# Сравнительный анализ методов автоматической классификации поэтических текстов на основе лексических признаков

© В.Б. Барахнин © О.Ю. Кожемякина © И.С. Пастушков

Институт вычислительных технологий СО РАН,
Новосибирский государственный университет,
Новосибирск, Россия

bar@ict.nsc.ru olgakozhemyakina@mail.ru pas2shkov.ilya@gmail.com

**Аннотация.** Проанализированы принципы формирования обучающих выборок для алгоритмов определения стилей и жанровых типов. Проведены вычислительные эксперименты с использованием корпуса текстов лицейской лирики А.С. Пушкина по выбору наиболее точного алгоритма классификации поэтических текстов, в том числе с использованием наиболее известных приемов ансамблирования базовых алгоритмов в композиции, таких, как взвешенное голосование, бустинг и стекинг, причем в качестве характеристических признаков стихотворений использовались одиночные слов, биграммы и триграммы. Рассмотренные алгоритмы показали свою работоспособность и могут быть использованы для автоматизации комплексного анализа русских поэтических текстов, существенно облегчая работу эксперта при определении их стилей и жанров путем предоставления соответствующих рекомендаций.

**Ключевые слова:** автоматический анализ поэтических текстов, определение жанров и стилей, алгоритмы классификации.

# Comparative Analysis of Methods of Automated Classification of Poetic Texts Based on Lexical Signs

© V.B. Barakhnin © O.Yu. Kozhemyakina © I.S. Pastushkov

Institute of Computational Technologies of SB RAS,
Novosibirsk, Russia
Novosibirsk State University,
Novosibirsk, Russia

bar@ict.nsc.ru olgakozhemyakina@mail.ru pas2shkov.ilya@gmail.com

**Abstract.** In this paper we analyze the principles of formation of the training samples for the algorithms of the definition of styles and genre types. The computational experiments with a corpus of texts of Lyceum lyrics of A. S. Pushkin at the choice of the most accurate algorithm of classification of poetic texts were conducted, including the usage of the best-known methods of assembling of the basic algorithms in the composition, such as weighted voting, boosting and stacking, and as a characteristic feature of the poems the single words, bigrams and trigrams were used. The considered algorithms showed their efficiency and can be used to automate the complex analysis of Russian poetic texts, significantly facilitating the work of the expert in determining of their styles and genres by providing the appropriate recommendations.

**Keywords:** automated analysis of poetic texts, the definition of genres and styles, classification algorithms.

## 1 Введение

В задачах автоматизированного анализа текстов на естественном языке возникает проблема их классификации по жанрам и стилям, которые являются важными атрибутами, используемыми при определении влияния низших уровней стиха на высшие (см., например, [1]).

Исследования в области автоматизированного определения жанрового типа текстов начаты недавно – в начале 2010-х годов. Так, в работе [2] предложены алгоритмы определения жанров оды,

**Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года**

песни, послания, элегии и эпитафии на материале английских поэтов–сентименталистов XVIII века: поскольку «несмотря на то, что в XVIII–XIX веках жанровые признаки стихотворных текстов постепенно начинают теряться ..., в английской литературе начала XVIII века жанры оды, песни, послания, элегии и эпитафии по соотношению своих формальных признаков еще достаточно хорошо разграничиваются».

В [3] изложен метод классификации текстов (по определенным жанрам и по авторам) на основе анализа статистических закономерностей буквенных распределений, т. е. вероятностей встречаемости букв и буквосочетаний, при этом подчеркнуто, что решение найдено без «вторжения в область литературы, т. е. без анализа синтаксиса, литературных приемов и схем взаимодействий персонажей». Однако в работе [4] сами авторы строят оригинальный контрпример к статистическому методу идентификации, что показывает необходимость использования, по крайней мере, методов морфологического анализа. Что же касается автоматизации определения стилистических характеристик текстов, то нам неизвестны исследования в этой области, по крайней мере, для текстов на русском языке.

В работе [5] нами показано, что метод опорных векторов (support vector machine, SVM) [6] позволил получить хорошие результаты при определении стилистической окраски поэтических текстов и удовлетворительные – при определении жанров.

