классифицированы как политические статьи, содержат мало ритмических средств, что обычно не характерно для их жанра.

Обсуждение результатов с лингвистической точки зрения

Полученные результаты по количеству ритмических средств в текстах разных жанров позволяет судить о ритмической специфике того или иного жанра в рамках одного языка. В русском языке, в котором по сравнению с другими языками, как показали ранние исследования [9], частотность ритмических средств ниже, чем в других языках, что подтверждают показатели по жанрам. Это безусловно связано с языковой спецификой, типологическими особенностями языка, в особенности с критериями частеречной классификации. Например, в английском языке значим как морфологический, так и синтаксический критерий в частеречной классификации, но для одноморфемных слов важен синтаксический критерий, их позиция в предложении. Кроме того, английский язык отличается большей степенью номинативности в сопоставлении с русским. Номинативность английского языка увеличивается также за счет герундия. Такое стремление к номинативности обусловлено этносоциокультурными факторами, а именно исторически сформировавшимся в рамках британской культуры уважением к факту и научной точности в разговоре [20]. Русский язык по сравнению с английским обладает большей степенью глагольности, при этом довольно много в русском языке глаголов, передающих эмоциональное состояние (волноваться, гневаться, раздражаться, радоваться и т. д.), что подчеркивает значимость эмоционального начала [21].

Данные особенности важны для характеристики ритмических средств, поскольку большая их часть выражается при помощи существительных, что обусловлено грамматической структурой

предложения, в котором второстепенные члены реже всего выражены глаголом, но чаще именем существительным, а также прилагательным или наречием. Таким образом, английский язык в целом обладает большей степенью ритмизации только на основе своей структуры, а именно на основе стремления к номинативности в морфологии.

Что касается различных показателей в различных жанрах, то несомненным является преобладание ритмических средств в художественной литературе, поскольку поэтичность изложения, творческий подход позволяют сосредоточиться на образах, которые могут реализовываться через различные типы повторов. Объяснением того, что научный текст близок к художественному, может быть определенная структурированная, отлаженная терминосистема, характерная для научных текстов в целом. В этом случае повторы обусловлены необходимостью оперирования конкретной для каждой дисциплины и отрасли знания лексики.

Что касается твитов, которые содержат наименьшее количество средств, то в качестве основной причины этого можно отметить их прагматическую функцию — выражение собственного мнения, четкое, краткое, не поэтизированное. То же и для отзывов, которые в русском языке также имеют низкий уровень ритмических средств. Однако в англоязычных отзывах, как и в рекламе в обоих языках, ритмические средства более активны, что также может объясняться большей номинативностью английского языка.

Заключение

В статье оценивалось влияние ритмических характеристик текстов на определение жанра. Задача была выполнена в два этапа: статистический анализ ритмических характеристик и классификация текстов по шести жанрам: художественные романы, научные статьи, политические статьи, рекламные статьи, отзывы, твиты. Визуализация статистических данных о ритмических характеристиках показала, что тексты различных жанров отличаются по маркерам стиля. При классификации текстов по жанрам с помощью этих характеристик и современных алгоритмов AdaBoost и LSTM были достигнуты достаточно высокие значения метрик качества: не менее 76 % F-меры для всех жанров, кроме рекламы. С наивысшим качеством около 98 % точности, полноты и F-меры были классифицированы художественные романы.

Перспективным направлением дальнейших исследований будет анализ ошибок классификации, который позволит лучше изучить ритмические особенности текстов и учесть их в моделях текстов.

References

[1] J. Worsham and J. Kalita, “Genre identification and the compositional effect of genre in literature”, in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 1963–1973.

[2] M. N. Melissourgou and K. T. Frantzi, “Genre identification based on SFL principles: The representation of text types and genres in English language teaching material”, *Corpus Pragmatics*, vol. 1, no. 4, pp. 373–392, 2017.

[3] L. A. Kochetova and V. V. Popov, “Research of Axiological Dominants in Press Release Genre based on Automatic Extraction of Key Words from Corpus”, *Nauchnyi dialog*, no. 6, 2019, In Russian.

