MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 29, NO. 4, 2022
journal homepage: www.mais-journal.ru

THEORY OF DATA

Classification of Russian Texts by Genres Based on Modern Embeddings and Rhythm

K. V. Lagutina[1]

DOI: 10.18255/1818-1015-2022-4-334-347

[1]P. G. Demidov Yaroslavl State University, 14 Sovetskaya str., Yaroslavl 150003, Russia.

MSC2020: 68T50
Research article
Full text in Russian

Received August 17, 2022
After revision November 4, 2022
Accepted November 9, 2022

The article investigates modern vector text models for solving the problem of genre classification of Russian-language texts. Models include ELMo embeddings, BERT language model with pre-training and a complex of numerical rhythm features based on lexico-grammatical features. The experiments were carried out on a corpus of 10,000 texts in five genres: novels, scientific articles, reviews, posts from the social network Vkontakte, news from OpenCorpora.
Visualization and analysis of statistics for rhythm features made it possible to identify both the most diverse genres in terms of rhythm: novels and reviews, and the least ones: scientific articles. Subsequently, these genres were classified best with the help of rhythm features and the neural network-classifier LSTM. Clustering and classifying texts by genre using ELMo and BERT embeddings made it possible to separate one genre from another with a small number of errors. The multi-classification F-score reached 99%. The study confirms the efficiency of modern embeddings in the tasks of computational linguistics, and also allows to highlight the advantages and limitations of the complex of rhythm features on the material of genre classification.

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

INFORMATION ABOUT THE AUTHORS

Ksenia Vladimirovna Lagutina
correspondence author
orcid.org/0000-0002-1742-3240. E-mail: lagutinakv@mail.ru
PhD.

Funding: The work is supported by the President of Russian Federation Scholarship for young scientists and postgraduates No. SP-2109.2021.5.

For citation: K. V. Lagutina, “Classification of Russian Texts by Genres Based on Modern Embeddings and Rhythm”, *Modeling and analysis of information systems*, vol. 29, no. 4, pp. 334-347, 2022.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 29, № 4, 2022
сайт журнала: www.mais-journal.ru

THEORY OF DATA

Классификация русскоязычных текстов по жанрам на основе современных эмбеддингов и ритма

К. В. Лагутина[1]

DOI: 10.18255/1818-1015-2022-4-334-347

[1]Ярославский государственный университет им. П. Г. Демидова, ул. Советская, д. 14, г. Ярославль, 150003 Россия.

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 17 августа 2022 г.
После доработки 4 ноября 2022 г.
Принята к публикации 9 ноября 2022 г.

В статье исследуются современные векторные модели текстов для решения задачи классификации русскоязычных текстов по жанрам. Модели включают эмбеддинги ELMo, языковую модель BERT с предобучением и комплекс числовых ритмических характеристик на основе лексико-грамматических средств. Эксперименты проводились на корпусе из 10 000 текстов пяти жанров: романы, научные статьи, отзывы, посты из социальной сети Вконтакте, новости из OpenCorpora.

Визуализация и анализ статистики для ритмических характеристик позволили выделить как наиболее разнообразные по ритму жанры: романы и отзывы, так и наименее – научные статьи. Именно эти жанры были впоследствии классифицированы лучше всего с помощью ритма и нейросети-классификатора LSTM. Кластеризация и классификация текстов по жанрам с помощью эмбеддингов ELMo и BERT позволила отделить один жанр от другого с небольшим количеством ошибок. F-мера мультиклассификации достигла 99%. Исследование подтверждает эффективность современных эмбеддингов в задачах компьютерной лингвистики, а также позволяет выделить достоинства и ограничения комплекса ритмических характеристик на материале классификации по жанрам.

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

ИНФОРМАЦИЯ ОБ АВТОРАХ

Ксения Владимировна Лагутина
автор для корреспонденции

orcid.org/0000-0002-1742-3240. E-mail: lagutinakv@mail.ru
кандидат технических наук.

Финансирование: 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.

Для цитирования: K. V. Lagutina, “Classification of Russian Texts by Genres Based on Modern Embeddings and Rhythm”, *Modeling and analysis of information systems*, vol. 29, no. 4, pp. 334-347, 2022.

Введение

Изучение стилевых особенностей русского языка — важная задача современной российской компьютерной лингвистики. Стилевые характеристики текста отражают в том числе и структурные особенности текстов, поэтому они являются маркерами различных жанров и часто используются для автоматического анализа в этой области [1].

Автоматическая классификация текстов по жанрам является одной из фундаментальных задач обработки текстов. Зная жанр текста, можно эффективнее решить и другие проблемы компьютерной лингвистики: определить часть речи или значение слова или словосочетания, найти подходящий по смыслу запроса документ и т. п. [2]. Как и многие задачи автоматического анализа текстов, для классификации англоязычных текстов по жанрам уже предложено несколько высококачественных решений [3, 4].

Русскоязычные тексты представляют собой более широкое поле для исследований. Учёные часто классифицируют тексты на основе стандартных характеристик или ограничиваются литературными жанрами [5]. Современные эмбеддинги, например, BERT, и вовсе остаются недоисследованными в области анализа жанров.

Автор статьи ставит перед собой задачу классифицировать русскоязычные тексты по жанрам на романы, научные статьи, отзывы, новости и посты в социальной сети Вконтакте с помощью современных числовых характеристик. Характеристики включают в себя эмбеддинги BERT и ELMo, а также разработанный автором ранее комплекс ритмических характеристик [6]. Также в цели исследования входит визуализация числовых характеристик для наглядного анализа жанров и интерпретации результатов классификации.

