МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Факультет информационных технологий
Кафедра общей информатики

Направление подготовки 09.03.01 Информатика и вычислительная техника
Направленность (профиль): Программная инженерия и компьютерные науки

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

Гудкова Степана Алексеевича

Тема работы:

ИССЛЕДОВАНИЕ ВРЕМЕННЫХ ИЗМЕНЕНИЙ СТИЛИСТИКИ А.С. ПУШКИНА МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

«К защите допущена»
Заведующий кафедрой,
д.ф.-м.н., доцент
Пальчунов Д.Е. /……..
(ФИО) / (подпись)
«31» мая 2024 г.

Руководитель ВКР
д.т.н., доцент, профессор
кафедры общей информатики
ФИТ
Барахнин В.Б. /……...
(ФИО) / (подпись)
«31» мая 2024 г.

Новосибирск, 2024

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)
Факультет информационных технологий
Кафедра общей информатики
(название кафедры)

Направление подготовки 09.03.01 Информатика и вычислительная техника
Направленность (профиль): Программная инженерия и компьютерные науки

УТВЕРЖДАЮ

Зав. кафедрой Пальчунов Д.Е.
(фамилия, И., О.)

…….
(подпись)
31 октября 2023 г.

ЗАДАНИЕ
НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ БАКАЛАВРА

Студенту Гудкову Степану Алексеевичу, группы 20206
(фамилия, имя, отчество, номер группы)

Тема «Исследование временных изменений стилистики А.С. Пушкина методами машинного обучения»
(полное название темы выпускной квалификационной работы)

утверждена распоряжением проректора по учебной работе № 0392 от 31.10.2023.

Срок сдачи студентом готовой работы: 20 мая 2024 г.

Исходные данные (или цель работы):

Разработка алгоритмов и методов по анализу стихотворных текстов полного собрания сочинений А.С. Пушкина для выявления в них признаков, описывающих изменение авторской стилистики.

Структурные части работы:

Анализ предметной области, обзор существующих исследований, разработка алгоритма извлечения стилеметрических признаков, реализация классификатора текстов по периодам творчества, описание полученных результатов.

Консультанты по разделам ВКР (при необходимости, с указанием разделов):

_____
(раздел, ФИО)

Руководитель ВКР
д.т.н., доцент, профессор
кафедры общей
информатики ФИТ
Барахнин В.Б. /……...
(ФИО) / (подпись)
31 октября 2023 г.

Задание принял к исполнению
Гудков С.А. /……...
(ФИО студента) / (подпись)
31 октября 2023 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Использование информационных систем при анализе текстов на естественном языке, в том числе и произведений художественной литературы, позволяет применять в исследовании методы статистики и машинного обучения. Для этого требуется автоматически извлекать из текста большое число разнообразных признаков, с помощью которых возможно построить количественное описание текста на различных уровнях: пунктуационном, синтаксическом, лексическом и фонетическом. Совокупность этих уровней во многом выражает особенности авторского стиля.

Происходящее со временем изменение стилистики автора отражается на статистических характеристиках текста. Определить зависимость распределения признаков от периода создания произведения возможно с помощью методов машинного обучения. Для этого производится периодизация творческого пути автора, и затем решается задача классификации произведений по периодам творчества. Те признаки, на основании которых проведена классификация, отражают изменение авторской стилистики. Применение методов объяснимого искусственного интеллекта позволяет определить значимость использованных при анализе признаков для решения задачи классификации. Полученные таким образом знания о тексте или корпусе текстов могут быть подвергнуты лингвистическому или филологическому анализу и получить объяснение на семантическом уровне.

В отличие от других исследований, ставящих целью сравнение авторской стилистики А.С. Пушкина со стилистикой других поэтов его эпохи или осуществление периодизации творческого пути автора с помощью методов математической статистики, в настоящей работе представлено исследование по выявлению стилистических признаков текста, поддающихся количественному описанию, распределение которых изменяется в зависимости от периода творчества. Предполагается, что существует непустое множество признаков, имеющих указанную зависимость. С помощью такого множества возможно описать изменение стилистики А.С. Пушкина.

Целью дипломной работы является разработка алгоритмов и методов анализа стихотворных текстов полного собрания сочинений А.С. Пушкина для выявления в них признаков, описывающих изменение авторской стилистики.

Для достижения цели поставлены следующие задачи:

1. Проанализировать существующие способы статистического описания текста.
2. Провести предобработку стихотворных текстов полного собрания сочинений А.С. Пушкина.
3. Разработать алгоритм извлечения признаков из текста.
4. Найти эмпирическое распределение для каждого признака и определить наличие зависимости распределения от года создания текста.
5. Разработать классификатор текстов по периодам творчества на основе извлечённых признаков.
6. Оценить работу классификатора, сделать выводы, визуализировать полученные результаты.

В соответствии с намеченной целью и поставленными задачами определены методы исследования:

1. Анализ и сравнение известных способов статистического описания текста.
2. Синтез технологических решений для разработки алгоритма извлечения признаков.
3. Формализация извлечённых признаков путём построения их эмпирического распределения.
4. Моделирование исследуемых текстов с использованием выделенных признаков, анализ полученных моделей методами машинного обучения.
5. Проведение экспериментов с применением разработанных программных средств, визуализация и анализ полученных результатов.

Научная новизна работы заключается в применении методов машинного обучения для анализа стихотворных текстов полного собрания сочинений А.С. Пушкина именно с целью исследования временных изменений авторской стилистики и в разработке алгоритмов, позволяющих эти методы применить.

Практическая ценность исследования состоит в разработке программного обеспечения для извлечения признаков из текста и классификации текстов А.С. Пушкина по периодам его творчества.

Структура и объём работы: введение, 3 главы и заключение. В первой главе дано определение предметной области, сделан обзор существующих исследований и применяемых в них методологий. Во второй главе описаны характеристические признаки, используемые в исследовании, алгоритм извлечения этих признаков из текста и его реализация в соответствующем программном модуле, сформулированы требования к предварительной обработке исследуемых текстов. В третьей главе рассмотрены существующие подходы к проблеме периодизации творческого пути А.С. Пушкина, описана реализация моделей машинного обучения – классификаторов текстов по периодам творчества – и объединение полученных моделей в итоговый ансамбль, для каждого классификатора приведены значения метрик, описывающих качество его работы, и построенное для него объяснение по методу Шепли, представлен анализ полученных результатов. В заключении сформулированы выводы работы.

Глава 1. Анализ предметной области

В первой четверти XX века началась активная экспансия идей и методов естественных наук и математики в гуманитарные науки и искусство и, наоборот, гуманизация естественных наук [1]. Лингвистический материал применялся в качестве иллюстраций к теоретическим исследованиям в области математической статистики [2; 3], а применение статистических методов в филологических исследованиях породило новую область науки – стилеметрию [1].

Стилеметрия является междисциплинарной областью, объединяющей литературную стилистику, статистику, машинное обучение и компьютерную лингвистику для изучения стиля документов различного назначения [4]. Хотя ещё в конце прошлого столетия стилеметрия понималась как “прикладная филологическая дисциплина, занимающаяся измерением стилевых характеристик” [1], в настоящее время круг решаемых стилеметрией задач значительно расширился: среди них и проблемы выявления плагиата или фальсификации различных документов, статей, электронных писем и сообщений в социальных сетях [4]. Классическими задачами стилеметрии являются задачи атрибуции авторства художественного текста [5; 6], составления частотных словарей [7] и метрических справочников [8] и исследования особенностей авторского стиля [7; 9].

Изучение временных изменений стилистики автора, как один из способов постановки задачи исследования особенностей авторского стиля, направлено на “выявление таких параметров, которые отражают единство стиля автора, с одной стороны, и его изменчивость – с другой” [10].

Рассмотрим несколько определений стиля. По определению Г. Хердана, стиль — общая характеристика индивидуального способа выражения личности в языке. Стиль понимается как подсознательный фактор, которому автор не может не подчиняться [11]. Согласно В. Винтеру, стиль может быть охарактеризован как система периодически повторяющихся выборок из перечня произвольных черт языка [12]. Н.Н. Журавлёва, синтезируя два этих

определения, предлагает следующее: “Стиль — это система периодически повторяющихся выборок, характеризующая индивидуальный способ выражения в языке конкретного человека,” – и замечает, что “именно такое определение оправдывает применение количественных методов при анализе авторского стиля” [13].

Методы стилеметрии основаны на сравнении вычислимых характеристик текстов. Их классифицируют на методы статистического анализа и методы машинного обучения [14]. В общем случае текст отображается в вектор вычисленных параметров, в качестве которых выбираются его статистические характеристики: частота использования определённых частей речи, слов, знаков препинания, количество и длина предложений (измеренная в словах, слогах, знаках), объём словаря и так далее [15]. Существуют и математические инструменты, предложенные специально для анализа текстов на естественном языке [16]: дельта Берроуза, введённая в [17], авторский инвариант, предложенный в [18], а также некоторые иные средства.

В работе [19] отмечается, что методы стилеметрии позволяют исследовать текст на пяти уровнях: пунктуационном, орфографическом, синтаксическом, лексико-фразеологическом, стилистическом.

- На пунктуационном уровне выявляются особенности употребления автором знаков препинания и характерные ошибки.
- На орфографическом уровне – характерные ошибки в написании слов.
- На синтаксическом уровне определяются особенности построения предложений, предпочтение тех или иных языковых конструкций, употребление времен, порядок слов.
- На лексико-фразеологическом уровне определяются словарный запас автора, особенности использования слов и выражений, склонность к употреблению редких и иностранных слов, диалектизмов, архаизмов, неологизмов, профессионализмов, навыки употребления фразеологизмов, пословиц, поговорок.

- На стилистическом уровне определяются жанр, общая структура текста, сюжет, характерные изобразительные средства, стилистические фигуры.

