Научная статья

УДК 81’42
DOI 10.25205/1818-7935-2026-24-1-74-86

Определение особенностей авторского стиля поэтов «озерной школы» методами машинного обучения

Владимир Борисович Барахнин[1,2]
Чжао Чжэньдун[2,3]
Елена Павловна Мачикина[1,4]

[1]Федеральный исследовательский центр информационных и вычислительных технологий
Новосибирск, Россия

[2]Новосибирский государственный университет
Новосибирск, Россия

[3]Московский государственный университет им. М. В. Ломоносова
Москва, Россия

[4]Сибирский государственный университет телекоммуникаций и информатики
Новосибирск, Россия

bar@ict.nsc.ru; https://orcid.org/0000-0003-3299-0507
c.chzhao2@alumni.nsu.ru
machikina@sibguti.ru; https://orcid.org/0009-0008-5512-4760

Аннотация

Статья посвящена исследованию стилистических особенностей поэзии представителей «озерной школы». Методология исследования включала в себя несколько этапов. Первоначально был сформирован англоязычный корпус текстов, состоящий из 99 поэтических произведений трех авторов, принадлежащих к данному литературному направлению. На данном этапе была осуществлена предварительная обработка текстов и извлечение релевантных признаков. Далее для представления поэтических текстов была применена модель пространства текстовых векторов. С целью анализа и идентификации уникальных стилистических характеристик каждого поэта была использована матрица признаков, проанализированная с применением пяти распространенных алгоритмов машинного обучения, включая метод опорных векторов (SVM) и случайный лес (Random Forest). Для решения проблемы несбалансированности данных была применена техника оверсэмплинга SMOTE (Synthetic Minority Oversampling Technique), что позволило повысить точность классификации. В частности, модифицированная модель SVM достигла F1-меры, превышающей 90 %, в задаче распознавания авторства поэтических текстов. На заключительном этапе исследования было проведено сопоставление комплексных стилистических характеристик поэзии трех авторов, что позволило выявить значимые различия в их индивидуальных стилях. Результаты исследования предоставляют ценную информацию для дальнейшего изучения в области анализа стилистических особенностей поэзии.

Ключевые слова

стилометрия, машинное обучение, классификация, идентификация авторства, технология SMOTE

Финансирование

Исследование выполнено в рамках государственного задания Минобрнауки России для Федерального исследовательского центра информационных и вычислительных технологий.

Для цитирования

Барахнин В. Б., Чжэньдун Ч., Мачикина Е. П. Определение особенностей авторского стиля поэтов «озерной школы» методами машинного обучения // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2026. Т. 24, № 1. С. 74–86. DOI 10.25205/1818-7935-2026-24-1-74-86

Determination of the «Lake School» Author Style Features by Machine Learning Methods

Vladimir B. Barakhnin[1,2]
Zhao Zhendong [2,3]
Elena P. Machikina[1,4]

[1]Federal Research Center for Information and Computational Technologies
Novosibirsk, Russian Federation

[2]Novosibirsk State University
Novosibirsk, Russian Federation

[3]Moscow State University named after M. V. Lomonosov
Moscow, Russian Federation

[4]Siberian State University of Telecommunications and Informatics
Novosibirsk, Russian Federation

bar@ict.nsc.ru; https://orcid.org/0000-0003-3299-0507
c.chzhao2@alumni.nsu.ru
machikina@sibguti.ru; https://orcid.org/0009-0008-5512-4760

Abstract

The article is devoted to the study of stylistic features of the Lake School poets. The research methodology included several stages. Initially, a corpus of texts was formed, consisting of 99 poetic works by three authors belonging to this literary trend. At this stage, preliminary text processing and extraction of relevant features were carried out. Further, a model of text vectors space was applied to represent poetic texts. In order to analyze and identify the unique stylistic characteristics of each poet, a feature matrix was applied, analyzed using five common machine learning algorithms, including the support vector machine (SVM) method and Random Forest. To solve the problem of data imbalance, SMOTE (Synthetic Minority Oversampling Technique) was applied, which improved classification accuracy. In particular, the modified SVM model achieved an F1 measure exceeding 90 % in the task of recognizing the authorship of poetic texts. At the final stage of the study, the complex stylistic characteristics of the poetry by the three authors were compared, which revealed significant differences in their individual styles. The results of the study provide valuable information for further research in the field of poetry stylistic features.

Keywords

stylometry, machine learning, classification, authorship identification, SMOTE technology

Funding

The research was carried out within the state assignment of the Ministry of Science and Higher Education, the Russian Federation, for the Federal Research Center for Information and Computational Technologies.