1. Обзор смежных работ

В научной литературе последних лет активно изучаются возможности предобученных языковых моделей и эмбеддингов, построенных на основе архитектуры Трансформер, а именно, подходы ELMo, GPT и BERT [7, 8]. Среди них ELMo и BERT хорошо подходят не только для генерации текстов, но и для классификационных задач.

Авторы исследования [4] применяют эмбеддинги GloVe, ELMo и BERT для классификации новостей по жанрам и структуре, используя нейронную сеть BiLSTM в качестве классификатора. ELMo и BERT достигают очень близких результатов по качеству: около 80 % точности для мультиклассификации. Эксперименты проводились на 853 англоязычных новостях. Следует отметить, что для достаточно небольшого корпуса текстов достигнуто отличное качество классификации.

Подобные высокие результаты часто достигаются для текстов на английском языке, поскольку для него разработано немало стандартных корпусов, размеченных для решения разнообразных задач компьютерной лигнвистики. Это позволяет проводить масштабные исследования с различными эмбеддингами: word2vec, GloVe, FastText, ELMo, BERT, — и демонстрировать высокую эффективность моделей на их основе сразу для нескольких задач классификации [9]. F-мера для классификации по темам стабильно достигает 80–90 %.

Для национальных языков современные эмбеддинги в области задач классификации также исследованы достаточно хорошо, но для исследования жанра или стиля практически не применяются.

Для русского языка модель RuBERT, являющаяся адаптацией BERT, предобучена и применена для классификации по тональности [10]. F-мера достигла 72 %.

Эмбеддинги BERT и ELMo применялись и участниками научного соревнования RuShiftEval по детекции изменения семантики слов в русскоязычном тексте [11]. Лучшие результаты (коэффициент корреляции Спирмена 80 %) показала модель XLM-RoBERTa, основанная на мультиязычной версии BERT, в комбинации с системой разрешения неоднозначности смысла слова.

Оригинальные версии ELMo и BERT для русского языка достигли только 50–55 % коэффициента корреляции Спирмена [12].

Глазкова [13] успешно классифицирует фрагменты биографических текстов на русском языке по десяти темам. Модель RuBERT достигает высокого значения F-меры в 93 % и превосходит мультиязычную версию BERT, word2vec и стандартный подход TF-IDF в сочетании с SVM.

В области классификации текстов по жанрам для русского языка применялись свёрточные нейронные сети и эмбеддинги word2vec [14]. Точность классификации на пять жанров: история, детективы, детская литература, поэзия, фантастика, — достигла около 78 %.

Автор статьи в своей предыдущей работе [5] вместе с коллегами исследовала комплекс ритмических характеристик для анализа шести жанров: художественные романы, научные статьи, политические статьи, рекламные статьи, отзывы, твиты. С помощью этих характеристик и классификаторов AdaBoost и LSTM были достигнуты достаточно высокие значения метрик качества: не менее 76 % F-меры для всех жанров, кроме рекламы.

Таким образом, в области классификации текстов по жанрам на русском языке актуальные языковые модели ELMo и BERT пока ещё почти не изучены. Но их высокие результаты для анализа жанров англоязычных корпусов, а также для других задач классификации русскоязычных текстов позволяют ожидать высокого качества решения и для проблемы, исследуемой в данной статье. Кроме того, сравнение нейросетевых и лингвистических моделей текстов, в частности, модели ритма, даст хороший материал для анализа достоинств и ограничений данных характеристик текста.

2. Корпус текстов пяти жанров

Для анализа русскоязычных текстов в различных стилях были выбраны пять жанров: художественные романы, научные статьи, отзывы, новости и посты в социальной сети Вконтакте. Каждому жанру соответствует 2 000 текстов, у текста может быть только один жанр.

Художественные тексты в данном исследовании — это фрагменты романов русскоязычных писателей XIX–XXI веков. Фрагменты содержат около 20 000 знаков и целое количество абзацев. Исходные тексты достаточно велики и существенно отличаются по объёму как друг от друга, так и от текстов других жанров, поэтому было принято решение выделить из них фрагменты заданного размера случайным образом, чтобы уравнять и уменьшить объёмы текстов. В результате тексты в среднем содержат 2 982 слова.

Научные статьи были собраны из журналов Грамота, Диалог и Кардиология. Они также были разделены на фрагменты объёмом около 2 000 знаков и целым количеством абзацев. В результате тексты в среднем содержат 191 слово.

Тексты отзывов включают в себя положительные, отрицательные и нейтральные русскоязычные отзывы на фильмы с интернет-ресурса Кинопоиск. Они в среднем содержат 406 слов.

Новостные тексты были взяты из корпуса текстов OpenCorpora [15]. Он содержит в том числе и новостные тексты из онлайн-медиа. Они в среднем содержат 433 слова.

Посты в социальной сети Вконтакте были собраны через API сайта из 50 различных групп, посвящённых разнообразным тематикам: наука, книги, кино и т. п. Выбирались не посты о новостях, которые часто являются дублями статей онлайн-медиа, а оригинальные авторские тексты, созданные именно для социальной сети. Они в среднем содержат 127 слов.

Кроме того, при сборе корпуса собирались только тексты, содержащие ритмические средства. Поэтому отсеивалось примерно 17 % текстов OpenCorpora, 2 % отзывов, 4 % научных статей, 30 % постов Вконтакте. Все художественные фрагменты содержали ритмические средства.

Таким образом, был сформирован корпус из 10 000 русскоязычных текстов пяти жанров, где каждый текст содержит хотя бы одно ритмическое средство.

3. Моделирование текстов

Для исследования жанров автор использует три современные векторные модели, где каждый текст представляется в виде кортежа из числовых характеристик. Это эмбеддинги, полученные с помощью языковой модели BERT, эмбеддинги ELMo и комплекс ритмических характеристик, которые ранее исследовались автором [6].