В работе [20] авторы также выделяют фонетический уровень, описывающий особенности интонации и мелодики, количество повторений слогов, гласных и согласных звуков для усиления выразительной силы и благозвучия, использование фраз, обеспечивающих ритм и гармонию.

Согласно [19], под авторским стилем обычно понимаются особенности текста на синтаксическом, лексическом и стилистическом уровнях, однако авторы [4] отмечают, что для формального вычислительного анализа эти уровни достаточно сложны, поэтому большинство исследователей используют для анализа фонетический, пунктуационный, лексический и синтаксический уровни.

На сегодняшний день нет единого мнения относительно оптимального набора стилеметрических признаков. Их выбор во многом случаен и часто зависит от применяемого классификатора [20].

1.1 Обзор существующих исследований

1.1.1 В работе [4] представлено исследование авторского стиля на основе русских поэтических текстов. Поставлена цель определить особенности авторского стиля А.С. Пушкина, а также поэтов пушкинской эпохи, используя методы машинного обучения. Задача сформулирована как задача бинарной классификации, в которой выделены два класса: стихи А.С. Пушкина и стихи других поэтов пушкинской эпохи: К.Н. Батюшкова, Е.А. Баратынского, П.А. Вяземского, Н.И. Гнедича, Д.В. Давыдова, А.А. Дельвига, В.А. Жуковского.

Выделены следующие группы признаков:

- распределение по частям речи и отношениям между словами;
- частоты знаков препинания;
- слова и их n-граммы;
- служебные слова;
- буквы и другие символы, а также их n-граммы;
- метроритмические особенности.

Распределение слов по частям речи и отношениям соответствует синтаксическому уровню; частота знаков препинания соответствует уровню пунктуации. Признаки, описывающие слова и их n-граммы, и служебные слова относятся к лексическому уровню. Признаки, описывающие распределение по буквам и другим символам, а также их n-граммы, во многом отражают мелодику текста и описывают его на фонетическом уровне. Кроме того, автор в меньшей степени может сознательно контролировать такое распределение, в отличие, например, от используемых слов. Метроритмические особенности можно отнести к фонетическому уровню. Таким образом, в исследовании рассмотрены все уровни описания текста, за исключением стилистического, однако авторы отмечают, что описание текста с использованием выделенного набора признаков не является исчерпывающим ни на одном уровне.

Лучший из полученных классификаторов основан на всех выделенных группах признаков: 2-, 3- и 4-граммах символов, словах, знаках препинания, грамматических отношениях и метроритмических характеристиках. Значение метрики AUC ROC для этой модели составило 0,924 при тестировании на выборке стихотворений, содержащих не менее 16 строк.

Сама классификация не отражает, какие признаки характерны или, наоборот, не характерны для творчества А.С. Пушкина. Для построения объяснения в рассматриваемой работе использован метод Шепли. Обнаружено, что наиболее значимыми признаками являются те, которые характеризуют предпочтения авторов в знаках препинания: Пушкин, в отличие от других поэтов, предпочитал использование точек и многоточий, а не восклицательных знаков, свойственных другим авторам. Также замечено, что Пушкин не был склонен употреблять междометие *ах*. При анализе значимых признаков, описывающих грамматические отношения, подтверждается, что пушкинскому творчеству не свойственно широкое употребление служебных слов. Кроме того, для А.С. Пушкина характерно использование таких грамматических конструкций, при которых значение существительного или местоимения

дополняется прилагательным. Метрическая характеристика стихотворения также входит в число значимых признаков.

В дальнейших исследованиях авторы подтвердили полученные результаты методами, связанными с вычислением энтропийных характеристик поэтического текста, а именно буквенной, звуковой и «эмоциональной» энтропии [21].

1.1.2 В исследовании [22] рассматривается вопрос о периодизации творческого пути А.С. Пушкина с применением методов математической статистики и компьютерных технологий. В частности, применён кластерный анализ, цель которого состоит в разбиении множества на некоторое количество таких подмножеств, элементы которых различаются между собой значительно меньше, чем с элементами из всех других подмножеств.

Анализ каждого стихотворения производился по 10 индексам (признакам). А именно:

1) разнообразие метрики (отношение количества размеров к количеству стихов);

2) номер размера – все размеры перенумерованы от 1 (самый частый – четырехстопный ямб) до 10 (группа самых редких размеров);

3) в хорее и ямбе количество стихов с пропуском ударения на 1-м икте по отношению к общему количеству стихов;

4) то же самое на 2-м икте;

5) в хорее и ямбе количество стихов с пропуском половины и более ударений на иктах по отношению к общему количеству стихов;

6) количество стихов, разорванных синтаксическими паузами, по отношению к общему количеству стихов;

7) количество стихов с неточными рифмами по отношению к общему количеству стихов;

8) то же для приблизительных рифм;

9) количество строф по отношению к количеству стихов;

10) количество разных строфических форм по отношению к количеству стихов.

Все стихотворения, написанные в течение года, объединялись и рассматривались как единый текст; в качестве значений индексов брались их средние арифметические. Исключения составляют 1817, 1820, 1824 года. Каждый из них был разбит на два массива, условно приравненных к году: 1817 (Лицей); 1817 (после Лицея); 1820 (до ссылки); 1820 (ссылка); 1824 (Юг); 1824 (Михайловское).

Полученная в результате исследования периодизация творческого пути Пушкина содержит 10 периодов: 1813 – 1815; 1816 – 1817 (Лицей); 1817 (после Лицея) – 1820 (до ссылки); 1820 (Юг) – 1821; 1822 – 1824 (Юг); 1824 (Михайловское) – 1826; 1827 – 1829; 1830; 1831 – 1832; 1833 – 1836.

Как отмечают авторы, результаты кластерного анализа подтвердили традиционную периодизацию, основанную на важнейших биографических фактах, и выявили существенные особенности творчества, остававшиеся в тени.

По такой же методике, добавив в рассмотрение меру разнообразия структуры книги как случайную величину и методы корреляционного анализа, авторами установлена периодизация творческого пути Н.С. Гумилёва и Б.Л. Пастернака.

Заметим, что для построения модели машинного обучения – классификатора произведений А.С. Пушкина по периодам творчества – полученная в рамках рассмотренного исследования периодизация творческого пути поэта не может быть использована в качестве разбиения множества произведений на классы, соответствующие периодам, поскольку процесс обучения модели требует большого числа текстов-образцов, представленных в обучающей выборке. Каждый из десяти периодов данной периодизации содержит около ста стихотворных произведений (а девятый – только лишь 25, это период, когда проза в творчестве Пушкина начинает превалировать над поэтическими жанрами). Оптимально количество классов, при котором каждый из них содержит по 200-250 произведений, то есть около четырёх классов.

Однако, безусловно, разбиение на большие по временной длительности и количеству произведений периоды должно опираться и на биографические сведения, и на результаты, полученные применением иных методов исследования проблемы периодизации творчества автора.

1.1.3 Материалом исследования [23] стала опубликованная в прижизненных сборниках ямбическая лирика Э.А. По, которая составляет ядро его творчества как поэта. На основании биографических данных выделены три периода его творчества: «ранний», «средний» и «зрелый». В работе рассматриваются морфологические, синтаксические и ритмические параметры, а также характеристики стихотворного синтаксиса.

Поскольку произведения имеют разный объём, данные по каждому тексту были нормированы путём деления на количество строк.

В качестве основного метода используется многомерный дискриминантный анализ. Как отмечают авторы, его преимуществом является выявление комплексов признаков, дифференцирующих сопоставляемые классы. Кроме того, полученные данные позволяют оценить степень удалённости групп объектов друг от друга в многомерном признаковом пространстве.

Проверка качества работы классификатора заключалась в автоматическом разбиении текстов на группы исключительно на основании полученных дифференциальных признаков. Затем состав полученных групп сравнивался с составом имеющихся периодов. В результате точность классификатора составила 100%.

Разработанная в ходе этого исследования методология применена авторами для исследования динамики индивидуального стиля Д. Райли [24] и Г. Лонгфелло [25].

1.1.4 Эта же методика анализа, используя в качестве параметров исключительно частеречные признаки, применяется к творчеству Д.Г. Лоуренса в исследовании [10]. В качестве параметров были привлечены следующие морфологические классы: «Существительное», «Глагол», «Модальный глагол», «Прилагательное», «Наречие», «Местоимение», «Личное местоимение»,

«Причастие I», «Причастие II», «Определенный артикль», «Неопределенный артикль», «Служебные слова», «Послелог» и «Междометие». Для всех произведений было определено количество единиц данных морфологических классов. Полученные абсолютные показатели были нормированы путём деления на число строк для каждого произведения.

Основными методами исследования являются многомерные статистические процедуры в рамках дискриминантного анализа.

В результате установлены признаки, выражающие различие между тремя периодами творчества поэта и признаки, определяющие единство стиля, то есть не дифференцирующие эти три этапа. На основе полученной признаковой модели для каждого этапа выделены произведения, составляющие его ядро, то есть имеющие минимальное расстояние до центра своего класса. Для измерения расстояний использована мера Махаланобиса.

1.1.5 В исследованиях, методы которых опираются на машинное обучение, в качестве меры близости применяется расхождение Кульбака – Лейблера, определённое в [26]. Обоснование применимости этого метода к задачам стилеметрии приведено в [27]. Множеством событий, расхождение между которыми измеряется в ходе исследования, является частота появления в текстах стоп-слов – предлогов, союзов, местоимений. Авторы отмечают, что именно “анализ встречаемости в текстах стоп-слов добавляет его результату стилевые особенности написания текста автором”. В процессе получения вероятностных мер проанализированы фрагменты художественных произведений объемом по 5 000 слов. Выбрано множество из 20 стоп-слов, которые присутствуют во всех исследуемых текстах. Были проанализированы фрагменты произведений Л.Н. Толстого «Война и мир» и «Анна Каренина», Ф.М. Достоевского «Преступление и наказание» и «Идиот», Н.В. Гоголя «Мертвые души» и «Вий». Обнаружено, что внутриавторские расхождения Кульбака – Лейблера меньше, чем межавторские.