В настоящей работе мы расширили используемые подходы, в частности, учитывая при построении характеристического вектора используемых в стихотворении лексем количество их вхождений, а также проводя эксперименты с характеристическими векторами биграмм и триграмм. Кроме того, нами был проведен сравнительный анализ целого ряда алгоритмов классификации поэтических текстов, в том числе с использованием наиболее известных приемов ансамблирования, т. е. построения композиций алгоритмов, в которых ошибки отдельных алгоритмов взаимно компенсируются. При ансамблировании рассматриваются алгоритмы, в которых функция, называемая алгоритмическим оператором, устанавливает соответствие между множеством объектов и пространством оценок, а функция, называемая решающим правилом, устанавливает соответствие между пространством оценок и множеством значений целевой функции. Таким образом, рассматриваемые алгоритмы имеют вид суперпозиции алгоритмического оператора и решающего правила. Многие алгоритмы классификации имеют именно такую структуру: сначала вычисляются оценки принадлежности объекта классам, затем решающее правило переводит эти оценки в номер класса. Значением оценки может быть вероятность принадлежности объекта классу, расстояние от объекта до разделяющей поверхности, степень уверенности классификации и т. п.

Таким образом, в статье проведен сравнительный анализ целого ряда методов автоматизированной классификации поэтических текстов, включая наиболее известные приемы ансамблирования базовых алгоритмов в композиции: взвешенное голосование, бустинг и стекинг.

## 2 Построение обучающей выборки

Наиболее эффективным подходом к автоматизации определения жанровых типов и стилистических характеристик является использование алгоритмов с обучением. Однако формирование обучающей выборки является отнюдь не банальной задачей. Наша попытка использовать в качестве обучающей выборки пушкинскую лирику зрелого периода (1828–1831 гг.) потерпела неудачу уже на раннем этапе работы, поскольку жанровое разнообразие пушкинского творчества этого периода, соотносясь со стилевыми особенностями произведений в особой пушкинской манере, не следует общепринятым законам. На данную черту указывал ране В.А. Грехнев: «Жанры и стиль не противостоят друг другу как враждебные, взаимоотрицающие начала, но между ними всегда существует внутреннее напряжение. Напряжение это возрастает там, где возрастают мощь и размах писательской индивидуальности» [7. С. 234]. Отсюда возникают жанрово-стилистические разновидности и варианты, во «внутреннем напряжении» между стилем и жанром берут начало неканонические жанры, и именно для обучающей выборки это становится критичным, поскольку возникают особенности, не попадающие в систему, следовательно, противоречащие по своей сути материалу для построения жанрово-стилевой системы. Вследствие этого мы решили остановиться на лицейской лирике (1814–1817 гг.), поскольку в ней наблюдаются использование наиболее строгих жанровых форм, стилистическое единство, а также следование правилам грамматик своего времени: «Почти вся лицейская лирика относится к возвышенному стилю, исключение – всего несколько стихотворений. Даже многие сатирические стихи написаны вполне в возвышенном стиле. Можно утверждать, что в ранних стихах Пушкина чувствуется влияние жестких правил «Грамматики» его лицейского учителя Н.Ф. Кошанского» [8. С. 24].

В свою очередь, использование именно лицейской лирики, как материала для создания обучающей выборки, оправдано и стилевым аспектом, поскольку стилевая дифференциация лексем – этап разработки классификатора. Для текстов на русском языке принято восходящее к трудам М.В. Ломоносова [9] деление текстов (прежде всего, художественных) на относящиеся к высокому, среднему и низкому стилям. Исторически каждый из них характеризуется специфическим соотношением использования старославянских (церковнославянских) и собственно русских слов (при этом отдельно рассматривается группа слов,

общих для старославянского и русского языков), долей архаизмов, а также употреблением определенных синтаксических конструкций.