Модель BERT была взята в версии для русского языка — RuBERT cased, т. е. RuBERT, учитывающий регистр букв. У данной модели, как и у оригинального BERT, имеется ограничение: длина исходного текста не может превышать 512 токенов. Токенами в данном случае считаются слова, знаки препинания и специальные маркеры BERT [CLS] и [SEP], обозначающие начало текста и конец предложения. Тексты некоторых жанров, например, фрагменты художественных романов, могут быть большего размера, поэтому для подсчёта эмбеддингов их необходимо разделить на части.

Автор разбивала тексты на абзацы и считала эмбеддинги отдельно для каждого. Таким образом текст представляется как матрица из 768 столбцов, в которой каждая строка содержит эмбеддинг для своего абзаца. 768 — это длина одного BERT-эмбеддинга. Далее для каждого столбца считалось среднее арифметическое, так получался итоговый вектор для текста.

Модель ELMo также была взята в версии для русского языка из Python-библиотеки DeepPavlov. ELMo строит эмбеддинги для каждого слова и работает существенно медленнее BERT, поэтому для ускорения подсчётов автор из каждого текста выбирала фрагмент меньшего объёма, состоящий из целого количества предложений и обладающий размером около 1000 знаков. Как будет показано далее, данного объёма текста будет вполне достаточно для успешной классификации.

Таким образом, ELMo для каждого текста строит матрицу, где строки — это эмбеддинги отдельных слов, а количество столбцов — 2560 (число является длиной стандартного эмбеддинга ELMo). Как и для BERT, здесь также считаются средние значения по столбцам, чтобы получить итоговый вектор для текста.

Третья модель текстов — комплекс ритмических характеристик. Для их получения сначала в текстах ищутся ритмические 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 автора [6, 16]. Комплекс чистовых характеристик на основе представленных средств для данного исследования был расширен, чтобы изучить структуру ритма более подробно. Также из набора была исключена характеристика “доля уникальных слов”, так как по результатам экспериментов она оказалась наименее полезной. Итоговый комплекс характеристик выглядит следующим образом:

- количество появлений в тексте конкретного средства, делённое на количество предложений;
- доли существительных, прилагательных, глаголов, наречий, имён 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 в количестве слов.

В обновлённой версии комплекса ритмических характеристик исследуются не только самостоятельные, но и служебные части речи, а также размеры ритмических средств.

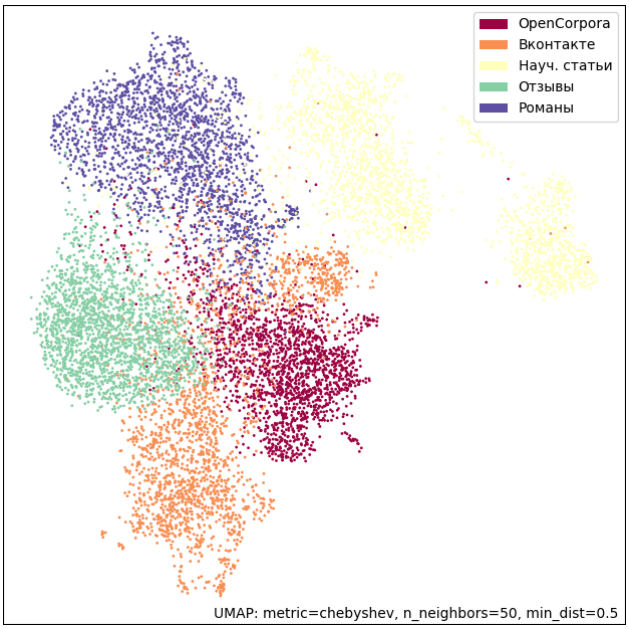
Таким образом, каждый текст представляется в виде эмбеддингов BERT (длина вектора — 768), эмбеддингов ELMo (длина вектора — 2560) или комплекса числовых характеристик, основанных на статистике ритмических средств (длина вектора — 24).

4. Визуализация числовых характеристик

Для того, чтобы визуализировать векторные модели и проанализировать особенности ритма в каждом жанре, для каждого набора характеристик были построены графики и диаграммы с разметкой по жанрам.

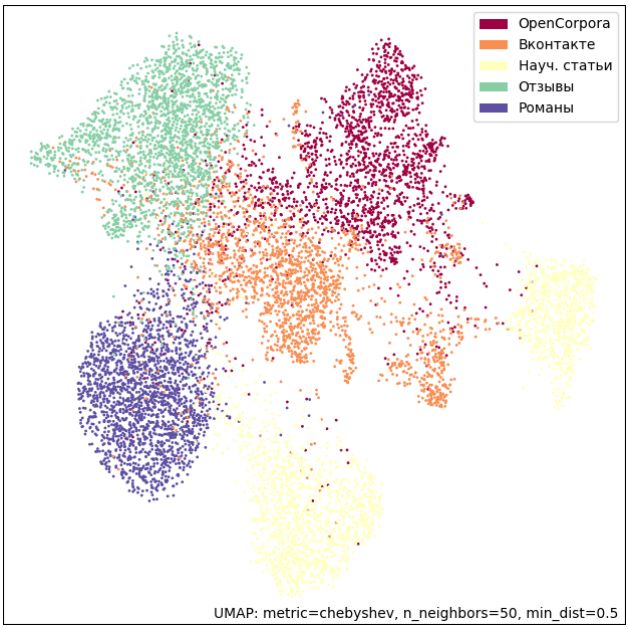
Так как эмбеддинги содержат числовые характеристики, которые невозможно интерпретировать отдельно друг от друга, было принято решение кластеризовать тексты на основе моделей BERT и ELMo и сопоставить полученные кластеры с исходными жанрами. Для визуализации был использован алгоритм UMap, который уменьшает размерность данных до двумерного пространства, кластеризуя объекты на основе их расстояния до k-ближайших соседей.