1.2 Требования к методологии исследования

На основании проведённого анализа существующих исследований сформулированы методологические принципы, применяемые в настоящей работе:

- выявление стилевых признаков, описывающих исследуемый текст на фонетическом, пунктуационном, лексическом и синтаксическом уровнях;
- выделение периодов в творчестве А.С. Пушкина не только на основании биографических сведений, но и в соответствии с существующими стилеметрическими исследованиями;
- применение методов объяснимого искусственного интеллекта для оценки значимости признаков при классификации произведений по периодам творчества.

Глава 2. Программный модуль извлечения признаков

В настоящем исследовании проведён анализ изменения авторской стилистики А.С. Пушкина на основе созданных им произведений: стихотворений, поэм, сказок, драматических произведений и романа в стихах “Евгений Онегин”. Все тексты взяты для анализа из Собрания сочинений [28], доступного в электронной версии на сайте Русской виртуальной библиотеки [29]. Осуществлена их временная разметка, то есть каждому тексту поставлен в соответствие год его создания. Согласно принятой при издании собраний сочинений А.С. Пушкина традиции 1817, 1820 и 1824 года поделены каждый на два периода, условно приравненных к году: 1817 до окончания Лицея; 1817 после выпуска из Лицея; 1820 до перевода из столицы на юг; 1820 – на юге; 1824 – на юге; 1824 – в Михайловском. Кроме того, два произведения – поэма “Руслан и Людмила” и роман “Евгений Онегин” – разбиты на песни и главы соответственно, рассматриваемые как независимые тексты. Это объясняется тем, что работа над различными фрагментами шла в разные годы и каждый из фрагментов достаточно велик по объёму. Год создания каждого фрагмента (его временная метка) уточнён по электронному научному изданию “ПУШКИН” [30]. Общее количество текстов, взятых для исследования, составляет 785 штук.

Проведён парсинг предназначенных для анализа текстов с использованием облачного сервиса UDPipe, который осуществляет описание структуры предложений, извлекая из текста универсальные зависимости. Универсальные зависимости – это структура для графического представления грамматики, включающая описание частей речи, морфологических и синтаксических признаков [31]. Зависимость – бинарное асимметричное отношение главного слова к зависимому слову или к главному слову зависимого выражения. Отличительной чертой универсальных зависимостей является отображение грамматических отношений между словами [31].

В работе использована модель russian-gsd-ud-2.6-200830, предназначенная для описания структуры русскоязычных текстов. Для каждого предложения каждого произведения построено дерево универсальных зависимостей, из

которого выделены признаки, характеризующие особенности авторского стиля: слова в начальных формах, части речи и грамматические отношения между словами. Фактически, это позволило дополнить текст произведения метаданными, содержащими информацию об универсальных зависимостях, извлечённых из этого текста. Вся дальнейшая работа велась с дополненным представлением текста произведений, хранящимся в файлах формата CoNLL-U.

2.1 Используемые стилеметрические характеристики

Разработан программный модуль на языке программирования Python, позволяющий для произвольного множества текстов построить частотный словарь слов в начальных формах, знаков препинания, частей речи, типов грамматических отношений между словами, *n*-грамм по произвольным символам, по буквам русского алфавита, по знакам препинания, по словам, длин слов, количества слов в предложении и количества знаков препинания в предложении; а также вычислить однородность текста и ранг входящих в него слов.

Однородность текста – величина, характеризующая появление новых слов в начале, в середине и в конце текста. Для вычисления однородности выполняется прямое и обратное прочтение текста по словам, и на каждом шаге вычисляется количество уникальных слов, встретившихся до этого шага. Для каждого прочтения можно построить график зависимости количества уникальных слов от числа сделанных шагов. Понятно, что это график неубывающей кусочно-постоянной функции (рисунок 1).

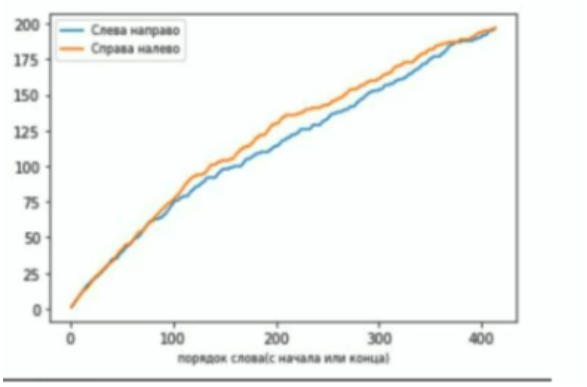

Рисунок 1 – Пример графика однородности текста

Однородность текста равна площади между графиками для прямого и обратного процесса. Так как графики – кусочно-постоянны, величина площади равна сумме по всем промежуткам между соседними целыми числами разниц значений при прямом и обратном прочтении. Кроме того, эту характеристику можно усреднить, поделив на количество слов в тексте.

Ранг слова определяется как порядковый номер этого слова в упорядоченном по убыванию частотном словаре слов, построенном по всему корпусу текстов. Вычисление ранга слов позволяет оценить, как часто автор использует уникальные слова. Для характеристики текста по этому показателю рассчитывается сумма рангов входящих в него слов, делённая на количество слов в тексте.

На основании извлечённых признаков получено описание исследуемых текстов на пунктуационном, синтаксическом, лексическом и фонетическом уровнях. Пунктуационному уровню соответствуют частоты знаков препинания, *n*-граммы по знакам препинания и количество знаков препинания в предложении. Синтаксическому уровню – распределение слов по частям речи и словосочетаний по типу грамматической связи. Уровню лексики – признаки, описывающие слова: их начальные формы, длина слов, *n*-граммы по словам, количество слов в предложении, а также однородность текста и ранг входящих в него слов. *N*-граммы по произвольным символам и по буквам русского алфавита во многом отражают мелодику текста и описывают его на фонетическом уровне.

2.2 Дополнительные возможности программного модуля

Разработанный программный модуль поддерживает расширение множества извлекаемых признаков. Для этого необходимо реализовать функцию, имя которой соответствует названию признака в информационной системе, а возвращаемое значение имеет тип словаря (class 'dict'). Функция может принимать произвольное число именованных аргументов. Иных ограничений на сигнатуру функции не накладывается. Кроме реализации

функции, указатель на неё необходимо внести в список доступных в системе признаков.

Разработанный программный модуль предоставляет возможность извлечения выбранного признака из отдельного текста, из всех текстов, соответствующих некоторому периоду (набору временных меток, присвоенных текстам), из всех текстов всех периодов. При извлечении признака из множества текстов он извлекается из каждого текста, и полученные частотные словари агрегируются – значения при каждом ключе суммируются по всем словарям. Результаты можно представить, кроме текстового представления в формате “ключ: значение”, в виде горизонтальной диаграммы, в которой ключи отсортированы по убыванию частоты их встречаемости; вертикальной диаграммы, подходящей для случаев, когда ключи являются числами – длиной слов, количеством слов и знаков препинания в предложении, однородностью текста и его рангом; и эмпирической функции распределения, подходящей для тех же случаев, что и вертикальная диаграмма, которая по смыслу близка к плотности эмпирического распределения.

2.3 Построение векторного представления текста

Для выполнения дальнейших задач – построения моделей машинного обучения – модуль извлечения признаков дополнен функциями преобразования текста в векторное представление. Они предоставляют возможность преобразовать в числовой вектор признаков отдельный текст и все тексты заданного периода. Во втором случае функция возвращает массив числовых векторов. Так как в обучающей выборке модели машинного обучения все векторы должны иметь одинаковую размерность, а результатом извлечения любого признака из текста является частотный словарь, необходима процедура преобразования содержания словаря в числовую последовательность заданной длины. Таких процедур в разработанном программном модуле предусмотрено две: для признаков, ключи в частотных словарях которых суть категории (слова, части речи, грамматические отношения, *n*-граммы), можно выбрать конечную упорядоченную последовательность ключей, каждому из которых в числовом

векторе будет соответствовать отношение значения при этом ключе к сумме значений при всех ключах словаря; вторая процедура предусмотрена для признаков, отражающих числовые характеристики отдельных слов или предложений, входящих в текст: длины слов, количества слов в предложении и количества знаков препинания в предложении. В этом случае в числовой вектор попадает среднее значение по всем элементам текста, которое, кроме того, можно нормировать, задав соответствующий коэффициент. Отметим, что есть два признака: однородность текста и ранг текста, – выражающиеся в виде одного числа, характеризующего весь текст целиком, которое и устанавливается в числовой вектор. Это число также при необходимости может быть нормировано путём деления на заданный, единый для всех преобразуемых в векторный вид текстов коэффициент. Множество функций преобразования содержания словаря в числовую последовательность при необходимости может быть расширено путём реализации новой функции, принимающей на вход преобразуемый словарь и произвольный набор неименованных аргументов, в том числе и пустой. Возвращаемое значение должно иметь тип либо вещественного числа (<class 'float'>), либо словаря, в качестве ключей содержащего набор аргументов, передаваемых функции, а в качестве значений – вещественные числа. Впрочем, количество допустимых типов возвращаемого значения может быть увеличено без потери обратной совместимости; требуется лишь, чтобы добавляемые типы отличались от уже допустимых (то есть, например, не может быть добавлен словарь, имеющий отличную от описанной выше структуру).