[4] S. E. Murphy, “Shakespeare and his contemporaries: Designing a genre classification scheme for Early English Books Online 1560-1640”, *ICAME Journal*, pp. 59–82, 2019.

[5] R. Malhotra and A. Sharma, “Quantitative evaluation of web metrics for automatic genre classification of web pages”, *International Journal of System Assurance Engineering and Management*, vol. 8, no. 2, pp. 1567–1579, 2017.

[6] D. DEJICA, "Understanding Technical and Scientific Translation: A Genre-based Approach", *Scientific Bulletin of the Politehnica University of Timisoara. Transactions on Modern Languages/Buletinul Stiintific al Universitatii Politehnica din Timisoara. Seria Limbi Moderne*, vol. 19, no. 1, pp. 56–66, 2020.

[7] V. Thakur and A. C. Patel, "An Improved Dictionary Based Genre Classification Based on Title and Abstract of E-book Using Machine Learning Algorithms", in *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security*, Springer, 2021, pp. 323–337.

[8] A. Cimino, M. Wieling, F. Dell'Orletta, S. Montemagni, and G. Venturi, "Identifying predictive features for textual genre classification: the key role of syntax", *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017*, pp. 107–112, 2017.

[9] K. Lagutina, A. Poletaev, N. Lagutina, E. Boychuk, and I. Paramonov, "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*, pp. 247–255, 2020.

[10] K. Lagutina, N. Lagutina, E. Boychuk, V. Larionov, and I. Paramonov, "Authorship verification of literary texts with rhythm features", *Proceedings of the 28th Conference of Open Innovations Association FRUCT*, pp. 240–251, 2021.

[11] A. Onan, "An ensemble scheme based on language function analysis and feature engineering for text genre classification", *Journal of Information Science*, vol. 44, no. 1, pp. 28–47, 2018.

[12] A. M. El-Halees, "Arabic Text Genre Classification", *Journal of Engineering Research and Technology*, vol. 4, no. 3, pp. 105–109, 2017.

[13] I. A. Batraeva, A. D. Nartsev, and A. S. Lezgyan, "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 vychislitelnaja tehnika i informatika*, no. 50, pp. 14–22, 2020, In Russian.

[14] V. B. Barahnin, O. Y. Kozhemyakina, E. V. Rychkova, I. S. Pastushkov, and Y. S. Borzilova, "Izvlechenie leksicheskih i metroritmicheskih priznakov, harakternyh dlya zhanra i stilya i ih kombinacij v processe avtomatizirovannoj obrabotki tekstov na russkom yazyke", *Sovremennye informacionnye tekhnologii i IT-obrazovanie*, vol. 14, no. 4, pp. 888–895, 2018, In Russian.

[15] O. A. Mitrofanova and A. D. Moskvina, "On the Role of Prepositional Statistics for Genre Identification of Russian texts", *International Journal of Open Information Technologies*, vol. 8, no. 11, pp. 91–96, 2020, In Russian.

[16] L. G. Gorbich and A. A. Zhivoderov, "Using statistical indexes to distinguish between scientific and popular science texts on the example of the works of A. E. Fersman", *Software & Systems*, vol. 33, no. 4, pp. 720–725, 2020, In Russian.

[17] A. R. Dubovik, "Automatic text style identification in terms of statistical parameters", *Komp'yuternaya lingvistika i vychislitel'nye ontologii*, no. 1, pp. 29–45, 2017, In Russian.

[18] A. Y. Antonova, E. S. Klyshinskij, and E. V. YAgunova, "Opredelenie stilevyh i zhanrovyh harakteristik kollekcij tekstov na osnove chasterechnoj sochetaemosti", *Otkrytye sistemy*, vol. 3, pp. 80–85, 2011, In Russian.

[19] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks", *Information processing & management*, vol. 45, no. 4, pp. 427–437, 2009.

[20] L. Kozlova, "Sravnitel'naya tipologiya anglijskogo i russkogo yazykov", *Barnaul: AltGPU*, no. 20019, p. 180, 2019, In Russian.

[21] A. Wierzbicka, *The semantics of grammar*. John Benjamins Publishing, 1988, vol. 18, p. 617.