Для реализации поставленной задачи мы идем от практики, делая выборку произведений Пушкина лицейского периода, с 1813 по 1817 гг., как материала, на котором вероятно построение наиболее точной теоретической модели жанрово-стилистических зависимостей, что, несомненно, делает конечный результат анализа наиболее точным и позволяет разработать наиболее адекватный классификатор, относящийся к стилевому аспекту. Так как мы решили ограничиться анализом жанров только малых стихотворных форм, то из анализа исключены поэмы, сказки, переводы, Dubia, и далее делаем список, включающий в себя стихотворения, как соответствующие системе жанров, приведенной в монографии Д.М. Магомедовой [10], так и не входящие в эту систему. В итоге рассмотрения списка произведений, взятого нами для анализа, мы выделяем следующие группы жанров.

Канонические: ода – 4 произведения, элегия – 27 произведений (в том числе одна историческая элегия – «Наполеон на Эльбе»), идиллия – 2 произведения, послание – 55 произведений, баллада – 3 произведения, неканонические, выделенные Д.М. Магомедовой (фрагмент, рассказ в стихах) – их нет.

Также мы добавляем жанры, которых нет в разработанной Д.М. Магомедовой системе канонических–неканонических: эпиграмма – 18 произведений, мадригал – 4 произведения, сонет – 1 произведение, романс – 1 произведение, анекдот – 1 произведение, притча – 2 произведения. Кроме этого, стихотворение «Безверие» (1817) определяется как элегия и философская ода [11]. Но для анализа мы определяем его как философскую оду. Жанровые типы этих произведений легли в основу классификатора (см. табл. 1): по одной оси мы разместили жанровые типы – в порядке возрастания «возвышенности»: ода, элегия, идиллия, послание и т. д., а по другой оси – традиционные стили.

На данном эмпирическом материале просматривается очевидная корреляция между жанровыми и стилистическими характеристиками текстов: ода, элегия и идиллия обычно написаны высоким стилем, в них не используется лексика, соответствующая низкому стилю, а для эпиграмм, напротив, характерно использование элементов лексики низкого стиля. Вообще говоря, стиль текста определяется по наиболее «низким» его лексемам, что особенно характерно для эпиграмм: наличие высокой лексики, употребляемой нередко в ироническом ключе, не должно вводить в заблуждение, ибо употребление одного–двух слов разговорной или откровенно обсценной лексики сразу характеризует авторский замысел. Тем не менее, для жанров, традиционно предполагающих возвышенную форму, прежде всего, мадригала, мы не считаем целесообразным относить принадлежащие к ним стихотворения, в которых с ироническим целями употреблено несколько «сниженных» (но не обсценных!) слов, к сниженному стилю. Следует отметить, что специфика стиля проявляется на лексическом уровне в гораздо большей степени, чем жанр.

В нашей выборке в силу ее специфических задач произведения, написанные в жанре притчи, отнесены: одно («Наездники») – к высокому стилю, второе («Истина») – к среднему, хотя, как известно, притча, будучи жанром, наиболее близким к басне, предполагает возможность написания ее в разных стилях, о чем свидетельствует, в частности, притча Пушкина «Сапожник», которую можно отнести, скорее, к низкому («разговорному») стилю.

**Таблица 1** Статистика по жанрово-стилевому соответствию

| | Высок. | Средн. | Низк. |
|---|---|---|---|
| Ода | 4 | - | - |
| Притча | 1 | 1 | - |
| Мадригал | 4 | - | - |
| Послание | - | 55 | 5 |
| Идиллия | - | 2 | - |
| Элегия | - | 37 | - |
| Романс | - | 1 | - |
| Баллада | - | 3 | - |
| Эпиграмма | - | - | 18 |
| Анекдот | - | - | 1 |

## 3 О возможности создания словаря стилистически дифференцированных лексем

Прежде, чем приступить к выбору алгоритмов определения стилистических и жанровых характеристик поэтических текстов, необходимо решить вопрос: возможно ли использовать для решения этой задачи априори составленные словари лексем, имеющих ту или иную стилистическую или жанровую окраску?