При формировании из множества преобразованных в векторный вид текстов обучающей, валидационной и тестовой выборок для построения модели машинного обучения осуществляется отбор текстов-образцов, имеющих достаточное (заданное как параметр) количество ненулевых координат в векторном представлении. Заметим, что нулевое значение почти всегда означает отсутствие признака. Действительно, если признак в векторе представлен как доля значений при некотором ключе в общем объёме словаря, равенство нулю

означает, что этот ключ (например, некоторое слово) ни разу не встретился в тексте; если значение признака – средняя величина одной из следующих характеристик: длины слов, количества слов в предложении, количества знаков препинания в предложении, – то равенство его нулю означает, что текст не содержит ни одного символа, но такие тексты в рассмотрение не приняты. Равенство нулю ранга текста означает то же – отсутствие в тексте хотя бы одного слова, а нулевое значение однородности – то, что слова, имеющие одинаковую начальную форму, расположены в тексте симметрично (что, как правило, выражается тем, что все слова имеют разную начальную форму). Из рассмотрения всех вариантов видно, что нулевое значение координаты в векторном представлении текста может означать либо отсутствие в этом тексте некоторого конкретного слова, части речи, *n*-граммы или грамматического отношения, соответствующего этой координате, либо равенство нулю однородности текста, которая в векторном представлении занимает не более одной координаты. Таким образом, если в векторе, соответствующем тексту, содержится достаточно много нулевых значений, то в этом тексте отсутствует достаточно большое количество признаков, что делает его нерепрезентативным. В большинстве случаев это связано с малым объёмом рассматриваемого текста: например, он может состоять лишь из одного четверостишия. При формировании выборок для обучения модели машинного обучения значение допустимого количества нулевых координат выбиралось примерно равным половине количества признаков, то есть половине размерности векторов представления текстов.

Другим важным моментом, который необходимо учитывать при построении выборок, является то, что объём классов, на которые разделяется массив исследуемых данных в зависимости от задачи, решаемой каждой отдельной моделью машинного обучения, может значительно отличаться. Необходимо, чтобы в каждой из выборок: обучающей, валидационной и тестовой – было одинаковое соотношение классов.

С учётом описанных выше замечаний реализована функция формирования для заданного набора классов, содержащих временные метки исследуемых текстов, и заданного набора признаков, по которым должно быть построено векторное представление текстов, обучающей, валидационной и тестовой выборок, составляющих заданную долю от общего числа репрезентативных текстов. После разделения данных на выборки в необходимом соотношении содержимое каждой из выборок перемешивается в случайном порядке. Реализована возможность сохранения полученных выборок в файл формата txt и загрузка их из файла.

Таким образом, реализованы все программные средства, необходимые для формирования данных для обучения моделей машинного обучения. Требуется только задать список используемых признаков, периоды, соответствующие классам, и вызвать функцию, выполняющую формирование обучающей, валидационной и тестовой выборок.

Глава 3. Классификатор произведений по периодам творчества

Для разработки моделей машинного обучения – классификаторов произведений А.С. Пушкина по периодам создания – необходимо произвести периодизацию творческого пути поэта.

3.1 Проблема периодизации творчества Пушкина

Проблема периодизации творчества Пушкина разными исследователями-пушкинистами решается исходя из различных оснований. В классических монографиях [32] и [33] эволюция пушкинского творчества изучается в рамках биографической схемы: Лицей — Петербург — Юг — Михайловское — После 1825 года — Последние (1830-е) годы. Основываясь на понятии стиля, Л.Я. Гинзбург в творчестве Пушкина выделила три основных периода: классицистический, романтический и реалистический (последний соответствует 1830-м годам) [34]. С.А. Фомичёв в исследовании [35], исходя из изменения жанровой системы и художественного метода Пушкина, отмечает такие периоды:

1) 1813—1816 гг.,
2) 1816—1820 гг.,
3) 1821—1823 гг.,
4) 1823—1828 гг.,
5) 1828—1833 гг.,
6) 1834—1837 гг.

Ф.М. Достоевский в речи о Пушкине изложил свой взгляд на вопрос о периодизации пушкинского творческого пути: “Я делю деятельность нашего великого поэта на три периода. <…> Замечу, однако же, мимоходом, что периоды деятельности Пушкина не имеют, кажется мне, твердых между собою границ. <…> Некоторые из произведений даже третьего периода могли, например, явиться в самом начале поэтической деятельности нашего поэта, ибо Пушкин был всегда цельным, целокупным, так сказать, организмом. <…> Но организм этот развивался, и периоды этого развития действительно можно обозначить и отметить, в каждом из них, его особый характер и постепенность

вырождения одного периода из другого” [36]. Временных границ, соответствующих периодам творчества Пушкина, Достоевский не выделяет.

Напомним, что в I главе настоящей работы было рассмотрено решение проблемы периодизации с применением методов математической статистики и компьютерных технологий.

Так как наше исследование опирается на методы машинного обучения, требующие для решения задачи классификации достаточно большого объёма классов, решено выделить четыре периода, границы между которыми отмечены в исследованиях, основанных на различных подходах (в скобках указано число текстов, созданных в этот период):

- с 1813 по лето 1820 года (216 текстов);
- со второй половины 1820 по лето 1824 года (133) – это период пребывания Пушкина на Юге;
- со второй половины 1824 по 1828 год (215);
- с 1829 по 1836 год (221).

С точки зрения биографического подхода, граница, проложенная по лету 1820 года, связана с переводом Пушкина из Петербурга на Юг; по лету 1824 года – со ссылкой в Михайловское; между 1828 и 1829 годами – с знакомством с Натальей Гончаровой (в декабре 1828) и с началом активных занятий прозаическим творчеством.

3.2 Описание классификаторов

Получен набор бинарных классификаторов, каждый из которых разделяет множество исследуемых текстов по одной из рассмотренных границ. В таблице 1 приведено описание этих классификаторов: все они реализуют архитектуру многослойного перцептрона, состоящего из 4 или 5 слоёв, кроме входного. Количество нейронов на каждом слое указано в скобках, на входном слое оно совпадает с количеством признаков.

Таблица 1 – Описание классификаторов

Название модели	model1	model2	model4	model4-1	model5
Класс 0	1813	1813	1829	1829	1820 на Юге
	пер. пол. 1824	пер. пол. 1820	1836	1836	пер. пол. 1824
Класс 1	1824 в Михайловском	1820 на Юге	1813	1813	1824 в Михайловском
	1836	1836	1828	1828	1828
Количество текстов	736	463	768	709	335
Количество признаков	24	22	21	23	20
Количество слоёв	4 (16-12-6-1)	5 (16-12-6-4-1)	5 (16-12-8-4-1)	4 (16-12-6-1)	4 (16-12-6-1)
Accuracy	0,804	0,817	0,812	0,803	0,691
AUC ROC	0,867	0,855	0,796	0,812	0,726
F1-мера	0,828	0,872	0,876	0,875	0,779

Выбор в качестве архитектуры нейронной сети многослойного перцептрона объясняется, во-первых и прежде всего, особенностью формирования векторного представления текстов: координаты вектора являются в значительной мере независимыми, отражающими свой признак. Если бы текст был представлен в виде последовательности чисел, по своей структуре соответствующей последовательности слов или символов в тексте, то было бы уместно использовать свёрточные и рекуррентные нейронные сети. Во-вторых, обучение многослойного перцептрона требует меньшего объёма вычислительных ресурсов, что позволяет произвести большее количество экспериментов для более точного подбора гиперпараметров модели.

Для получения моделей машинного обучения, проведения экспериментов и визуализации полученных результатов реализованы функции-обёртки, оперирующие средствами фреймворка Keras, позволяющие сконструировать многослойный перцептрон с заданным количеством нейронов для каждого слоя, обучить построенную модель по заданной обучающей выборке с графическим выводом зависимости функции потерь и точности модели от числа эпох (если задана валидационная выборка, то график будет построен и для неё), сформировать для обученной модели матрицу ошибок, вычислить F1-меру и величину площади под ROC-кривой и построить графики Precision-Recall- и ROC-кривых.

Процесс построения классификаторов начинался с выбора признаков, распределение которых заметно отличается на исследуемых классах текстов. Затем происходило формирование обучающей выборки при значении допустимого количества нулевых координат в векторе, равном половине числа признаков. Осуществлялось обучение моделей с различным набором полносвязных слоёв. Также проводились эксперименты по обучению на выборке текстов, увеличенной или уменьшенной за счёт соответствующего изменения допустимого количества нулевых координат в векторном представлении. Таким образом подбирался оптимальный объём обучающей выборки. Полученные модели сравнивались по трём метрикам качества, представленным в таблице 1, причём большее внимание уделялось площади под ROC-кривой и F1-мере, так как эти метрики более полно отражают работу модели на классах неравного объёма. Так были определены лучшие модели.

При построении нейронных сетей с названием “model1” и “model4” (то есть разделяющих множество исследуемых текстов на те же классы, что и соответствующие модели в таблице 1) применена такая методика: для лучшей модели было построено объяснение по методу Шепли (более подробно о нём – далее), и затем пересмотрено множество признаков – признак, показавший самую низкую значимость, удалён, и для “model1” добавлено 5 новых признаков, для “model4” – 3 новых признака. По пересмотренным наборам

признаков сформированы новые обучающие выборки и обучены новые нейронные сети. В случае с “model1” повышение метрик качества оказалось незначительно – порядка 0,01, а в случае с “model4” значение площади под ROC-кривой увеличилось на 0,016, однако точность снизилась на 0,009, величина F1-меры при этом не изменилась. Для новой модели “model4” также построено объяснение. Сравнение с объяснением модели предыдущей версии показало, что важность признаков во многом перераспределилась, что говорит о том, что нейронные сети выявили различные аспекты зависимости признаков от периода создания текста, поэтому решено в дальнейших исследованиях с применением ансамблевого подхода использовать обе эти модели. При построении других моделей методика пересмотра множества признаков на основе объяснения нейронной сети не применялась.

