# СТИЛОМЕТРИЯ

НГУ, кафедра Математического Моделирования

Лисин Владислав

10 мая 2023 г.

## Содержание



## Аннотация

В данной работе представлено исследование авторского стиля А.С. Пушкина на основе сопоставления его поэтических текстов с текстами поэтов-современников. Целью данной работы является определение особенностей авторского стиля А.С. Пушкина с помощью методов машинного обучения. В данной работе описано построение нескольких классификаций на основе разных групп признаков, а также классификация на основе комбинированного набора признаков из разных групп. Также анализируется качество всех построенных классификаций; особое внимание уделено интерпретации нейросетевого решения и выявлению особенностей авторского стиля. В работе также используются современные архитектурные решения на основе сетей с механизмом внимания и глубокие рекуррентные сети с использованием предобученных векторных представлений слов.

## 1 Постановка задачи

Задача стилометрии в компьютерной лингвистике и обработке естественного языка заключается в автоматическом анализе текстов с целью идентификации авторства или стиля написания, используя различные стилистические и лингвистические характеристики текста.

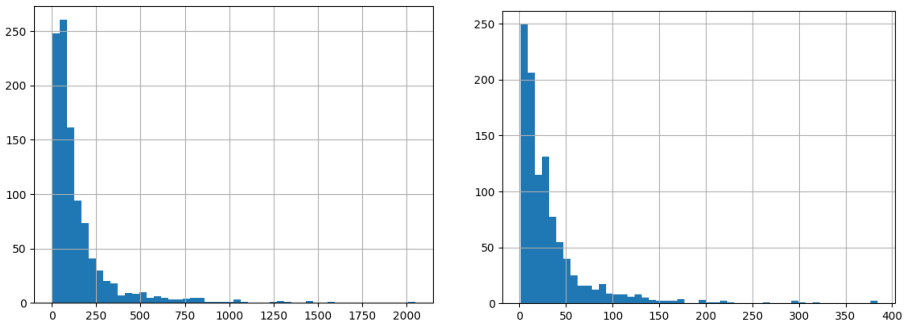
В данной работе для решения задачи стилометрии используются методы машинного обучения и анализа неструктурированных данных, которые позволяют выявлять свойства текста, такие как синтаксические зависимости, использование определенных слов и словосочетаний (n-gram'ы слов), пунктуация и другие характеристики. Эти уникальные признаки анализируются с помощью различных алгоритмов, которые позволяют построить модели для идентификации авторства или стиля написания. Конкретно в данной работе эти методы, в том числе нейростевые, будут использованы для того, чтобы определить отличительные особенности написания поэм известным отечественным писателем А.С.Пушкиным.

Задача стилометрии имеет широкий спектр применений, включая раскрытие авторства текстов, обнаружение плагиата, анализ литературных и исторических данных, а также анализ и сравнение текстов различных жанров и стилей. Она также может быть полезной в контексте анализа социальных медиа, где можно использовать стилометрию для идентификации фейковых аккаунтов или анализа политических кампаний.

# 2 Предобработка, метрики и модели

## 2.1 Корпус текстов

В распоряжении есть корпус текстов (поэм) А.С.Пушкина и его современников-писателей: Батюшков, Боратынский, Жуковский, Дельвиг. Корпус содержит 1031 поэму, размером от 2-ух строчной безымянной поэмы Жуковского до 384-ех строчной поэмы "Странствователь и домосед"Батюшкова. Длина поэм также весьма различается, самые малые произведения содержат не более 9 слов, наибольшая насчитывает 2047 слов.




(a) Y: поэмы, X: слова

(b) Y: поэмы, X: строки

Рис. 1: Описание набора данных: (а) гистограмма распределения стихов по количеству строк; (б) Гистограмма распределения стихов по количеству слов.

Распределение поэм по классам, если считать поэмы А.С.Пушкина положительным, а все остальные отрицательным, также неравномерное. С значительной разницей в почти три десятка процентов доминирует поэзия современников Пушкина.