Большое внимание вопросам стилистической дифференциации слов уделено в монографии О.С. Ахмановой «Очерки по общей и русской лексикологии» [12]. Приведены списки слов «разговорных», со «сниженной» стилевой характеристикой и с «повышенной» стилевой характеристикой. Однако эти списки далеко не полны и носят, скорее, иллюстративный характер, более того, автор признаёт, что «далеко не все из включенных в них слов будут одинаково убедительными (многие, несомненно, покажутся спорными)», и, наконец, стилистическая окраска некоторых лексем менялась со временем, т. е. эта характеристика, взятая из монографии [12], могла быть иной как для языка XIX века, так и для современного. Поэтому для соотнесения слова с тем или иным стилем в той же монографии предложено использовать анализ их структурно-семантической формы. Так, существительные с суффиксом *-к-*а в разнообразных структурно-семантических

вариантах, а также с различными суффиксами со значением «лица» относятся к «разговорной» или «сниженной» лексике; для «разговорной», в отличие от «сниженной», лексики характерно большое число наречий; для «книжной» лексики характерны заимствованные слова, а для «возвышенной» – славянские со сложной структурой, а также архаизмы и т. п.

Однако все эти наблюдения носят весьма частный характер. Так, слова с суффиксом *-к*-а *пытка, речка, шутка* и т. д. встречаются в стихах Пушкина, относящихся отнюдь не к «низкому» или «разговорному» стилю, то же самое относится к словам *бочка, кружка, пушка* и т. д., в которых *–к* является частью корня, но установление этого факта требует нетривиального этимологического анализа, плохо поддающегося автоматизации. Заимствованные слова с течением времени становятся достоянием всех стилей, и это касается не только «древних» заимствований вроде *лошадь* или *собака*, но и новых: *велосипед*, *танк* и т. п. Славянизмы, в том числе со сложной структурой, могли использоваться, в том числе, для придания стихотворению иронического оттенка (например, «Ода его сиятельству графу Д.И. Хвостову» Пушкина и многочисленные сатирические стихи А.К. Толстого).

Ситуация осложняется еще и тем, что нередко «разговорным» или «сниженным» является не все слово в целом, а лишь один из его лексико-семантических вариантов, а также обретением словом той или иной окраски лишь при вхождении в состав фразеологизма.

Таким образом, вхождение в текст отдельных лексем не может служить достаточно надежным критерием отнесения текста к определенному стилистическому типу.

Тем более, четкое выделение жанровой принадлежности отдельных слов представляется совершенно бесперспективной задачей, и нам неизвестны сколько-нибудь удовлетворительные попытки ее разрешения хотя бы на теоретическом уровне.

Именно поэтому нам представляется наиболее целесообразным определять стилистические и жанровые характеристики поэтических текстов на основании вхождения в них совокупности лексем (включая *n*-граммы), определяемых на базе обучающей выборки.

## 4 Описание численного эксперимента

Для эксперимента использовался описанный выше корпус текстов лицейской лирики Пушкина, состоящий из 121 стихотворения, размеченных экспертом по жанрам и стилям.

При обучении была проведена лемматизация всех уникальных слов, встречающихся в текстах, и создан словарь их исходных форм. Отдельно был составлен словарь имен собственных, которые удалялись из словаря всех слов, поскольку гипотезы, подобные той, что имена из древнегреческого пантеона присущи только высокому стилю, были опровергнуты, в частности, при подготовке данных для экспериментов. Каждый текст кодировался последовательностью цифр, соответствующей количеству вхождений в него слов из словаря: 0 ставился, если слова нет в тексте, 1 – если слово встречается 1 раз, 2 – если 2 и т.д. Помимо лексических признаков, первоначально предполагалось использование стихотворных характеристик (рифма, размер, стопность и т. п.), но это привело к серьёзному ухудшению качества классификации, поэтому было решено от них отказаться.

Также были собраны словари *n*-грамм (*n*=2, 3), которые не содержали имён собственных, причем *n*-граммы были не упорядоченными внутри себя, поскольку в поэзии очень часто встречается обратный порядок слов.

Далее опишем применявшиеся нами приемы ансамблирования, то есть комбинирования алгоритмов, взаимно улучшающего их свойства.

Во-первых, это – два варианта взвешенного голосования с использованием нескольких классификаторов, в случае hard-голосования решение о классификации того или иного объекта принимается на основании заключения большинства используемых классификаторов, в случае soft-голосования результат определяется, исходя из аргумента максимизации вероятности отнесения классифицируемого объекта к некоторому классу.