На рис. 1 представлены результаты кластеризации при k равном 50 и метрике Чебышёва для измерения расстояния. Жанры текстов отмечены различными цветами. И BERT, и ELMo позволяют отделить все жанры друг от друга: у каждого жанра имеется собственный кластер, мало пересекающийся с другими. Для научных статей появляется сразу два кластера, причём кластеры BERT ближе друг к другу, чем кластеры ELMo. Посты Вконтакте, наоборот, у BERT также делятся на два кластера (или один протяжённый, на который накладывается кластер OpenCorpora), у ELMo можно выделить один кластер. ELMo Новости из OpenCorpora и посты Вконтакте наиболее часто смешиваются друг с другом и с остальными жанрами. В целом, по результатам кластеризации можно ожидать высокого качества классификации с помощью эмбеддингов.



a)

Fig. 1. UMap clusters for embeddings
a) BERT, b) ELMo



b)

Рис. 1. Кластеры UMap для эмбеддингов
a) BERT, b) ELMo

Комплекс ритмических характеристик более разнообразен по содержанию, поэтому он визуализируется по нескольким группам характеристик на отдельных диаграммах.

Гистограмма на рис. 2 создана на основе количеств появлений в тексте конкретных средств и визуализирует для каждого жанра средний процент каждого средства среди всех ритмических средств. Справа в легенде перечислены все ритмические средства сверху вниз, на гистограмме их доли отмечены снизу вверх. Каждый столбец соответствует жанру текстов, числа — это проценты конкретных средств, число над столбцом — процент апозиопезиса среди всего ритма.

Гистограмма демонстрирует, что самыми частыми ритмическими средствами являются диакопа и многосоюзие, их проценты самые большие в каждом жанре: 48–84 % и 7–19 % соответственно. Значительные доли в некоторых жанрах имеют анафора, эпифора, эпаналепсис, эпизевксис и повторяющиеся вопросительные и восклицательные предложения: от 2 до 11 % в большинстве случаев. Наиболее разнообразен по римту жанр художественных романов, на второе место по этому параметру можно поставить жанр отзывов. Научные статьи и новости OpenCorpora, наоборот, содержат мало различных типов средств, а количественно в них преобладает диакопа.

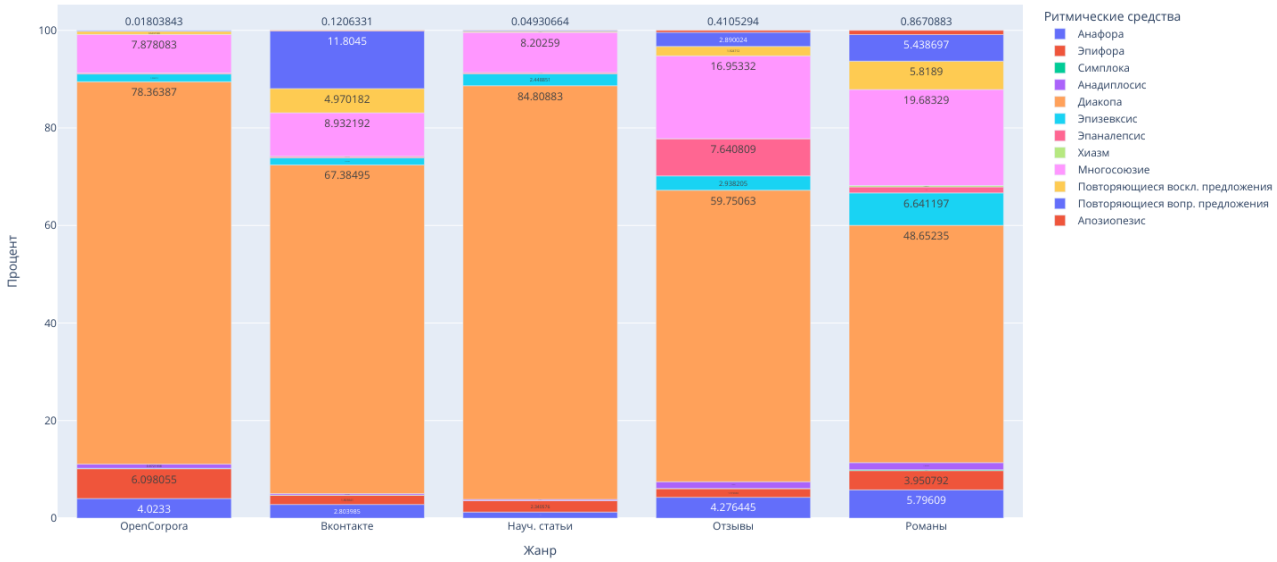


Fig. 2. The histogram with percents of rhythm features

Рис. 2. Гистограмма с процентами ритмических средств

Наиболее часто встречающиеся части речи отображены на тепловой карте на рис. 3. Это существительные (NOUN), прилагательные (ADJS), глаголы (VERB), наречия (ADVB), местоимения (PRON) и соединительные союзы (CCONJ). В столбцах тепловой карты указаны части речи, в строках — жанры, в ячейках — средний процент данной части речи в текстах данного жанра. Справа диапазон значений сопоставляется с диапазоном цветов: тёмные оттенки соответствуют меньшим долям, светлые — большим.

Тепловая карта показывает, что в романах и отзывах две части речи наиболее часто образуют ритмические средства: существительные (25 % и 22 %) и соединительные союзы (22 % и 19 %). В романах достаточно часто встречаются и глаголы: 16 %. В остальных жанрах самый большой процент среди всех частей речи занимают существительные, а соединительные союзы, хоть и имеют небольшой процент 9–10 %, но так же занимают второе место по количеству. Как следует из предыдущей гистограммы, существительные соответствуют диакопам, а многосоюзия — соединительным союзам.