3.3 Итоговый ансамбль

Классификаторы, представленные в таблице 1, объединены в итоговый ансамбль, осуществляющий классификацию исследуемых произведений по четырём выделенным ранее периодам. Решено проводить тестирование ансамбля на всём корпусе текстов, поскольку для каждого классификатора обучающая, валидационная и тестовая выборки формировались независимо. Так как наборы признаков, на которых основан каждый из классификаторов, различны (здесь важно не только само множество признаков, но и их порядок в векторном представлении текста), для тестирования ансамбля тексты необходимо преобразовать к виду массива векторов – для каждой модели из ансамбля свой вектор. Реализованы функции формирования тестовой выборки для ансамбля и получения предсказаний от каждой модели. Функционирование ансамбля заключается в присвоении тексту одной из четырёх меток, соответствующих периоду, на основе сопоставления предсказаний, полученных от каждой модели ансамбля. В результате тестирования сформирована матрица ошибок, представленная в таблице 2. В первом столбце отмечены действительные метки текстов, в первой строке – предсказанные ансамблем. Ячейки отражают долю текстов, в действительности относящихся к периоду,

соответствующему строке, отнесённых классификатором к периоду, соответствующему столбцу.

Таблица 2 – Матрица ошибок ансамбля

	1813 - 1820	1820 Юг - 1824	1824 Мих. - 1828	1829 - 1836
1813 - 1820	0,48	0,25	0,2	0,07
1820 Юг - 1824	0,23	0,43	0,23	0,11
1824 Мих. - 1828	0,13	0,2	0,44	0,22
1829 - 1836	0,07	0,14	0,32	0,48

При анализе результатов важно помнить, что классы в данном случае не являются некоторыми независимыми категориями, они – суть участки непрерывного процесса творческого развития Пушкина. Заметим, что, чем далее предсказанный период расположен относительно действительного, тем меньшее количество текстов получает в качестве предсказания ансамбля этот период. Это отражает снижение стилистической схожести текстов при увеличении временного промежутка между их созданием.

3.4 Оценка влияния признаков

Для оценки влияния признаков на результаты классификации построено объяснение каждой модели, вошедшей в итоговой ансамбль, с помощью библиотеки shap языка программирования Python. Она предоставляет возможность применить метод Шепли как метод объяснимого искусственного интеллекта. В качестве объясняющей функции использован KernelExplainer. Полученные диаграммы представлены в приложении А в том же порядке, что и соответствующие модели в таблице 1. Признаки отсортированы по убыванию значимости их влияния на результаты классификации. Точки с отрицательным значением абсцисс соответствуют текстам, получившим в качестве предсказания модели класс 0, точки с положительным значением абсцисс –

текстам, получившим в качестве предсказания класс 1. Точки синего цвета отражают малые значения признака, красного – высокие значения. Для каждого признака линия на диаграмме имеет толщину, пропорциональную количеству точек, имеющих соответствующую абсциссу.

Анализ полученных диаграмм, отражающих влияние признаков на результаты работы классификаторов, позволяет описать временные изменения авторской стилистики А.С. Пушкина в форме характерных черт каждого периода.

Раннему творчеству (1813 – 1820 гг.) характерно широкое употребление имён собственных, словосочетаний существительного с существительным и дополнений как членов предложения, зачастую соответствующих таким словосочетаниям.

Восклицательный знак в ранних произведениях встречается очень часто. В Южный период намечается снижение частоты употребления, и, начиная с середины 1820-х годов, восклицательный знак не характерен творчеству Пушкина.

В период пребывания на Юге (1820 – 1824 гг.) более, чем в другие периоды, Пушкин склонен использовать слово *что*. При этом соответствующая буквенная 3-грамма чаще встречается в текстах 30-х годов. Возможно, это связано с повышением разнообразия составных союзов, включающих в себя *что*. Также в произведениях Южного периода реже встречается вопросительный знак.

В произведениях периода 1824 – 1828 гг. высока доля буквенных 3-грамм *ень*, зачастую встречающихся в суффиксе существительного вместо *ени*: например, мгновенье, вдохновенье и т.п., в V главе “Евгения Онегина”, также относящейся к этому периоду, употребляется даже форма *Евгенья* – в родительном падеже.

Позднему творчеству (1829 – 1836 гг.) характерно более частое использование местоимений. Это выражается в значимости высокой частоты употребления как местоимений относительно других частей речи, так и личных

местоимений *я*, *мы*, *вы*, *он*, *они* относительно всех слов в тексте. Кроме того, высокие значения частоты употребления слова *я* имеют большую значимость и для произведений второй половины 1820-х годов. С другой стороны, личные местоимения *ты* и *она* чаще употребляются в первые два периода – с 1813 по 1824 год. Это, безусловно, взаимосвязано с высокой значимостью для этих же периодов слов *любовь* и *друг*, что в некоторой степени отражает и основные темы поэзии Пушкина первой половины его творчества. Отход от этих тем намечается уже в середине 1820-х годов, что, в частности, отражается в снижении доли буквенных 3-грамм *люб* в третий из выделенных нами периодов.

Также в поздний период увеличивается доля буквенных 3-грамм *ста*, входящих в слова, близкие по значению или происхождению к словам стать, ставить, старость.

В процессе творческого развития происходит расширение разнообразия лексики (это отражают однородность – homogeneity на диаграммах – и ранг текста), учащается использование предлогов, обстоятельств, снижается средняя длина слов, частота употребления однородных членов предложения. Снижается и доля прилагательных, что отражается, в том числе, и в таких признаках: буквенных 3-граммах *ный* и *ной* и 3-грамме по всем символам *ой_* (здесь знаком _ обозначен пробел). Кроме того, упрощается структура предложений: позднему периоду характерно меньшее число знаков препинания в предложении. Количество слов в предложении снижается в 1820-е годы в сравнении с ранним периодом, но затем, в 1830-е годы, снова повышается.

Отметим также, что ранг текста оказался самым значимым признаком для всех классификаторов, что демонстрирует высокую эффективность этого статистического инструмента.

Таким образом, получен и протестирован классификатор стихотворных текстов А.С. Пушкина по периодам творчества. Классификатор разработан с применением ансамблевого подхода на основе нейронных сетей с архитектурой многослойного перцептрона. Применение методов объяснимого искусственного интеллекта позволило выделить характерные черты каждого периода.

ЗАКЛЮЧЕНИЕ

В настоящем исследовании рассмотрены временные изменения авторской стилистики А.С. Пушкина. Результаты сформулированы в виде характерных черт четырёх периодов творчества поэта, выделенных на основании синтеза различных подходов, разработанных в пушкинистике. Для выявления закономерностей, описывающих особенности авторского стиля в каждый из периодов, использованы методы машинного обучения, а именно получен ансамбль, состоящий из пяти нейронных сетей, реализующих архитектуру многослойного перцептрона, осуществляющий классификацию поэтических текстов Пушкина по выделенным периодам. Для каждой нейронной сети, вошедшей в ансамбль, построено объяснение с помощью метода Шепли, позволившее оценить влияние признаков на результаты классификации и тем самым выявить характерные черты каждого периода творческого развития Пушкина.

Таким образом, достигнута цель исследования: разработаны алгоритмы и методы анализа стихотворных текстов полного собрания сочинений А.С. Пушкина для выявления в них признаков, описывающих изменение авторской стилистики. В ходе работы все задачи, поставленные для достижения цели, решены. Описание разработанной программы и руководство оператора содержатся в приложениях Б и В соответственно.

В дальнейших исследованиях могло бы быть перспективным рассмотрение другого способа периодизации творческого пути А.С. Пушкина для классификации произведений по периодам творчества или, с другой стороны, применение регрессионных нейронных сетей, предсказывающих не период, а непосредственно временную метку, в нашем случае в промежутке от 1813 до 1836 года. Также возможна постановка задачи кластеризации, например, на основании признаков, выявленных в рамках представленного исследования.

Выпускная квалификационная работа выполнена мной самостоятельно и с соблюдением правил профессиональной этики. Все использованные в работе

материалы и заимствованные принципиальные положения (концепции) из опубликованной научной литературы и других источников имеют ссылки на них. Я несу ответственность за приведенные данные и сделанные выводы.

Я ознакомлен с программой государственной итоговой аттестации, согласно которой обнаружение плагиата, фальсификации данных и ложного цитирования является основанием для не допуска к защите выпускной квалификационной работы и выставления оценки «неудовлетворительно».

Гудков Степан Алексеевич
ФИО студента

_____
Подпись студента

« ____ »_____2024 г.
(заполняется от руки)

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ И ЛИТЕРАТУРЫ

1. Мартыненко Г. Я. Стилеметрия: возникновение и становление в контексте междисциплинарного взаимодействия //Структурная и прикладная лингвистика. – 2014. – Вып. 10. – С. 3–23.
2. Марков А. А. Опыт статистического исследования текста романа «Евгений Онегин» / Изв. Имп. Академии наук. СПб., 1913. Серия 6. Т. 7.
3. Дж М., Кендалл А. Стьюарт //Теория распределений. М:«Наука». – 1966.
4. Barakhnin V., Kozhemyakina O., Grigorieva I. Determination of the Features of the Author’s Style of AS Pushkin’s Poems by Machine Learning Methods //Applied Sciences. – 2022. – Т. 12. – №. 3. – С. 1674.
5. Dittenberger W. Sprachliche Kriterien für Chronologie der Platonische Dialoge. Hermes 16. Berlin, 1881. S. 321–345.
6. Морозов Н. А. Лингвистические спектры: Средство для отличия плагиатов от истинных произведений того или другого известного автора: Стилеметрический этюд / Известия Отделения рус. языка и словесности Имп. Академии наук. СПб., 1915. Т. XX. Кн. 4. С. 93–134.
7. Куницкий В. Н. Язык и слог комедии "Горе от ума": К столетию дня рождения А. С. Грибоедова, 4 янв. 1795 г. — 4 янв. 1895. — К.: Тип. И.И. Чоколова, 1894. – 57 с.
8. Лапшина Н. В., Романович И. К., Ярхо Б. И. Метрический справочник к стихотворениям А.С. Пушкина. — М.; Л.: Academia, 1934.
9. Jule G. U. Oe Statistical study of Literary Vocabulary. Camdrige, 1944.
10. Новокщёнова Н. И. Морфологические маркеры динамики индивидуального стиля Д.Г. Лоуренса //Известия Смоленского государственного университета. – 2011. – №. 2. – С. 228-237.
11. Herdan G. The advanced theory of language as choice and chance. – Berlin : Springer, 1966. – С. 118.
12. Winter W. Styles as dialects //Lubomir Dole zel and Richard W. Bailey, editors, Statistics and Style. – 1962. – С. 3-9.
13. Журавлева Н. Н. Применение количественных методов при анализе стиля автора и решении проблем атрибуции //Вестник Тюменского государственного

университета. Гуманитарные исследования. Humanitates. – 2012. – №. 1. – С. 150-155.