Во-вторых, это – бустинг, идея которого состоит состоит в жадном выборе очередного алгоритма для добавления в композицию так, чтобы он лучшим образом компенсировал имеющиеся на этом шаге ошибки. Две основные эвристики бустинга – это фиксация $a_1 b_1(x), \dots, a_{t-1} b_{t-1}(x)$ при добавлении $a_t b_t(x)$, где $a_t = \ln \frac{1-p_t}{p_t}, t = 1, \dots, T$, $p_t$ – частота ошибки базового алгоритма $b_t$, и гладкая аппроксимация пороговой функции потерь.

Нами были применены наиболее известные примеры бустинга – AdaBoost [13], использующий экспоненциальную аппроксимацию функции потерь, и градиентный бустинг (Gradient boosting) [14]. Среди прочих нами был применён метод опорных векторов (Support Vectors Machine, SVM) [6], усиленный AdaBoost.

Наконец, в-третьих, это – стекинг [15], который основан на применении базовых классификаторов для получения предсказаний (метапризнаков) и использовании их как признаков низшего ранга для некоторого «обобщающего» алгоритма (мета-алгоритма). Иными словами, основной идеей стекинга является преобразование исходного пространства признаков задачи в новое пространство, точками которого являются предсказания базовых алгоритмов. В данном исследовании в качестве мета-алгоритма была взята логистическая регрессия над SVM, градиентным бустингом, многослойным персептроном и голосованиями.

Отметим, что в процессе решения рассматриваемой задачи нам пришлось столкнуться

с проблемой миноритарных классов, которые ясно обозначены в таблице 1. Для решения этой проблемы были применены случайное дублирование элементов миноритарных классов, а также стратегия SMOTE [16], которая основана на идее генерации некоторого количества искусственных примеров, которые были бы «похожи» на имеющиеся в миноритарном классе, но при этом не дублировали их. Для создания новой записи вычисляют разность $d = X_b - X_a$, где $X_a, X_b$ – векторы признаков «соседних» примеров $a$ и $b$ из миноритарного класса, которые находят, используя алгоритм ближайшего соседа [17]. В данном случае необходимо и достаточно для примера $b$ получить набор из $k$ соседей, из которого в дальнейшем будет выбрана запись $b$. Остальные шаги алгоритма ближайшего соседа не требуются. Далее из $d$ путем умножения каждого его элемента на случайное число в интервале (0, 1) получают $\hat{d}$. Вектор признаков нового примера вычисляется путем сложения $X_a$ и $\hat{d}$. Алгоритм SMOTE позволяет задавать количество записей, которое необходимо искусственно сгенерировать. Степень сходства примеров $a$ и $b$ можно регулировать путем изменения значения $k$ (числа ближайших соседей).

Программное приложение для классификации поэтических текстов реализовано на языке Python с использованием библиотек sklearn (реализация алгоритмов, их композиций и кросс-валидации), imblearn (реализация SMOTE), xgboost (наиболее эффективная реализация градиентного бустинга) и pymorphy2 [18] для приведения слов к нормализованному виду, а также для отсечения имен собственных.

В таблицах 2–7 приведены результаты работы классификаторов и их композиций, полученные при трехэтапной кросс-валидации (трехкратное разбиение корпуса на обучающее и тестовое множества, каждый раз классификатор обучался на обучаемом и оценивался на тестовом множестве). Из таблицы результатов был исключен рекомендуемый при работе со SMOTE метод ближайших соседей, так как он показывал очень низкую точность.

**Таблица 2** Лексические признаки + SMOTE для определения стиля

| **Классификатор** | **Средн.** | **Max** | **Min** |
|---|---|---|---|
| SVM AdaBoost | 0.88 | 0.91 | 0.84 |
| XGBoost | 0.83 | 0.9 | 0.81 |
| Многосл. перс. | 0.85 | 0.95 | 0.67 |
| Голосование, hard | 0.94 | 0.95 | 0.92 |
| Голосование, soft | 0.94 | 0.95 | 0.92 |
| Стекинг | 0.94 | 0.97 | 0.92 |