For citation

Barakhnin V. B., Zhao Zhendong, Machikina E. P. Determination of the «Lake School» author style features by machine learning methods. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2026, vol. 24, no. 1, pp. 74–86. DOI 10.25205/1818-7935-2026-24-1-74-86

Введение

Стилометрия – область компьютерной лингвистики, связанная с количественным анализом стилистических особенностей текста для определения авторства, датировки, атрибуции и других характеристик литературных произведений, связанных со стилем автора. Основой

анализа текстов является извлечение количественно измеримых характеристик текста, таких как энтропия и информационный объем, частота и длина слов, что обеспечивает объективное описание стилистических особенностей текста. С развитием методов машинного обучения возможности стилистического анализа литературных произведений значительно расширились, позволяя, например, определять авторство анонимных текстов и выявлять стилистические различия между авторами, работающими в рамках одной тематики. Однако существенная часть исследований фокусируется на длинных прозаических формах, таких как романы и статьи, в то время как традиционная поэзия остается менее изученной. Настоящая работа представляет собой попытку применения методов машинного обучения для идентификации стилевых особенностей поэзии представителей «озерной школы».

«Озерная школа», сформировавшаяся в конце XVIII – начале XIX века в Северо-Западной Англии, объединила группу поэтов-романтиков, в том числе Уильяма Вордсворта (William Wordsworth, 1770–1850), Роберта Саути (Robert Southey, 1774–1843) и Сэмюэля Тейлора Кольриджа (Samuel Taylor Coleridge, 1772–1834). Поэзия этих авторов характеризуется глубокой любовью к природе и повышенным вниманием к человеческим чувствам. Их стилистическая манера, отличающаяся свободой, лаконичностью и ясностью, оказала значительное влияние на последующее развитие поэзии. Произведения поэтов «озерной школы» не только занимают важное место в истории литературы, но и отражают идеологические и культурные тенденции социально-политического контекста эпохи, что обусловливает актуальность их дальнейшего изучения.

1. Обзор литературы и постановка задачи

Идея о возможности количественного анализа стиля текстов возникла давно. Еще в XIX веке исследователи обращали внимание на частотность использования определенных слов и фраз, а также на особенности рифмы и метра, но без строгой математической базы. В настоящий момент в стилометрических исследованиях литературных текстов можно выделить два основных направления: статистические исследования различных признаков текстов и изучение текстов с использованием методов машинного обучения. Оба направления активно развиваются. Например, в статье [Андреев, 2003] рассматривалась классификация стихотворных текстов американских поэтов-романтиков на основе множества разноуровневых признаков при помощи дискриминантного анализа, а в другой работе [Андреев, 2022] автор использует методы статистики для получения объективных индексов, позволяющих дать оценку тексту автора и его стилю. В работе [Oakes, 2018] для сопоставления стилей различных авторов использована модель *N*-грамм, основанная на предположении о зависимости появления *N*-го слова только от предыдущих *N*-1 слов. Исследователи подсчитали 5 000 наиболее часто встречающихся 4-грамм и провели сравнение текстов Льюиса с текстами других авторов. Применительно к поэтическим текстам, исходя из выделенных Ю. М. Лотманом [Лотман, 1998] уровней парадигматики поэтического текста, в статье [Барахнин, 2012] подчеркивается необходимость анализа значительного объема стихотворных текстов для изучения влияния структуры низшего уровня (фонетика, ритм, рифма, лексика) на структуру высшего уровня (тематика, литературный жанр). Предлагаются методы и техники для проведения комплексного анализа русской поэзии, при этом каждый уровень структуры стиха рассматривается как отдельная семантическая единица, взаимодействующая с другими характеристиками. В статье [Орехов, 2021] был применен метод Delta Берроуза для анализа особенностей стиля текстов романов В. Набокова и их переводов.

Имеется весьма немного исследований по атрибуции авторства и стилометрическому анализу поэтических текстов с применением методов машинного обучения и статистического анализа, например [Plecháč, 2018; Timofeeva, 2021].

Если говорить о применении для решения задач стилометрии современных методов машинного обучения, то можно отметить следующие работы.

В [Romanov, 2021] отмечается возрастающая значимость задачи идентификации авторства текстов в контексте цифровизации общества и переноса коммуникативной активности в онлайн-среду. Авторы применяли методы машинного обучения, включая метод опорных векторов (SVM), архитектуры глубокого обучения (долгая краткосрочная память (LSTM), сверточные нейронные сети (CNN) с механизмом внимания) и Transformer для определения авторства русскоязычных текстов. Была достигнута высокая точность классификации: SVM показал 96 %, а глубокие нейронные сети – 93 %. Кроме того, была проведена оценка устойчивости моделей относительно попыток анонимизации текста. Результаты экспериментов продемонстрировали, что преднамеренные изменения, направленные на сокрытие авторства, существенно ухудшают качество работы метода SVM.

В [Abbasi, 2022] предложена новая система идентификации авторов, основанная на комбинации двух техник извлечения характеристик и использовании различных алгоритмов машинного обучения (RF, XGB, MLP, LR, Ensemble) для классификации авторов неструктурированных текстовых документов. Авторы статьи акцентируют внимание на обсуждении и сравнении механизмов идентификации и классификации авторства.