14. Романов А. С. Методика и программный комплекс для идентификации автора неизвестного текста //Автореферат. Томск.-2010. – 26 с. – 2010.
15. Батура Т. В. Формальные методы определения авторства текстов //Вестник Новосибирского государственного университета. Серия: Информационные технологии. – 2012. – Т. 10. – №. 4. – С. 81-94.
16. Мамаев Н. К. и др. Метод Дельты Бёрроуза для определения авторства анонимных и псевдонимных литературных произведений на русском языке //Proceedings of the R. Piotrowski's Readings in Language Engineering and Applied Linguistics.—Санкт-Петербург: RWTH Aahen University. – 2018. – С. 1-14.
17. Burrows J. 'Delta': a measure of stylistic difference and a guide to likely authorship //Literary and linguistic computing. – 2002. – Т. 17. – №. 3. – С. 267-287.
18. Фоменко В. П., Фоменко Т. Г. Авторский инвариант русских литературных текстов //Методы количественного анализа текстов нарративных источников. М.: Ин-т истории СССР. – 1983. – С. 86-109.
19. Батура Т. В. Формальные методы установления авторства текстов и их реализация в программных продуктах //Программные продукты и системы. – 2013. – №. 4. – С. 286-295.
20. Lagutina K. et al. A survey on stylometric text features //2019 25th Conference of Open Innovations Association (FRUCT). – IEEE, 2019. – С. 184-195.
21. Kozhemyakina O. et al. The Question of Studying Information Entropy in Poetic Texts //Applied Sciences. – 2023. – Т. 13. – №. 20. – С. 11247.
22. Баевский В. С., Семенова Н. А. Эволюция лирического стиха как основа периодизации творческой биографии поэта: кластерный и корреляционный анализ (Пушкин, Гумилев, Пастернак) //Славянский стих. VII: Лингвистика и структура стиха/под ред. М.Л. Гаспарова, Т.В. Скулачевой. – М.: Языки славянской культуры. – 2004. – С. 421-436.
23. Андреев В. С. Динамика индивидуального стиля Э.А. По: стилеметрический подход //Труды института русского языка им. В.В. Виноградова. – 2017. – Т. 14. – С. 306-316.

24. Андреев В. С. Опыт стилеметрического подхода к проблеме сопоставления индивидуальных стилей //Известия Смоленского государственного университета. – 2017. – №. 4. – С. 183-189.
25. Андреев В. С. Вариативность индивидуального стиля как стилеметрическая проблема //Русская филология: ученые записки Смоленского государственного университета. – 2018. – Т. 18. – С. 227-235.
26. Кульбак С. Теория информации и статистика. – Москва: Наука, 1967.
27. Алябышева Ю. А., Веряев А. А. Фиксация авторских особенностей текста с использованием информационной меры Кульбака-Лейблера //Информатизация образования и методика электронного обучения: Материалы III Международной научной конференции – 2019. – С. 8-13.
28. А. С. Пушкин. Собрание сочинений в 10 томах. М.: ГИХЛ, 1959—1962.
29. Русская виртуальная библиотека: [Электронный ресурс]. URL: https://rvb.ru/pushkin/toc.htm (Дата обращения: 04.02.2024).
30. Фундаментальная электронная библиотека. Электронное научное издание “ПУШКИН”: [Электронный ресурс]. URL: https://feb-web.ru/feb/pushkin/default.asp (Дата обращения: 25.03.2024).
31. Marneffe, M.-C.; Manning, C.; Nivre, J.; Zeman, D. Universal Dependencies. Comput. Linguist. 2021, 47, 255–308.
32. Томашевский Б. В. Пушкин. – М.—Л., 1961.
33. Благой Д. Д. Творческий путь Пушкина. – М.—Л., 1950.
34. Гинзбург Л. Я. К постановке проблемы реализма в пушкинской литературе //Пушкин: Временник Пушкинской комиссии.-Москва, Ленинград: Издательство АН СССР. – 1936. – №. 2. – С. 387-401.
35. Фомичев С. А. Периодизация творчества Пушкина (к постановке проблемы) //Пушкин: Исследования и материалы. – 1982. – Т. 10. – С. 11-27.
36. Достоевский Ф. М. Т. 10 //Собрание сочинений в 10 томах. М., 1958, с. 442.

ПРИЛОЖЕНИЕ А

Распределение влияния признаков

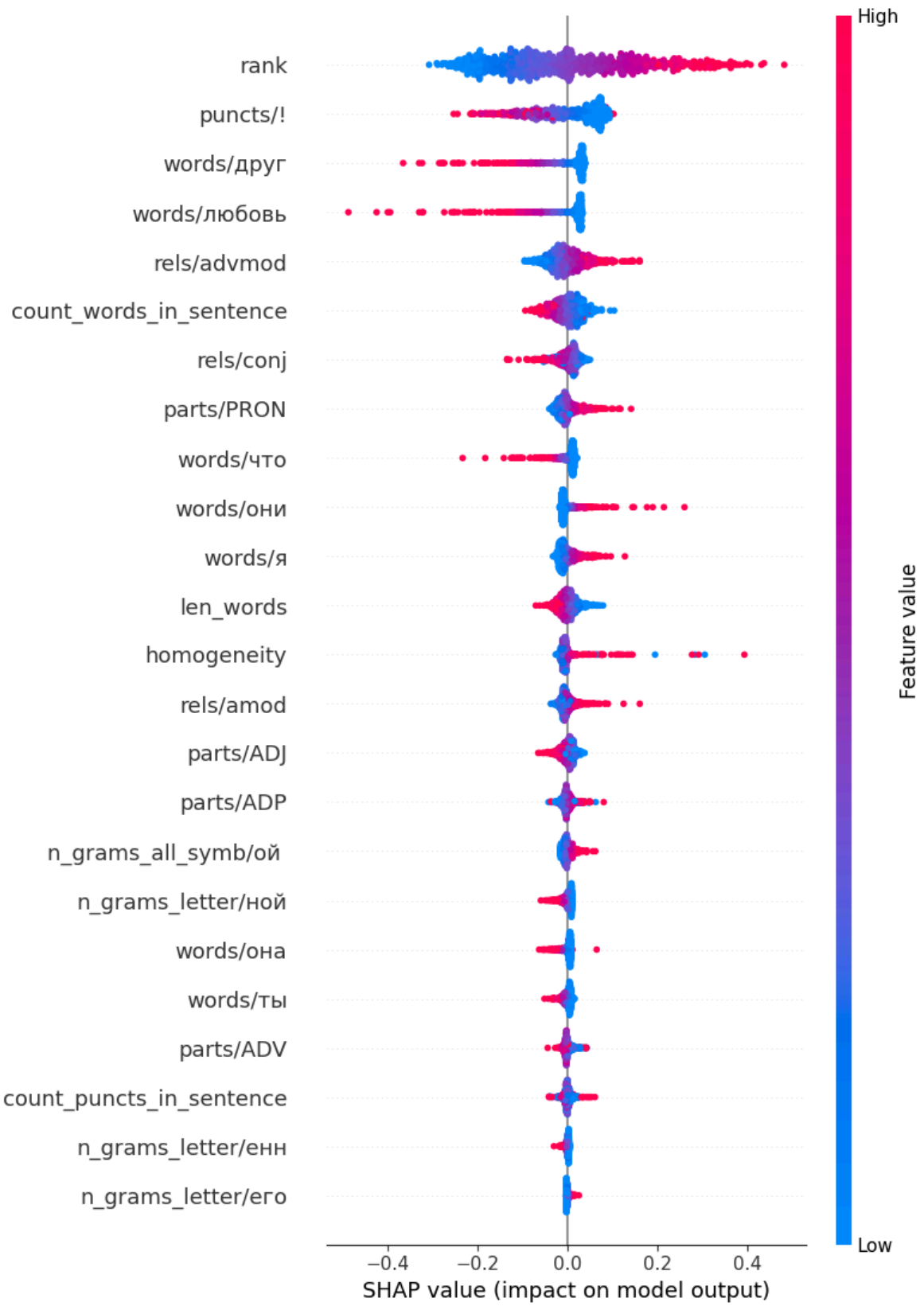

Рисунок А.1 – Распределение влияния признаков на результаты “model1”