Максимальное и среднее расстояния между первым и последним повторяющимся в средстве словом визуализированы на коробчатых диаграммах, см. рис. 4. Для каждого жанра отображено собственное распределение значений характеристики. Прямоугольник обозначает первый и третий квартили, вертикальная линия внутри него — медиану, белый круг — среднее значение, чёрные отрезки — минимальное и максимальное значение. Выбросы исключены из диаграммы.

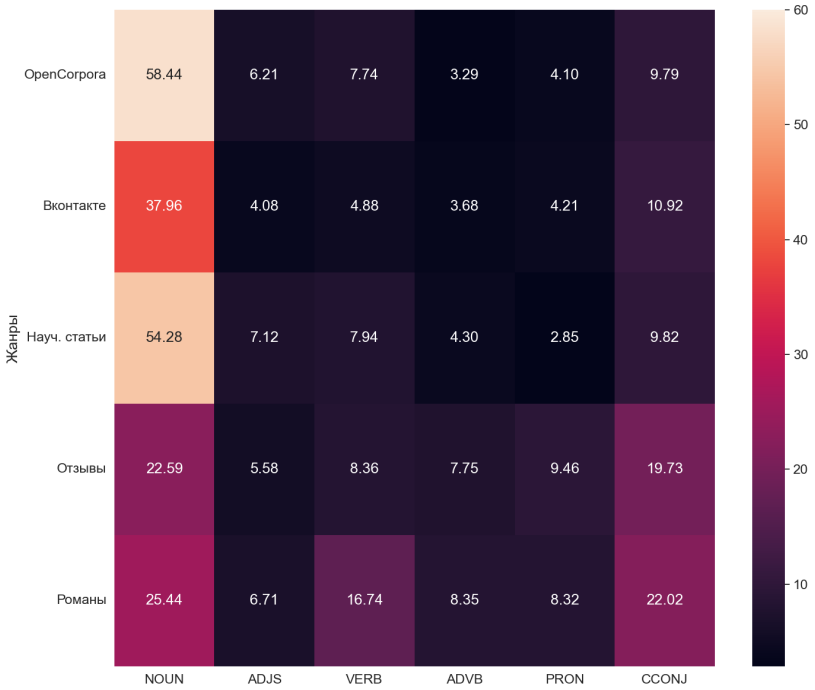
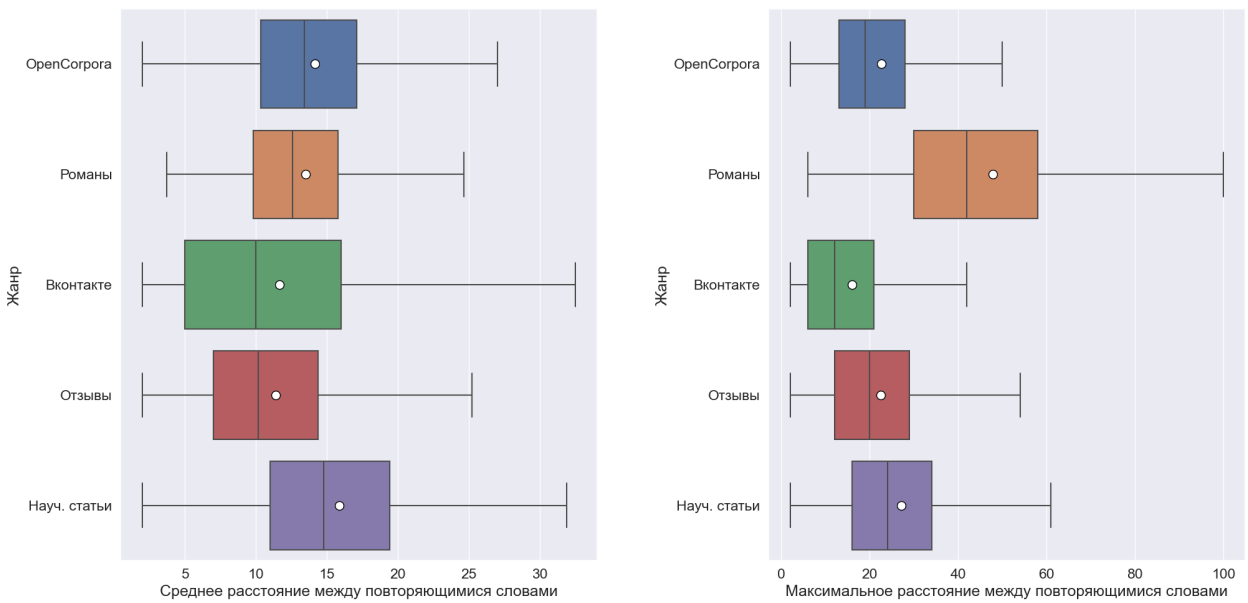


Fig. 3. The heat map with percents of the most frequent parts of speech

Рис. 3. Тепловая карта с процентами наиболее частых частей речи



a) b)

Fig. 4. The boxplot for a) average, b) maximal distance among words in rhythm features

Рис. 4. Коробчатая диаграмма для a) среднего, b) максимального расстояния между словами в ритмических средствах

Диаграммы показывают, что медианный размер ритмического средства составляет 10–15 слов от первого до последнего повторяющегося слова включительно. В целом жанры по статистике средних расстояний похожи друг на друга, только посты Вконтакте и научные статьи имеют чуть

больший диапазон значений, в том числе и между первым и третьим квартилем. По максимальному расстоянию ярко выделяются романы: все их статистические параметры больше, чем у других жанров.