В [Barakhnin, 2022] продемонстрирована эффективность использования частоты знаков препинания и *N*-грамм слов в качестве признаков для классификации стихов русских поэтов. Наилучшие показатели классификации достигаются при комбинировании различных групп признаков с использованием ансамблевых методов, основанных на деревьях решений и логистической регрессии. Особое внимание уделяется значимости восклицательного знака как дифференцирующего признака.

Целью данного исследования является разработка и оценка методов машинного обучения для автоматической классификации стихотворений по авторам. Исследование направлено на создание модели, способной точно и надежно определять авторство стихотворения на основе анализа стилометрических особенностей поэтов «озерной школы» или решать задачу классификации текста по автору.

В следующих разделах статьи будут описаны основные этапы исследования, а именно: сбор и подготовка корпуса стихотворений; выбор признаков, характеризующих стиль автора; преобразование текстовых данных в числовой формат, пригодный для использования в моделях машинного обучения; выбор и обучение моделей машинного обучения; оценка производительности и анализ полученных результатов.

2. Сбор данных и предварительная обработка текстов

Для исследования стилистических особенностей поэтов «озерной школы» использованы все стихотворения небольшого англоязычного собрания, размещенные на веб-сайте www.poemtryfoundation.org. В табл. 1 приведено распределение количества стихотворений по авторам.

Таблица 1

Характеристики корпуса стихотворений

Table 1

Characteristics of the poetry corpus

Автор	Количество стихов	Доля, %	Общее количество слов
Вордсворт	59	59,6	25211
Кольридж	33	33,4	15600
Саути	7	7	2564

Отобранные стихотворные тексты подверглись предварительной обработке. Основные этапы предварительной обработки:

- Приведение теста к нижнему регистру. Такое преобразование гарантирует, что слова с символами разных регистров будут идентичными, и позволяет уменьшить сложность словаря.
- Удаление стоп-слов. На этом этапе происходит удаление стоп-слов, таких как «a», «an», «the», «and» и т. д. Исключение стоп-слов, как показано, например, в [Kaur, 2018] помогает снизить размерность пространства признаков и упростить структуру исследуемого текста.
- Удаление знаков препинания и специальных символов. При анализе характеристик автора знаки препинания обычно не содержат большого объема семантической информации, а некоторые специальные символы, такие как пробел и конец строки, могут увеличивать размерность при анализе текстовых особенностей. Однако в дальнейшем будет проводиться отдельный частотный анализ знаков препинания для выявления эмоциональной энтропии текстов.

Таким образом, после предварительной обработки получен набор текстовых файлов, пригодных для извлечения набора признаков текста, который будет использоваться при машинном обучении моделей классификации.

3. Извлечение признаков стихотворных текстов

В качестве набора признаков поэтического текста были выбраны статистика знаков препинания в стихотворении, общее количество слов и строк в стихотворении, средняя длина предложения и средняя длина стихотворения, лексические характеристики и частеречные характеристики. Таким образом, каждому стихотворному тексту должен быть сопоставлен вектор численных признаков $f = (f_1, f_2,..., f_{20})$. Общую совокупность векторов численных признаков корпуса поэтических текстов обозначим как S.

Теперь более подробно про процесс определения каждого признака текста. Предварительно тексты были приведены к нижнему регистру. Такое преобразование гарантирует, что слова с символами разных регистров будут идентичными.

Признаки $f_1, f_2,..., f_{12}$ содержат количество вхождений основных знаков препинания и специальных символов типа скобок и кавычек (!”(),-.:;?`´) в стихотворении. Использование знаков препинания обычно связано с семантикой, и предпочтения в их использовании различаются у разных поэтов. Например, из статистических данных мы видим, что у всех трех поэтов самым часто используемым знаком препинания является запятая. Однако, помимо этого, Вордсворт и Кольридж предпочитают использовать восклицательный знак, в то время как Саути чаще всего прибегает к использованию точки с запятой. Это может свидетельствовать о том, что его способы выражения эмоций не так прямолинейны, как у двух других поэтов, и что он предпочитает использовать метафорическое или косвенное выражение чувств. После определения признаков $f_1, f_2,..., f_{12}$ знаки препинания, скобки и кавычки удаляются из текста, определение остальных признаков происходит по тексту без знаков препинания, скобок и кавычек.