**Таблица 3** Лексические признаки + случайное дублирование миноритарных классов для определения жанра

| **Классификатор** | **Средн.** | **Max** | **Min** |
|---|---|---|---|
| SVM AdaBoost | 0.88 | 0.89 | 0.86 |
| XGBoost | 0.90 | 0.92 | 0.89 |
| Многосл. перс. | 0.93 | 0.95 | 0.91 |
| Голосование, hard | 0.92 | 0.95 | 0.88 |
| Голосование, soft | 0.92 | 0.96 | 0.88 |
| Стекинг | 0.90 | 0.93 | 0.87 |

**Таблица 4** Биграммы + SMOTE для определения стиля

| **Классификатор** | **Средн.** | **Max** | **Min** |
|---|---|---|---|
| SVM AdaBoost | 0.95 | 0.98 | 0.92 |
| XGBoost | 0.92 | 0.97 | 0.88 |
| Многосл. перс. | 0.96 | 0.98 | 0.93 |
| Голосование, hard | 0.95 | 0.98 | 0.91 |
| Голосование, soft | 0.94 | 0.97 | 0.88 |
| Стекинг | 0.95 | 0.98 | 0.90 |

**Таблица 5** Биграммы + случайное дублирование миноритарных классов для определения жанра

| **Классификатор** | **Средн.** | **Max** | **Min** |
|---|---|---|---|
| SVM AdaBoost | 0.94 | 0.96 | 0.90 |
| XGBoost | 0.97 | 1.00 | 0.93 |
| Многосл. перс. | 0.97 | 0.99 | 0.94 |
| Голосование, hard | 0.94 | 1.00 | 0.88 |
| Голосование, soft | 0.93 | 1.00 | 0.88 |
| Стекинг | 0.96 | 1.00 | 0.89 |

**Таблица 6** Триграммы + SMOTE для определения стиля

| **Классификатор** | **Средн.** | **Max** | **Min** |
|---|---|---|---|
| SVM AdaBoost | 0.83 | 0.98 | 0.88 |
| XGBoost | 0.90 | 0.94 | 0.87 |
| Многосл. перс. | 0.95 | 0.99 | 0.93 |
| Голосование, hard | 0.93 | 0.98 | 0.89 |
| Голосование, soft | 0.91 | 0.98 | 0.88 |
| Стекинг | 0.94 | 0.99 | 0.89 |

**Таблица 7** Триграммы + случайное дублирование миноритарных классов для определения жанра

| **Классификатор** | **Средн.** | **Max** | **Min** |
|---|---|---|---|
| SVM AdaBoost | 0.95 | 1.00 | 0.86 |
| XGBoost | 0.94 | 1.00 | 0.84 |
| Многосл. перс. | 0.97 | 0.99 | 0.95 |
| Голосование, hard | 0.96 | 1.00 | 0.91 |
| Голосование, soft | 0.96 | 1.00 | 0.91 |
| Стекинг | 0.96 | 1.00 | 0.88 |

Из полученных данных можно сделать следующие выводы:

- стекинг не всегда даёт наилучшее (т. е. наиболее соответствующее экспертной оценке) решение (табл. 3);
- при увеличении контекста признаков (от одного слова к би- и триграммам) XGBoost становится более точным, чем многослойный персептрон;

- увеличение ширины контекста приводит к улучшению качества, но только до определённого момента (использование тетраграмм дало заметное ухудшение результатов). Отметим, что применение популярной концепции word2vec [19] дало очень слабый результат (0.83–0.85), и при этом время подсчёта увеличилось в несколько раз;
- на основе лексических признаков или *n*-грамм можно получить хороший результат даже с помощью простых классификаторов;
- исходя из критерия максимизации минимальной точности, следует использовать многослойный персептрон, а в качестве лексических характеристик стихотворений – триграммы.