Таким образом, визуализация демонстрирует, что жанры существенно отличаются друг от друга и по ритму, и по эмбеддингам на основе языковых моделей BERT и ELMo. Следовательно, векторные модели на основе данных характеристик могут быть хорошими маркерами жанров и обеспечивать качественную классификацию.

5. Классификация по жанрам

Постановка экспериментов

Тексты были классифицированы по жанрам на основе нескольких векторных моделей: эмбеддингов ELMo и BERT, ритмических характеристик, а также дополнительно комбинации BERT с ритмическими характеристиками.

Классификация проводилась двумя способами:

- мультиклассовая классификация на пять жанров;
- бинарная классификация для каждого жанра, когда тексты классифицировались на принадлежащие и не принадлежащие конкретному жанру.

Для обоих способов применялись одни и те же классификаторы, включающие в себя два стандартных классификатора машинного обучения и две нейронные сети:

- классификатор AdaBoost — мета-алгоритм машинного обучения, который объединяет результаты 50 классификаторов-деревьев решений, корректирующих неправильно классифицированные тексты;
- классификатор RandomForest — мета-алгоритм машинного обучения, который усредняет результаты 50 классификаторов-деревьев решений;
- 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.

Данные алгоритмы были выбраны как лучшие алгоритмы по классификации текстов как с помощью ритма [5], так и с помощью ELMo [7] и BERT [8].

Для классификации корпус был разделён случайным образом на обучающую и тестовую выборки в отношении 4:1, чтобы провести пятикратную кросс-валидацию. Оценка качества выполнялась с помощью трёх стандартных мер: точность, полнота и F-мера. Во всех случаях стандартное отклонение для всех мер не превысило 4 % для алгоритма RandomForest и 2 % для остальных классификаторов, что свидетельствует о высокой стабильности классификации по жанрам.

Мультиклассификация жанров

Результаты мультиклассовой классификации представлены в таблице 1. Обе модели на основе эмбеддингов превосходят ритмические характеристики на 20–30 % по точности, полноте и F-мере. Лучшей по качеству моделью можно назвать BERT, так как она в комбинации с LSTM практически безошибочно различает жанры (99 % F-меры). Комбинация эмбеддингов с ритмом не повышает результаты классификации, что проиллюстрировано строками BERT + Ритм.

Модель ELMo практически так же хороша, как и BERT: 95–96 % F-меры, что всего на 3–4 % меньше лучших результатов.

Table 1. Multi-class text classification by genres

Таблица 1. Мультиклассовая классификация текстов по жанрам

Классификатор	Модель	Точность	Полнота	F-мера
AdaBoost	Ритм	64.6	62.8	63.7
AdaBoost	ELMo	84.9	84.8	84.8
AdaBoost	BERT	91.9	91.1	91.5
AdaBoost	BERT + Ритм	90.5	88.7	89.6
RandomForest	Ритм	63.2	57.2	60.0
RandomForest	ELMo	89.3	89.4	89.4
RandomForest	BERT	**94.9**	**94.8**	**94.8**
RandomForest	BERT + Ритм	94.3	93.6	93.9
LSTM	Ритм	67.8	67.3	67.6
LSTM	ELMo	96.6	96.6	96.6
LSTM	BERT	**99.2**	**99.2**	**99.2**
LSTM	BERT + Ритм	**99.3**	**99.2**	**99.3**
GRU	Ритм	68.7	67.8	68.2
GRU	ELMo	95.3	95.3	95.3
GRU	BERT	**98.7**	**98.7**	**98.7**
GRU	BERT + Ритм	**98.6**	**98.6**	**98.6**

Что касается классификаторов, нейронные сети LSTM и GRU одинаково обеспечивают высокий уровень классификации до 98–99 %. Среди стандартных классификаторов RandomForest превосходит AdaBoost на 3–4 % и достигает почти 95 % F-меры, что также является очень хорошим результатом, лишь немногим ниже уровня нейросетей.

Ритмические характеристики не обеспечивают высокого качества для пяти жанров: лучшие результаты с классификатором GRU достигают всего лишь 68 % F-меры.

Ошибки мультиклассификации для BERT и ритмических характеристик представлены в виде тепловых карт на рис. 5. Тепловые карты 5a и 5b указывают, сколько текстов из жанра в строке были приняты за жанр в столбце, а в последнем столбце указано общее количество ошибок для каждого жанра. Тепловые карты 5c и 5d демонстрируют те же данные, но в процентном соотношении относительно общего числа ошибок. На главной диагонали условно указаны нули.

При классификации с помощью BERT все романы отнесены правильно к своему жанру, а девять новостей классифицированы как другие жанры. Пять новостей и постов Вконтакте приняты за отзывы. За научную статью ошибочно не принят ни один текст.

При классификации с помощью ритмических характеристик романы классифицируются лучше всего: 19 ошибок, 52 % которых отнесены к новостям из OpenCorpora. Новости классифицируются с наибольшим количеством ошибок, 152, причём тексты этого жанра путаются со всеми остальными жанрами. Среди научных статей тоже много ошибок, 142, и это преимущественно отнесение к жанру новостей (59 % ошибок). За новости из OpenCorpora принято наибольшее количество текстов из других жанров в процентном соотношении: 40–59 %.

Таким образом, наиболее качественно классифицируются романы, а наименее — новости из OpenCorpora. Тем не менее, эмбеддинги BERT классифицируют все жанры с высокой точностью, полнотой и F-мерой.