При анализе стихотворных текстов мы обнаружили, что длина стихов трех поэтов часто варьируется, их стили сильно различаются. Стихи, написанные Вордсвортом, обычно содержат от 30 до 200 строк, в то время как у Кольриджа встречаются как пяти-шестистрочные короткие стихи, так и длинные стихи, превышающие 600 строк. Стихи Саути, напротив, обычно содержат менее 100 строк. Поэтому общее количество слов и строк в стихах рассматриваются как характеристики поэтического стиля. Признак f_{13} содержит длину стихотворения или общее количество символов в тексте, т. е. общее количество символов в поэтическом тексте после удаления всех знаков препинания и специальных символов, включая пробелы и знаки

конца строки. Признак f_{15} содержит общее количество строк стихотворения, а f_{16} – общее количество слов в стихотворном тексте после удаления стоп-слов, знаков препинания и лишних пробелов.

Средняя длина предложения представляет собой более продвинутую статистическую характеристику, чем длина текста, и отражает сложность стихотворения. Средняя длина предложений и общая длина стихотворения в символах обычно обладают определенной стабильностью у разных стихотворений одного автора. Поэтому средняя длина предложения (признак f_{14}), рассчитанная на основе количества слов в предложении после удаления стоп-слов, знаков препинания и лишних пробелов, будет рассматриваться как одна из характеристик стиля поэта.

Богатство лексики хорошо отражает мастерство и стиль автора. Для измерения степени лексического разнообразия в тексте используется показатель Type-Token Ratio (TTR), который определяется как отношение числа различных слов (типов лексики) к общему числу слов (лексическим токенам) в тексте. Для вычисления показателя богатства языка f_{17} тексты подвергались токенизации, далее проводилась лемматизация всех слов и удаление стоп-слов. Для всех этапов применялась библиотека NLTK, для удаления стоп-слов использовался стандартный список стоп-слов для английского языка из библиотеки NLTK. В полученном тексте подсчитывалось количество уникальных лемм и вычислялся признак f_{17}.

Частеречные характеристики включают в себя частоту использования слов трех различных частей речи: глаголов, прилагательных и существительных (признаки f_{18}, f_{19}, f_{20} соответственно). Стиль письма различных авторов различен, и частота использования слов разных частей речи также различна, поэтому частеречные характеристики были включены в набор признаков текста. Частеречные характеристики определялись после лемматизации по следующему алгоритму:

- Проводилась токенизация текста
- Все слова приводились к леммам с помощью WordNetLemmatizer
- Выполнялась частеречная разметка через pos_tag()
- Подсчитывалось количество различных частей речи: существительных (теги NN*), глаголов (теги VB*), прилагательных (теги JJ*)

Фрагмент полученной матрицы S признаков стихотворных текстов представлен на рис. 1. Строки матрицы S соответствуют вышеперечисленным признакам, а столбцы – стихотворным текстам. Для стихотворных текстов указано только авторство, конкретные названия стихотворений опущены.

Далее полученные данные необходимо подвергнуть нормализации для использования в машинном обучении. Процедура нормализации позволяет отобразить значения данных в диапазоне от 0 до 1, устранить влияние различных масштабов признаков и повысить точность разрабатываемой модели. Для получения нормализованных значений $\tilde{f} = \left(\tilde{f}_1, \tilde{f}_2, ..., \tilde{f}_{20}\right)$ применяется формула $f_i = \dfrac{f_i - \min(f_i)}{\max(f_i) - \min(f_i)_i}$, где максимум и минимум берется среди всех значений компоненты f_{13}.

N-грамма – это базовый метод представления текста в обработке естественного языка, который использует разбиение текста на последовательности из N слов для захвата локальной информации в тексте. Например, во фразе «Солнце светит ярко» можно выделить две 2-граммы – «Солнце светит» и «светит ярко». Анализ использования поэтом N-грамм в тексте может отражать его предпочтения в использовании определенных слов или фраз. Обычно для анализа выбирают невысокие значения (N = 1, 2, 3), поскольку слишком большие значения N могут значительно увеличить объем словаря N-грамм.

Для определения частот N-грамм (N = 1, 2, 3) используется модель мешка слов (Bag of Words Model) и последовательность этапов обработки:

- Токенизация текста.

f	описание признака	Вордсв	Вордсв	Кольрид	Вордсво	Вордсвс	Вордсв	Кольрд	Саути	Вордсвор
f1	частота !	0	8	4	2	1	10	1	14	19
f2	частота "	0	0	0	0	0	0	0	0	31
f3	частота (	0	0	0	0	0	0	0	0	2
f4	частота)	0	0	0	0	0	0	0	0	2
f5	частота ,	6	27	14	5	29	33	10	72	115
f6	частота –	1	3	0	0	3	2	2	9	12
f7	частота .	0	2	4	4	3	6	3	10	30
f8	частота :	2	1	2	1	1	8	1	4	19
f9	частота ;	1	3	0	4	12	16	2	2	39
f10	частота ?	0	2	1	0	0	0	1	0	9
f11	частота ´	0	0	0	0	0	0	0	0	0
f12	частота `	0	0	0	0	0	0	0	0	0
f13	длина стихотворения	274	1386	628	481	977	1953	591	3082	7166
f14	средняя длина предложения	5,50	6,89	6,67	5,22	5,71	6,03	7,4	6,43	6,99
f15	общее количество строк	8	33	15	17	31	57	15	76	175
f16	общее количество слов	44	262	120	94	177	374	111	566	1279
f17	богатство языка	0,435	0.363	0,381	0,415	0,382	0,32	0,414	0,3	0,282
f18	количество глаголов	7	45	17	22	22	58	19	83	224
f19	количество прилагательных	6	22	11	11	18	40	12	67	158
f20	количество существительны	17	72	40	25	57	105	26	154	381