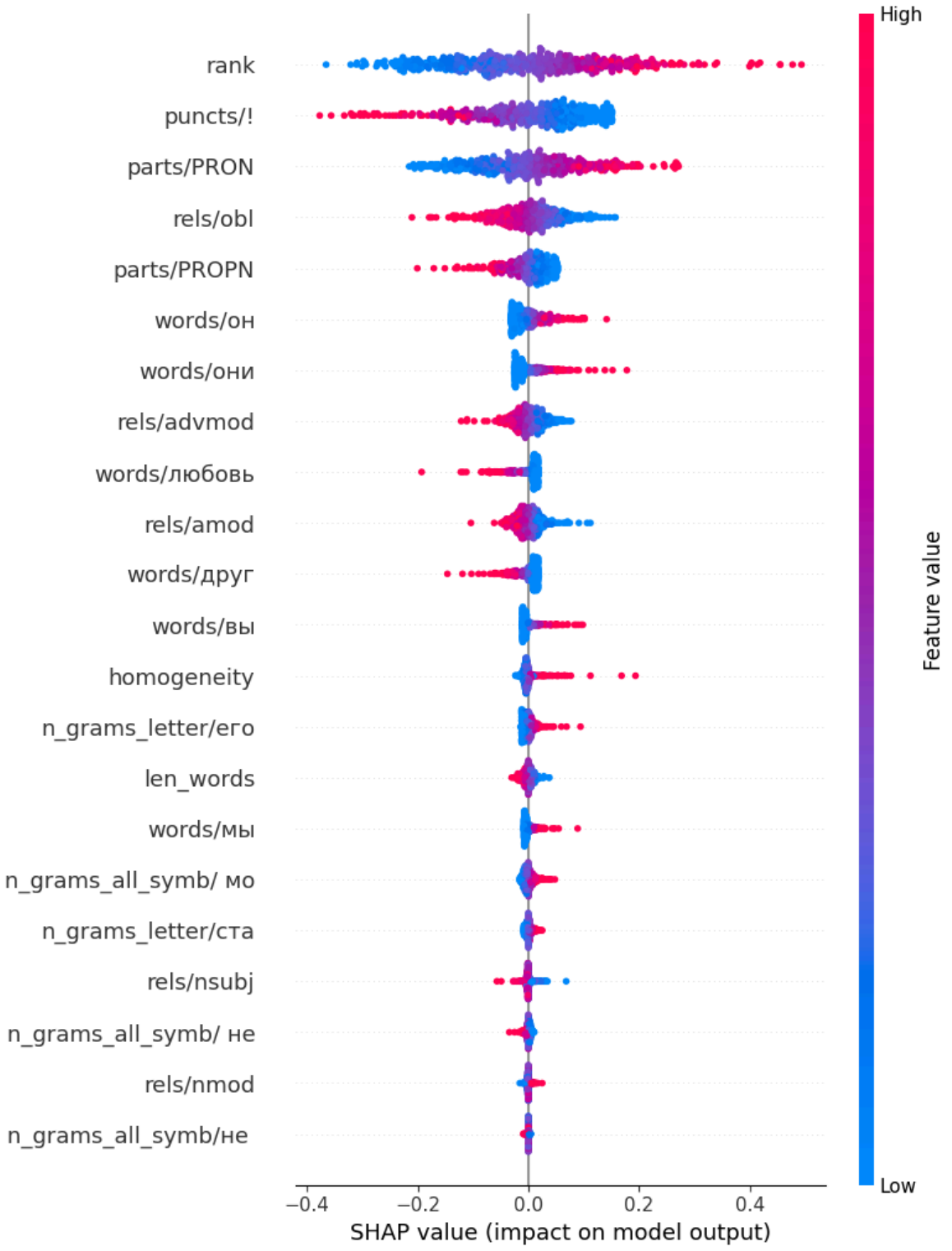

Рисунок А.2 – Распределение влияния признаков на результаты “model2”

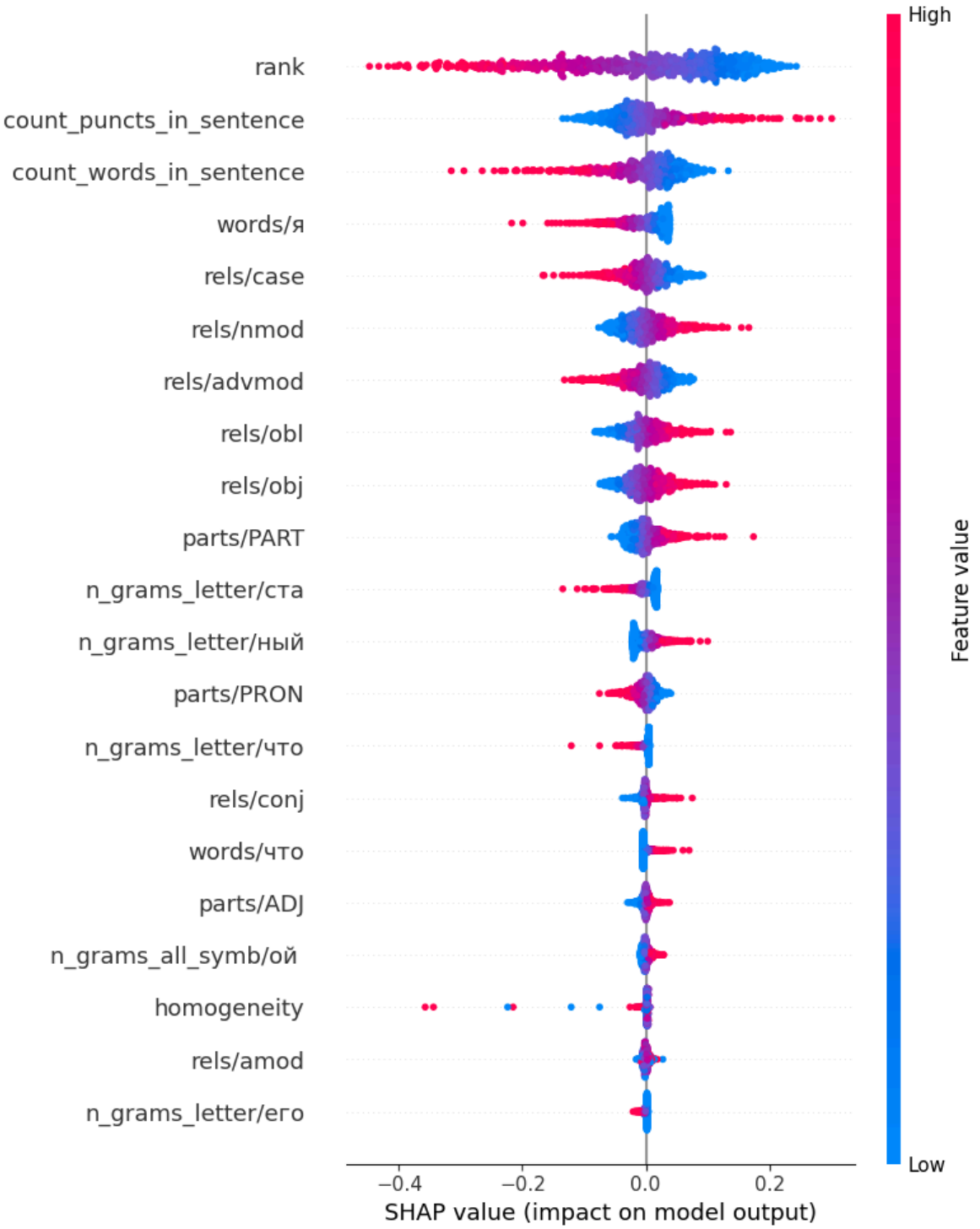

Рисунок А.3 – Распределение влияния признаков на результаты “model4”

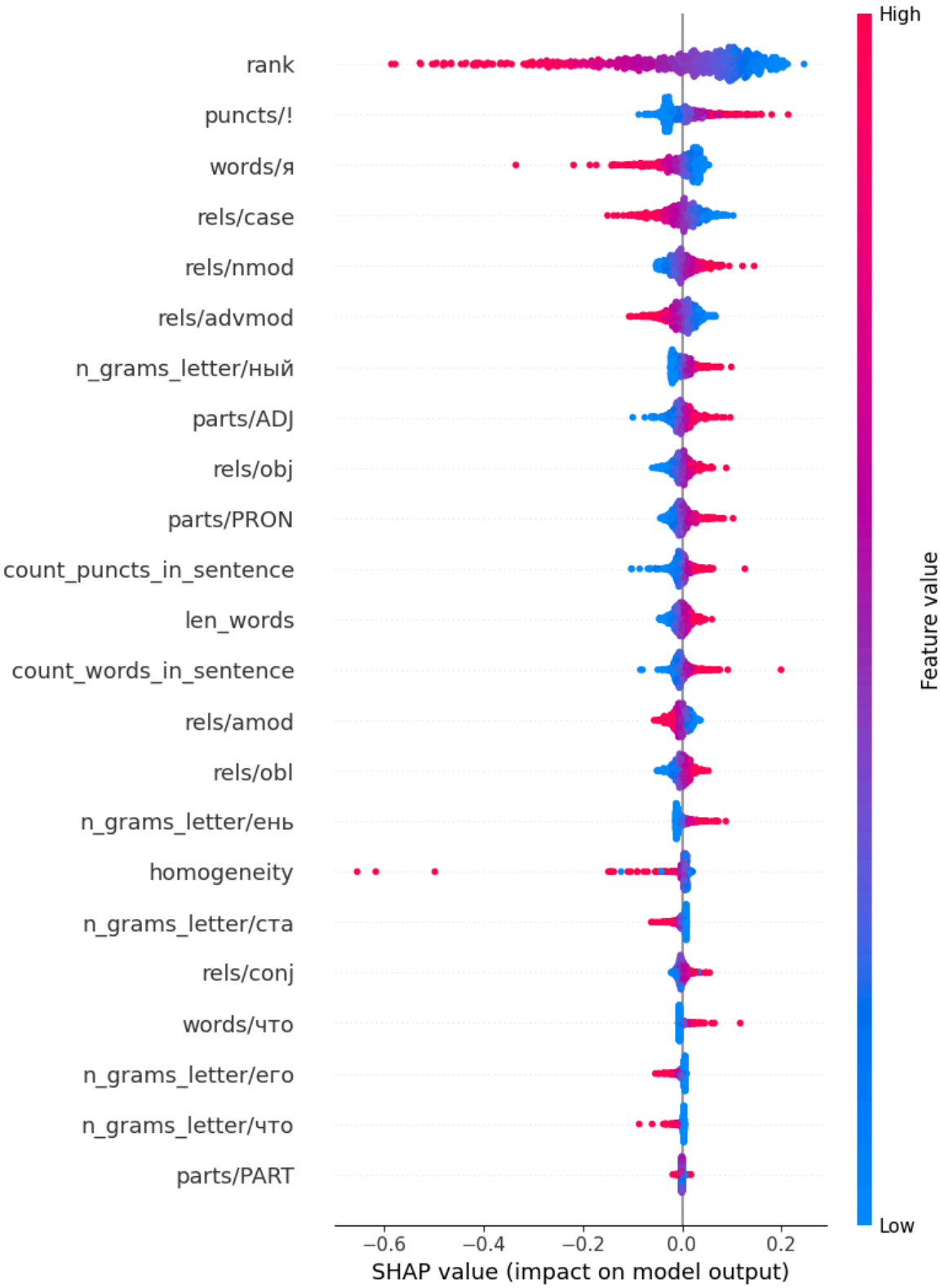

Рисунок А.4 – Распределение влияния признаков на результаты “model4-1”