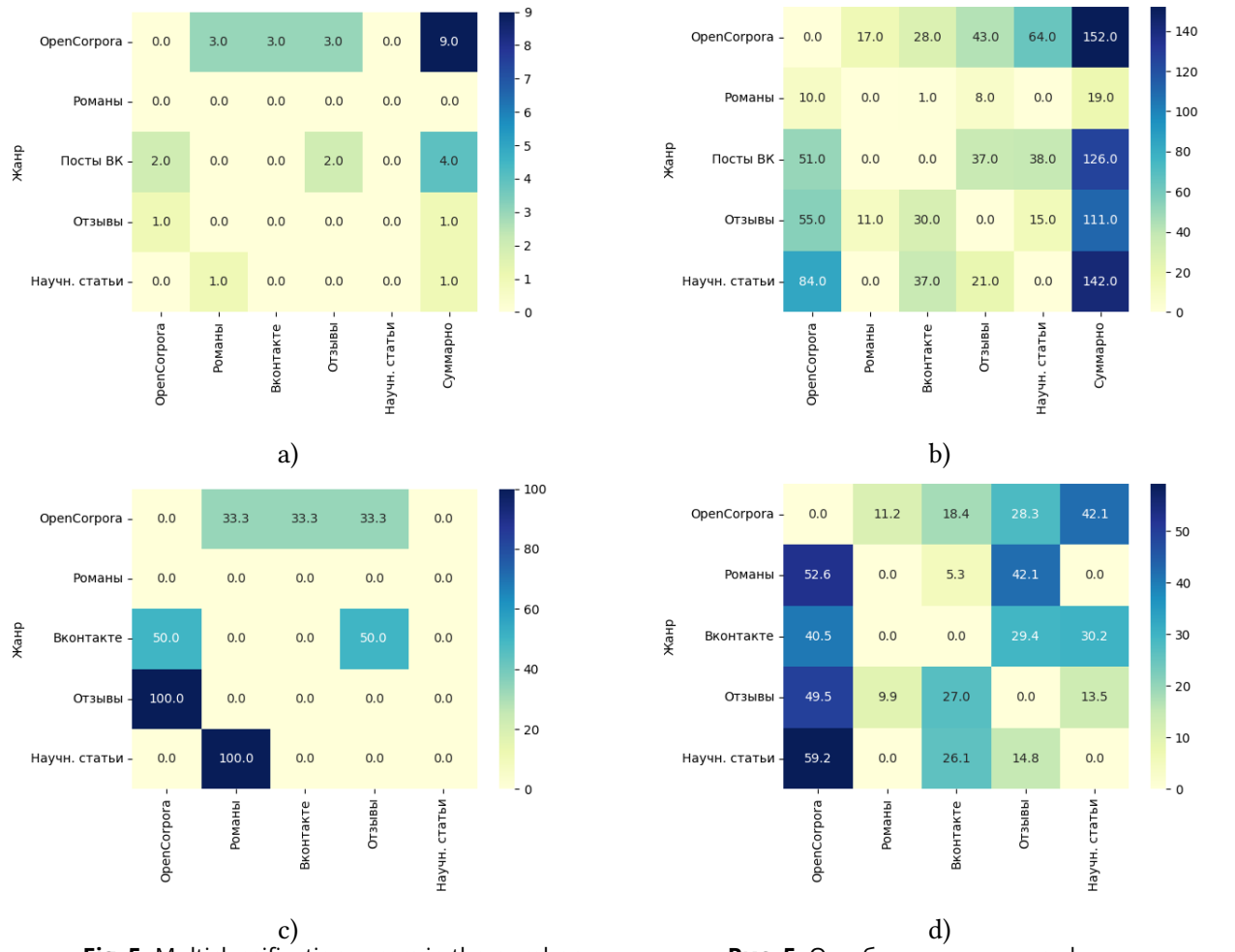


Fig. 5. Multiclassification errors in the number of texts for the a) BERT, b) rhythm, and in percent for the c) BERT, d) rhythm

Рис. 5. Ошибки мультиклассификации в количестве текстов для a) BERT, b) ритма, и в процентах для c) BERT, d) ритма

Верификация жанров

Для того, чтобы разобраться глубже в достоинствах, недостатках и ограничениях моделей текста, была проведена верификация жанров текстов для лучшей и худшей модели: эмбеддингов BERT и ритмических характеристик. Алгоритмом классификации была нейросеть LSTM. Её результаты представлены в таблице 2.

В целом, результаты верификации подтверждают результаты обработки ошибок: наиболее хорошо отделяются от остальных романы (96.0–99.8 % F-меры), наименее — новости из OpenCorpora (67.3–99.1 % F-меры). Это верно и для комплекса ритмических характеристик, и для эмбеддингов BERT. Кроме того, с высоким качеством верифицируются отзывы: 84.2–99.8 % F-меры.

BERT хорошо отделяет любой жанр от остальных. Комплекс ритмических характеристик лучше всего верифицирует романы и отзывы, а тексты OpenCorpora верифицируются слабее.

Чтобы определить набор жанров, для которых мультиклассификация с помощью ритма будет наиболее эффективной, из корпуса были исключены новости из OpenCorpora, а для остальных жанров была проведена мультиклассификация и верификация. Далее были исключены посты Вконтакте как худшие по верификации из четырёх жанров, и мультиклассификация с верификацией повторились. Любопытно, что среди трёх жанров: романы, отзывы, научные статьи, — наименьшее качество верификации оказалось у отзывов (85.3 % F-меры), а качество верификации научных статей

Table 2. Text verification by genres

Таблица 2. Верификация текстов по жанрам

Модель	Жанр	Точность	Полнота	F-мера
Ритм	OpenCorpora	73.9	61.8	67.3
Ритм	Романы	**95.0**	**97.0**	**96.0**
Ритм	Вконтакте	86.0	73.6	79.3
Ритм	Отзывы	**85.2**	**83.1**	**84.2**
Ритм	Научн. статьи	82.5	76.5	79.4
BERT	OpenCorpora	99.3	99.0	99.1
BERT	Романы	**99.6**	**99.9**	**99.8**
BERT	Вконтакте	99.6	99.4	99.5
BERT	Отзывы	**99.8**	**99.8**	**99.8**
BERT	Научн. статьи	99.7	99.7	99.7

оказалось вторым (92.3 % F-меры). И как завершение данного эксперимента, была проведена бинарная классификация для романов и научных статей.