Рис. 1. Матрица *S* признаков стихотворных текстов
Fig. 1. Matrix *S* of poetic text features

- Создание мешка слов, который включает все слова и сочетания, встречающиеся во всех стихотворениях трех поэтов, чтобы гарантировать уникальный индекс для каждой *N*-граммы.
- По полученному мешку слов и сочетаний формируется вектор, длина которого равна размеру мешка слов и сочетаний, каждая позиция соответствует слову или сочетанию в словаре, значение соответствующей позиции равно метрике TF-ID для каждого стихотворения из коллекции.

Вычисление метрики TF-ID для *N*-граммы является широко распространенным статистическим методом обработки текста на естественном языке. Показатель TF-IDF оценивает значимость слова в стихотворении на основе данных о всей коллекции стихотворений. Важность слова (сочетания слов) пропорциональна частоте его появления в стихотворном тексте, но обратно пропорциональна частоте его появления в корпусе текста.

Для *N*-граммы t и стихотворения d метрика TF-IDF определяется как $TF - IDF(t, d) = TF(t, d) \cdot IDF(t)$, где $TF(t, d)$ – частота *N*-граммы t в стихотворении d, а $IDF(t)$ – логарифм доли стихотворений, где встречается *N*-грамма t, среди всех стихотворений корпуса Известно, что больший вес в метрике TF-IDF получают слова (комбинации слов), которые являются наиболее характерными в рамках одного документа корпуса, поскольку обладают высокой частотой в пределах конкретного документа и низкой частотой употреблений в других документах корпуса.

Наборы векторов с показателями TF-IDF для всех слов, биграмм и триграмм для каждого стихотворения из корпуса текстов обозначим G1, G2, G3 соответственно. После проведенной описанной выше обработки данных получены нормализованная матрица *S* признаков текстовых векторов, включающая структурные, лексические и грамматические особенности каждого стихотворения, а также матрицы признаков G1, G2, G3, содержащие информацию о значимости *N*-грамм для каждого стихотворения.

3. Обучение и оценка результатов классификации текстов

Для решения поставленной задачи предлагается использовать пять распространенных алгоритмов классификации текста: метод опорных векторов (SVM), случайный лес (RF), логистическую регрессию (LR), наивный байесовский классификатор (NB) и многослойный персептрон (MLP), чтобы обучить и оценить модели. Обучение и оценивание качества моделей будет проводиться на извлеченных данных трех типов:

- нормализованная матрица признаков стихотворений (матрица *S*);
- матрицы *N*-грамм, взвешенных по TF-IDF (матрицы G1, G2, G3);
- комбинация матриц признаков (S + G1, G2, G3) для получения общей матрицы признаков авторского стиля.

Поскольку рассматриваемый корпус текстов не очень большой, то логично использовать метод кросс-валидации для анализа извлеченных трех матриц признаков. В этом методе данные разбиваются на пять одинаковых по размеру подмножеств. На каждой итерации обучения одно из подмножеств остается для тестирования, а остальные четыре подмножества используются для обучения модели. Этот процесс повторяется пять раз, причем каждый раз для тестирования выбирается подмножество, ранее не использованное. Таким образом, каждый образец данных используется как для обучения, так и для тестирования модели, что исключает зависимость от конкретного разбиения данных.

Для каждого выбранного алгоритма машинного обучения происходит обучение соответствующей модели классификатора на обучающем наборе данных. Для моделей метода опорных векторов (SVM), случайного леса (RF), логистической регрессии (LR) и многослойного персептрона (MLP) применяется техника поиска по сетке для настройки гиперпараметров с целью оптимизации производительности модели.

На тестовом наборе данных проверяются три оценки качества классификации каждой из моделей – точность (precision), полнота (recall) и оценка F1 (F1 Score). Под точностью понимается, какая доля из предсказанных моделью положительных результатов составляет истинные положительные результаты, т. е. precision = TP / (TP + FP), где TP – истинно положительные результаты (True Positives), FP – ложно положительные результаты (False Positives). Чем выше точность, тем выше доля истинных положительных результатов среди результатов, которые модель предсказывает как положительные. Полнота (recall) измеряет, какая доля истинно положительных образцов среди всех реальных положительных образцов успешно предсказана моделью, т. е. recall = TP / (TP + FN). Чем выше значение полноты, тем правильнее модель способна определять реальные положительные образцы. Оценка F1 (F1-score) учитывает как точность, так и полноту, и представляет собой гармоническое среднее между точностью и полнотой. F1-оценка объединяет оценки точности (Precision) и полноты (Recall) для модели, и более точно оценивает производительность модели на несбалансированных наборах данных при решении задач классификации. Анализ и сравнение метрики качества моделей каждого алгоритма на тестовом наборе данных позволяет определить классификатор, который наилучшим образом подходит для поставленной задачи.