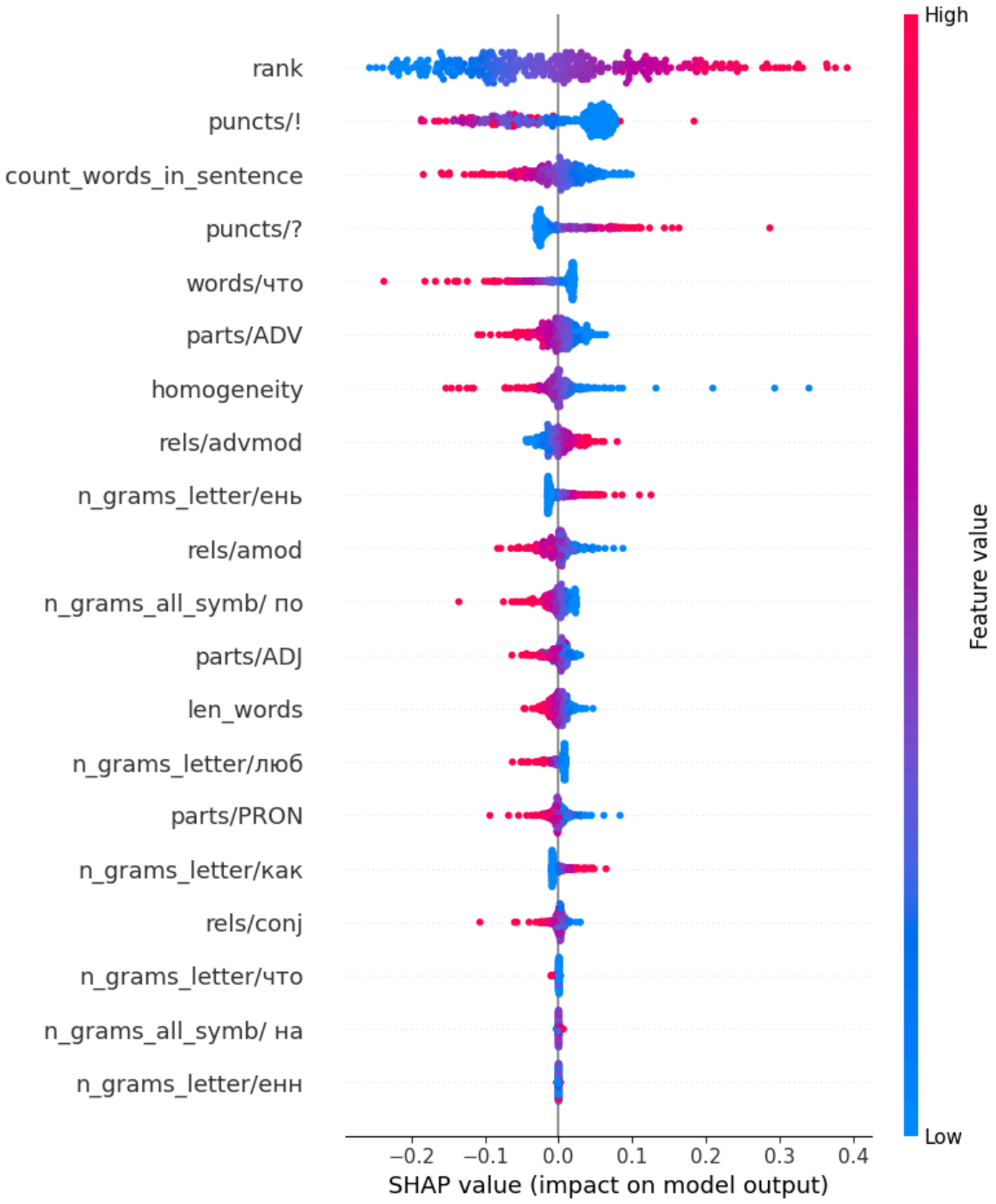

Рисунок А.5 – Распределение влияния признаков на результаты “model5”

ПРИЛОЖЕНИЕ Б

ПРОГРАММА «Анализ
изменения стиля Пушкина»
ОПИСАНИЕ ПРОГРАММЫ

Листов 10

Новосибирск, 2024

СОДЕРЖАНИЕ

АННОТАЦИЯ

В данном программном документе приведено описание программной системы “Анализ изменения стиля Пушкина”, которая реализует применение стилеметрических методов для исследования изменения авторской стилистики А.С. Пушкина.

Исходным языком программной системы является Python.

Основная функциональность программной системы – извлечение и формальное представление стихотворных произведений А.С. Пушкина и их анализ методами машинного обучения.

Оформление программного документа «Описание программы» произведено по требованиям ЕСПД (ГОСТ 19.402-78, ГОСТ 19.105-78).

1 Общие сведения

1.1 Обозначение и наименование программы

Программа обозначается и именуется “Анализ изменения стиля Пушкина”.

1.2 Программное обеспечение, необходимое для функционирования программы

Для функционирования программы необходим интерпретатор Python.

1.3 Языки программирования, на которых написана программа

Исходным языком программирования является Python.

2 Функциональное назначение

Разработанная программа предлагает пользователю выбрать набор признаков, на основании которых должно быть построено векторное представление текстов стихотворных произведений А.С. Пушкина, созданных в заданный пользователем период. Эти данные могут быть сохранены в файл или использованы для обучения нейронной сети.

3 Описание логической структуры

3.1 Алгоритм, структура программы и используемые методы

Программа состоит из модуля извлечения признаков, который осуществляет парсинг текстов произведений, содержащихся в файлах формата CoNLL-U, и модуля обучения нейронных сетей средствами фреймворка Keras.

3.2 Связи программы с другими программами

Для запуска программы необходимы интерпретатор языка Python. Программа использует фреймворк Keras для построения моделей машинного обучения.

4 Используемые технические средства

Программа может быть запущена на любом вычислительном устройстве, где установлен интерпретатор Python 3.

5 Вызов и загрузка

Запуск программной системы осуществляется с помощью вызова необходимой пользователю функции в интерактивном блокноте Jupyter или Python-скрипте.

6 Входные и выходные данные

Для функций извлечения признаков входными данными являются файлы формата CoNLL-U, содержащие текст произведения с его метаданными, выходными данными – массив векторов чисел.

7 Лист регистрации изменений

Таблица 1 – Лист регистрации изменений в программном документе «Описание программы»

Лист регистрации изменений									
Номера листов (страниц)					Всего листов (страниц) в документе	№ документа	Входящий № сопроводительного документа	Подпись	Дата
Номер изм.	измененных	замененных	новых	аннулированных					

ПРИЛОЖЕНИЕ В

ПРОГРАММА “Анализ
изменения стиля Пушкина”
РУКОВОДСТВО ОПЕРАТОРА

Листов 8
Новосибирск, 2024

СОДЕРЖАНИЕ

АННОТАЦИЯ

В данном программном документе приведено руководство оператора по применению и эксплуатации программной системы “Анализ изменения стиля Пушкина”, которая реализует применение стилеметрических методов для исследования изменения авторской стилистики А.С. Пушкина. В руководстве оператора описано функциональное и эксплуатационное предназначение программной системы, условия выполнения программы и описан процесс выполнения программы.

Оформление программного документа «Руководство оператора» произведено по требованиям ЕСПД (ГОСТ 19.505-79, ГОСТ 19.105-78).

1 Назначение программы

1.1 Функциональное назначение программы

Программа “Анализ изменения стиля Пушкина” предлагает пользователю выбрать набор признаков, на основании которых должно быть построено векторное представление текстов стихотворных произведений А.С. Пушкина, созданных в заданный пользователем период. Эти данные могут быть сохранены в файл или использованы для обучения нейронной сети.

1.2 Эксплуатационное назначение программы

Программная система предназначена для извлечения стилеметрических характеристик из текстов стихотворных произведений А.С. Пушкина.

1.3 Состав функций

• Извлечение стилеметрических признаков.

• Представление распределения признаков в виде диаграммы.

• Построение векторного представления текстов на основе выбранного набора признаков.

• Обучение модели машинного обучения.

2 Условия выполнения программы

2.1 Минимальный состав аппаратных средств

Для функционирования программной системы требуется:

• Оперативная память объёмом не менее 4 Гб;

• Жёсткий диск объёмом не менее 128 Гб;

• Монитор;

• Клавиатура;

• Мышь.

2.2 Минимальный состав программных средств

Для функционирования программной системы требуется наличие следующих программных средств:

• Любая операционная система, для которой существует реализация интерпретатора Python 3.

• Интерпретатор Python 3.

2.3 Требования к оператору

Оператор программы должен обладать практическими навыками работы с языком Python 3.

3 Выполнение программы

3.1 Загрузка и запуск программы

Запуск программной системы осуществляется с помощью вызова необходимой пользователю функции в интерактивном блокноте Jupyter или Python-скрипте.

3.2 Выполнение программы

Выполнение программы не предполагает никаких дополнительных операций.

3.3 Завершение работы программы

Завершение программы происходит автоматически после окончания вычислений.

4 Сообщения оператору

В ходе работы программной системы никаких сообщений для оператора не предусмотрено.

5 Лист регистрации изменений

Таблица 1 – Лист регистрации изменений в программном документе «Руководство оператора»

Лист регистрации изменений									
Номера листов (страниц)					Всего листов (страниц) в документе	№ документа	Входящий № сопроводительного документа	Подпись	Дата
Номер изм.	измененных	замененных	новых	аннулированных					