Результаты каждой мультиклассификации и итоговой бинарной классификации приведены на рис. 6. На диаграмме по горизонтали указано количество классов, по вертикали — значения точности, полноты и F-меры. Уже исключение текстов OpenCorpora увеличивает качество классификации на 10 %, до 77 % для всех характеристик. А романы, научные статьи и отзывы вместе классифицируются отлично: 91.5 % F-меры. Романы и научные статьи классифицируются с ошибкой в одном тексте, когда фрагмент научной статьи принят за фрагмент романа. Стоит отметить, что в данном фрагменте было несколько цитат, написанных в литературном стиле.

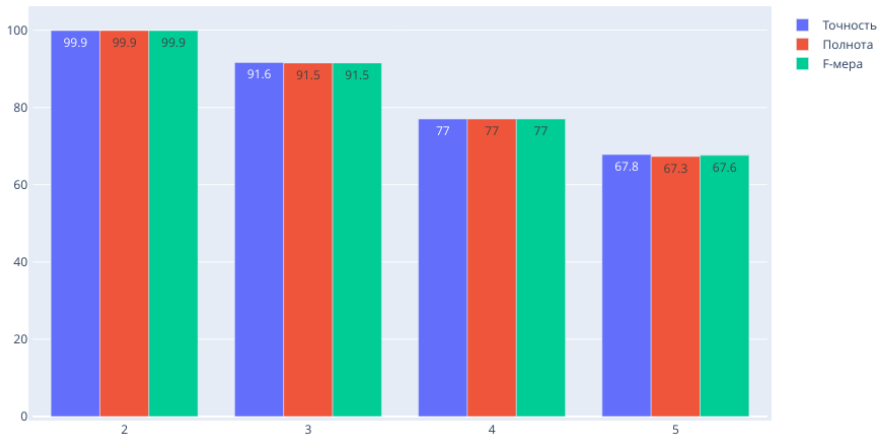


Fig. 6. Quality of multiclassification while reducing the number of classes

Рис. 6. Качество мультиклассификации при уменьшении количества классов

Таким образом, можно выделить три жанра: романы, научные статьи и отзывы, которые наиболее хорошо классифицируются с помощью комплекса ритмических характеристик. Наиболее вероятная причина таких результатов заключается в том, что в этих жанрах имеются общепринятые рекомендации к стилю письма, что может влиять и на особенности ритма. Новости и посты социальной сети Вконтакте более разнообразны по авторскому стилю.

Заключение

В данной статье автор классифицировала корпус из 10 000 русскоязычных текстов на пять жанров с помощью трёх векторных моделей: BERT, ELMo и комплекса ритмических характеристик. Лучшее качество было достигнуто с помощью комбинации эмбеддингов BERT с нейросетевым классификатором LSTM: 99 % F-меры как для мультиклассификации, так и для верификации отдельных жанров. Эмбеддинги ELMo показали близкий результат: 96 % F-меры.

Комплекс ритмических характеристик оказался полезен для романов, научных статей и отзывов на фильмы. Романы и отзывы наиболее разнообразны по ритмическим средствам среди всех жанров, как показали их визуализация и анализ. Кроме того, для данных жанров, как и для научных статей, существуют общепринятые рекомендации к их написанию, которые и могли повлиять на сходство стилей текстов в одном жанре. Новости OpenCorpora и посты Вконтакте гораздо более разнородны по стилю и тематике.

Результаты визуализации и классификации жанров для комплекса ритмических характеристик дают широкие возможности для их интерпретации с филологической точки зрения, что может быть направлением для будущих исследований. Также в дальнейшем можно рассмотреть и другие корпуса данных, включающие в себя большее количество Интернет-текстов различных жанров, чтобы протестировать современные эмбеддинги на более трудных задачах классификации длинных текстов.

References

[1] 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.

[2] B. Kessler, G. Numberg, and H. Schütze, “Automatic detection of text genre”, in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 1997, pp. 32–38.

[3] 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.

[4] Z. Dai and R. Huang, “A Joint Model for Structure-based News Genre Classification with Application to Text Summarization”, in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 3332–3342.

[5] 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.

[6] 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.

[7] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep Contextualized Word Representations”, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 2227–2237.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “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, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[9] C. Wang, P. Nulty, and D. Lillis, "A comparative study on word embeddings in deep learning for text classification", in *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, 2020, pp. 37–46.

[10] Y. Kuratov and M. Arkhipov, "Adaptation of deep bidirectional multilingual transformers for Russian language", in *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, 2019, pp. 333–339.

[11] A. Kutuzov, L. Pivovarova, *et al.*, "RuShiftEval: a shared task on semantic shift detection for Russian", in *Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference "Dialogue" (2021)*, vol. 20, 2021, pp. 533–545.

[12] J. Rodina, Y. Trofimova, A. Kutuzov, and E. Artemova, "ELMo and BERT in semantic change detection for Russian", in *International Conference on Analysis of Images, Social Networks and Texts*, Springer, 2020, pp. 175–186.

[13] A. V. Glazkova, "Topical classification of text fragments accounting for their nearest context", *Automation and Remote Control*, vol. 81, no. 12, pp. 2262–2276, 2020.

[14] 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.

[15] V. Bocharov, S. Alexeeva, D. Granovsky, E. Protopopova, M. Stepanova, and A. Surikov, "Crowdsourcing morphological annotation", in *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". Volume 1*, 2013, pp. 109–114.

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