Оценки качества классификации стихотворений с использованием кросс-валидации на векторах признаков стихов *S* приведены в табл. 2

Из результатов классификации видно, что логистическая регрессия и наивный байесовский классификатор показывают более высокие результаты, превышающие 70 % точности, а остальные три классификатора также достигают более 60 % точности, что говорит об относительной точности выбора стилевых характеристик поэтов.

В табл. 3 показаны результаты классификации лучшими моделями по матрицам G1, G2, G3.

Таблица 2

Оценки качества классификации (SVM, RF, LR, Naive Bayes, MLP)

Table 2

Classification quality assessments (SVM, RF, LR, Naive Bayes, MLP)

Модель	precision	recall	F1 score
SVM	0,608	0,633	0,5854
RF	0,505	0,623	0,553
LR	0,681	0,733	0,7057
Naive Bayes	0,688	0,7	0,649
MLP	0,634	0,665	0,637

Таблица 3

Результаты классификации (1-gram, 2-gram, 3-gram)

Table 3

Classification results (1-gram, 2-gram, 3-gram)

N-gram	лучшие модели	precision	recall	F1 score
1-gram	LR	0,444	0,667	0,533
2-gram	SVM	0,593	0,7	0,607
3-gram	SVM	0,604	0,739	0,665

Из оценок, приведенных в таблице, видно, что результаты классификации для N-грамм не идеальны. При анализе фактических результатов классификатора было обнаружено, что поэт Вордсворт имеет гораздо больше стихов, чем два других поэта, что составляет более 70 % от всего набора данных. Это приводит к крайне неравномерному распределению данных, и классификатор легко ошибается, относя стихи других двух поэтов к поэзии Вордсворта. Это также приводит к плохим результатам классификации для двух матриц признаков, поэтому для решения проблемы неравновесия в наборе данных применен метод пересэмплирования SMOTE. SMOTE (Synthetic Minority Over-sampling Technique) [Chawla, 2002] – это широко используемая техника пересэмплирования данных, предназначенная для решения проблемы несбалансированных наборов данных. Метод SMOTE улучшает производительность моделей машинного обучения путем балансировки набора данных за счет генерации синтетических образцов для миноритарного класса. SMOTE идентифицирует образцы миноритарного класса и генерирует вокруг них новые синтетические образцы, увеличивая количество образцов миноритарного класса и приближая его к количеству образцов мажоритарного класса, что помогает снизить предвзятость модели в пользу образцов мажоритарного класса при прогнозировании, улучшая производительность и обобщающую способность модели.

Для лучшей модели SVM проводится тестирование на новом объединенном наборе данных (S + G1, G2, G3), обработанном с использованием SMOTE и оптимизированном после поиска по сетке. Результаты тестирования приведены в табл. 4.

Путем создания матрицы ошибок для 1-грамм и 3-грамм, которые показали лучшие результаты классификации, было выявлено, что при использовании 1-грамм два стихотворения Вордсворта были ошибочно отнесены к Кольриджу, в то время как в стихах Кольриджа четыре были ошибочно отнесены к Вордсворту. При использовании 3-грамм одно стихотворение

Вордсворта было неправильно классифицировано как произведение Кольриджа, и для двух стихотворений Кольриджа неправильно классифицированы как авторство Вордсворта.

Таблица 4

Результаты классификации по объединенным наборам данных (S + G1, G2, G3)

Table 4

Classification results for the combined data set

	N-gram	Wordsworth	Southey	Coleridge
precision	1-gram	0,80	1,00	0,87
	2-gram	0,72	0,95	0,69
	3-gram	0,89	1,00	0,94
recall	1-gram	0,89	1,00	0,78
	2-gram	0,72	1,00	0,61
	3-gram	0,94	0,88	1,00
f1-score	1-gram	0,84	1,00	0,81
	2-gram	0,72	0,97	0,67
	3-gram	0,92	0,91	1,00