## 5 Заключение

В работе проанализированы принципы формирования обучающих выборок для алгоритмов определения стилей и жанровых типов. Проведены вычислительные эксперименты с использованием корпуса текстов лицейской лирики А.С. Пушкина по выбору наиболее точного алгоритма классификации поэтических текстов, в том числе с использованием наиболее известных приемов ансамблирования базовых алгоритмов в композиции, таких, как взвешенное голосование, бустинг и стекинг, причем в качестве характеристических признаков стихотворений использовались одиночные слов, биграммы и триграммы. Рассмотренные алгоритмы показали свою работоспособность (при этом, исходя из критерия максимизации минимальной точности, следует использовать многослойный персептрон, а в качестве лексических характеристик стихотворений – триграммы) и могут быть использованы для автоматизации комплексного анализа русских поэтических текстов, существенно облегчая работу эксперта при определении их стилей и жанров путем предоставления соответствующих рекомендаций.

## Поддержка

Работа выполнена при частичной поддержке Президиума РАН (проект 2016-PRAS-0015) и Президентской программы «Ведущие научные школы РФ» (грант 7214.2016.9).

## Литература

[1] Барахнин, В.Б., Кожемякина, О.Ю. Об автоматизации комплексного анализа русского поэтического текста. CEUR Workshop Proceedings, 934, сс. 167-171 (2012)

[2] Лесцова, М.А.: Определение ядра и периферии жанров оды, песни, послания, элегии и эпитафии на материале английских поэтов-сентименталистов XVIII века. Вестник Челябинского государственного пед. университета, 4, сс. 196-205 (2014)

[3] Орлов, Ю.Н., Осминин, К.П.: Определение жанра и автора литературного произведения статистическими методами. Прикладная информатика, 26 (2), сс. 95-108 (2010)

[4] Орлов, Ю.Н., Осминин, К.П.: Методы статистического анализа литературных текстов. Эдиториал УРСС, Москва (2012)

[5] Barakhnin, V., Kozhemyakina, O., Pastushkov, I.: Automated Determination of the Type of Genre and Stylistic Coloring of Russian Texts. ITM Web of Conferences 10, 02001, 4 p. (2017). doi: 10.1051/itmconf/20171002001

[6] Vapnik,V.N.: The Nature of Statistical Learning Theory. Springer-Verlag (1995)

[7] Грехнев, В.А.: Лирика Пушкина. О поэтике жанров. Горький: Волго-Вятское книжное издательство (1985)

[8] Барахнин, В.Б., Кожемякина, О.Ю.: К проблеме аутентичности фонетического анализа в связи с возможными особенностями авторской орфографии (на примере чередования окончаний -ой/-ый в лирике А.С. Пушкина). Вестник Томского государственного университета. Филология, 13 (2), сс. 5-28 (2016)

[9] Ломоносов, М.В.: Предисловие о пользе книг церковных в российском языке. Ломоносов, М. В. Полн. собр. соч. 7, сс. 585-592. М.–Л.: Изд-во АН СССР (1952)

[10] Магомедова, Д.М.: Филологический анализ лирического стихотворения. М.: Издательский центр «Академия» (2004)

[11] Свободина, С.Ф.: К вопросу о философской направленности и жанровых особенностях стихотворения А.С. Пушкина «Безверие». Пушкинский музеум: альманах, 6, сс. 261-270. Всероссийский музей А.С. Пушкина, Санкт-Петербург (2014)

[12] Ахманова, О.С.: Очерки по общей и русской лексикологии. М.: Учпедгиз (1957)

[13] Freund, Y., Schapire, R.E.: A Short Introduction to Boosting. J. of Japanese Society for Artificial Intelligence, 14 (5), pp. 771-780 (1999)

[14] Friedman, J.H.: Stochastic Gradient Boosting. Computational Statistics and Data Analysis, 38 (4), pp. 367-378 (2002)

[15] Wolpert, D.H.: Stacked Generalization. Neural Networks, 5 (2), pp. 241-259 (1992)

[16] Chawla, N.V.: Data Mining for Imbalanced Datasets: An Overview. Data Mining and Knowledge Discovery Handbook, pp. 875-886. Springer-Verlag (2010)

[17] Cover, T.M., Hart, P.E.: Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, 13, pp. 21-27 (1967)

[18] Korobov, M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science, 542, pp. 320-332 (2015)

[19] Mikolov, T., Kai, Chen, Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. Computation and Language, 12 p., (2013). https://arxiv.org/pdf/1301.3781.pdf