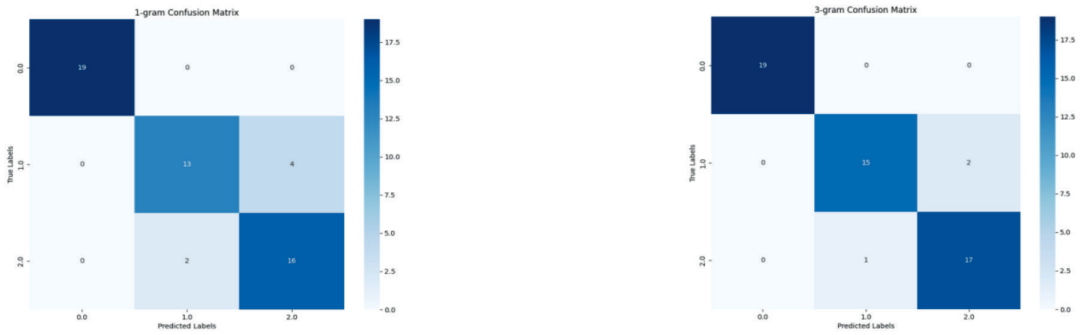

Рис. 2. Матрица ошибок (2,0 : Вордсворт 1,0: Кольридж 0,0 Саути)
Fig. 2. The Error Matrix (2.0 : Wordsworth 1.0: Coleridge 0.0 Southey)

В то же время стихотворения Саути успешно классифицированы на разных N-граммах. Эти результаты указывают на то, что улучшенный набор данных чувствителен к стихам Саути, и все стихи были успешно идентифицированы (включая «ложные стихи», созданные с помощью технологии SMOTE). Стиль отдельных стихотворений Вордсворта и Кольриджа проявляет определенную схожесть, однако в целом стиль всех трех поэтов успешно распознан.

Заключение

Это исследование использует пять распространенных алгоритмов машинного обучения для распознавания стилевых характеристик поэтов «озерной школы». Затем мы применяем технику SMOTE для балансировки набора данных и применяем новый набор данных к ранее

хорошо себя зарекомендовавшей модели опорных векторов (SVM). В лучшей группе признаков 3-грамм точность и f1-мера превышают 90 %, что свидетельствует о превосходстве модели.

В следующей работе мы попытаемся оценить важность каждого отдельного признака для распознавания стиля поэтов и исследовать новые признаки, такие как ритм и служебные слова, для более точного отображения характеристик стиля авторов.

Список литературы

Андреев В. С. Классификация стихотворных текстов методом дискриминантного анализа // Математическая морфология: электронный математический и медико-биологический журнал. 2003. Т. 5, № 1. С. 58–70.

Андреев В. С. Статистическая оценка динамики стиля: лирика В. Набокова в зеркале частей речи // Известия Смоленского государственного университета. 2022. Т. 60, № 4. С. 81–91. DOI: 10.35785/2072-9464-2022-60-4-81-91

Барахнин В. Б., Кожемякина О. Ю. Об автоматизации комплексного анализа русского поэтического текста // CEUR Workshop Proceedings, Pereslavl-Zalessky, 15–18 октября 2012 года. С. 167–171.

Лотман Ю. М. Структура художественного текста // Лотман Ю. М. Об искусстве. – СПб.: Искусство– СПБ, 1998. С. 14–288.

Орехов Б. В. Текст и перевод Владимира Набокова через призму стилеметрии // Новый филологический вестник. 2021. Т. 58, №3. С. 200–213.

Abbasi A., Javed A.R., Iqbal F., Jalil Z., Gadekallu T.R., Kryvinska N. Authorship identification using ensemble learning. // Sci Rep. 2022. 12, 9537. DOI: 10.1038/s41598-022-13690-4

Barakhnin V., Kozhemyakina O., Grigorieva I. Determination of the Features of the Author's Style of A.S. Pushkin's Poems by Machine Learning Methods. //Appl. Sci. 2022. 12, 1674. DOI: 10.3390/app12031674

Chawla N. V., Bowyer K., Hall L. O., Kegelmeyer W. P. SMOTE: Synthetic Minority Oversampling Technique // Journal of Artificial Intelligence Research. 2002. Vol. 16, No. 1. P. 321–357. DOI:10.1613/jair.953

Kaur J., Buttar P.K. Stopwords Removal and its Algorithms Based on Different Methods // Int. J. of Adv.Research in Comp.Sci. 2018. Vol. 9. No. 5. P. 80–88. DOI: 10.26483/ijarcs.v9i5.6301

Oakes M. P., Computer stylometry of C. S. Lewis's The Dark Tower and related texts // Digital Scholarship in the Humanities. 2018. Vol. 33. No. 3. P. 637–650. DOI: 10.1093/llc/fqx043

Plecháč P.; Bobenhausen K.; Hammeric B. Versification and authorship attribution. A pilot study on Czech, German, Spanish, and English poetry // Studia Metr. Poet. 2018, Vol. 5. No. 12. P. 29–54. DOI: 10.12697/smp.2018.5.2.02

Romanov A., Kurtukova A., Shelupanov A., Fedotova A., Goncharov V. Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks. //Future Internet. 2021. Vol. 13. No. 1. DOI:10.3390/fi13010003

Timofeeva M. Comparative Analysis of Reasoning in Russian Classic Poetry //Appl. Sci. 2021. 11, 8665. DOI: 10.3390/app11188665

References

Andreev V. S. Classification of Verse Texts by Means of Discriminant Analysis. *Matematika morfologio elektrona matematika kaj medicina-biologia revuo*. 2003, vol. 5, no. 1, pp. 58–70. (in Russ)

Andreev V. S. Statistical Evaluation of Style Dynamics: Lyrical Poems by V. Nabokov in the Mirror of Parts of Speech. *Izvestia of Smolensk State University*, 2022, vol. 60, no. 4, pp. 81–91. DOI: 10 .35785/2072-9464-2022-60-4-81-91 (in Russ.)

Barakhnin V., Kozhemyakina O. About the Automation of the Complex Analysis of Russian Poetic Text. *CEUR Workshop Proceedings*, 2012, vol. 934, pp. 167–171. (in Russ.)

Lotman Yu. M. The Structure of the Artistic Text. In: *Lotman Yu. M. About Art*. St. Petersburg: Iskusstvo-SPB, 1998, pp. 14–288. (in Russ.)

Orekhov B. V. Text and Translation by Vladimir Nabokov through the Prism of Stylemetry. *The New Philological Bulletin*, 2021, vol. 58, no. 3, pp. 200–213. (in Russ.)

Abbasi A., Javed A. R., Iqbal F., Jalil Z., Gadekallu T. R., Kryvinska N. Authorship Identification Using Ensemble Learning. Sci Rep, 2022, vol. 12, p. 9537. DOI: 10.1038/s41598-022-13690-4

Barakhnin V., Kozhemyakina O., Grigorieva I. Determination of the Features of the Author's Style of A. S. Pushkin's Poems by Machine Learning Methods. *Appl. Sci*., 2022, vol. 12, p. 1674. DOI: 10.3390/app12031674

Chawla N. V., Bowyer K., Hall L. O., Kegelmeyer W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 2002, vol. 16, no. 1, pp. 321–357. DOI: 10.1613/jair.953

Kaur J., Buttar P. K. Stopwords Removal and Its Algorithms Based on Different Methods. *Int. J. of Adv. Research in Comp. Sci*., 2018, vol. 9, no. 5, pp. 80–88. DOI: 10.26483/ijarcs.v9i5.6301

Oakes M. P. Computer Stylometry of C. S. Lewis's The Dark Tower and Related Texts. *Digital Scholarship in the Humanities*, 2018, vol. 33, no. 3, pp. 637–650. DOI: 10.1093/llc/fqx043

Plecháč P., Bobenhausen K., Hammeric B. Versification and Authorship Attribution: A Pilot Study on Czech, German, Spanish, and English Poetry. *Studia Metr. Poet*., 2018, vol. 5, no. 12, pp. 29–54. DOI: 10.12697/smp.2018.5.2.02

Romanov A., Kurtukova A., Shelupanov A., Fedotova A., Goncharov V. Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks. *Future Internet*, 2021, vol. 13, no. 1. DOI: 10.3390/fi13010003

Timofeeva M. Comparative Analysis of Reasoning in Russian Classic Poetry. *Appl. Sci*., 2021, vol. 11, p. 8665. DOI: 10.3390/app11188665

Информация об авторах

Барахнин Владимир Борисович, доктор технических наук, заведующий лабораторией информационных ресурсов Федерального исследовательского центра информационных и вычислительных технологий; заведующий кафедрой математического моделирования Новосибирского государственного университета

Чжао Чжэньдун, выпускник бакалавриата Новосибирского государственного университета; магистрант Московского государственного университета им. М. В. Ломоносова

Мачикина Елена Павловна, кандидат физико-математических наук, старший научный сотрудник Федерального исследовательского центра информационных и вычислительных технологий; доцент кафедры прикладной математики и кибернетики Сибирского государственного университета телекоммуникаций и информатики

Information about the Authors

Vladimir B. Barakhnin, Doctor of Technical Sciences, Head of the Laboratory of Information Resources at the Federal Research Center for Information and Computing Technologies, Head of the Department of Mathematical Modeling at Novosibirsk State University

Zhao Zhendong, Bachelor's Degree Graduate from Novosibirsk State, Master's Student at Moscow State University

Elena P. Machikina, Candidate of Physical and Mathematical Sciences, Senior Researcher Federal Research Center for Information and Computing Technologies, Associate Professor of the Department of Applied Mathematics and Cybernetics, Siberian State University of Telecommunications and Informatics

Статья поступила в редакцию 11.09.2025;
одобрена после рецензирования 26.11.2025; принята к публикации 28.11.2025

The article was submitted 11.09.2025;
approved after reviewing 26.11.2025; accepted for publication 28.11.2025