На основе этих данных произведён анализ отличительных особенностей авторского стиля Пушкина. Некоторые замечания по представленной статистике и, в общем, по датасету. Во-первых, корпус достаточно мал, особенно в контексте определения авторского стиля, что затрудняет использование современных методов решения поставленной задачи как с помощью классических алгоритмов машинного обучения, так и (на самом деле, тем более) при помощи нейронных сетей. Во-вторых, данные является несбалансированными, процент поэм Пушкина от общей массы составляет около 31 процента, что, конечно, мало, но не катастрофично. Заметим также, что это обстоятельство не позволяет перепрофилировать проблему из бинарной классификации в поиск аномалий в данных.

Таблица 1: Результаты классификации на основе слов и их n-grams

| Автор | Число поэм | Процент от всего корпуса |
|---|---|---|
| Пушкин | 323 | 0.3132 |
| Жуковский | 210 | 0.2036 |
| Дельвиг | 199 | 0.1930 |
| Боратынский | 197 | 0.1910 |
| Батюшков | 102 | 0.0989 |

Эта особенность будет, в некоторой мере, влиять на дальнейший процесс обучения моделей, а также метрики их оценивания. В-третьих сами поэмы весьма различны по размерам, некоторые из них недостаточно продолжительны и не содержат потенциальных отличительных признаков определенного автора, другие же, наоборот, избыточны и могут быть укорочены в угоду производительности без потери точности на метриках, однако позволив сократить размерность пространства признаков, например, для Word2Vec представлений. Заметим, что как первая, так и вторая группа поэм не многочисленны в общей массе произведений. Делаем вывод, что основной проблемой для нас, всё же, будет являться первая - малый объем корпуса текстов.

## 2.2 Определение групп признаков

С точки зрения филологии индивидуальный (письменный) стиль представляет собой сложное понятие, отражающее общественно-исторический характер, этнические, психологические, морально-этические особенности автора [1]. Разные авторы выделяют разные уровни анализа текста [1, 2], которые можно использовать в стилометрии. В данной работе использованы следующие уровни: пунктуационный, синтаксический и лексический.

Были выделены следующие группы признаков, каждая из которых соответствует определенному филологическому уровню:

Пунктуация - важная группа признаков, которая помогает выявлять эмоциональный окрас текста, а также определяет четкую структуру произведения уникальную для данного автора. Поэтому, знаки препинания, такие как точки, запятые, двоеточия, вопросительные и восклицательные знаки и так далее, могут быть полезны для определения авторства текста или его стиля. В данной работе сформирована группа признаков, учитывающая распределение знаков препинания (частотность), а также использование определенных знаков препинания в тексте: определенные знаки препинания, такие как восклицательные знаки или вопросительные знаки, могут указывать на наличие эмоций в тексте или на то, что автор хочет подчеркнуть определенные мысли, к ним же отнесём использование кавычек и многоточий. Очевидно, что данная группа признаков относится к пунктуационному уровню.

Группа признаков "части речи"позволяет анализировать использование определенных слов и их роли в тексте. Части речи - это классификация слов на основе их грамматических характеристик, таких как падеж, число, время и т.д. Датасет опирается на следующие особенности: количество слов каждой части речи (это может помочь выявить, какие части речи автор чаще всего использует в тексте и как это влияет на стиль письма) и частотность употребления определенных частей речи: некоторые части речи, такие как глаголы, могут указывать на наличие действия в тексте, а прилагательные могут указывать на наличие описания.

Группа признаков синтаксические зависимости позволяет анализировать связи между словами в тексте и определять структуру предложения, выделять ключевые идеи в произведении. В данном датасете основное внимание отдаётся структуре предложения: анализ синтаксических зависимостей может помочь выявить

структуру предложения и определить, какие слова являются подлежащими, сказуемыми, дополнениями и т.д. Также выявляются особенности использования активного и пассивного залога, вводных слов и фраз, а также сочинительных и подчинительных союзов. Группы признаков "части речи"и синтаксические зависимости относятся соответственно к синтаксическому уровню

Группа признаков n-gram'ы слов и символов относятся к последнему лексическому уровню и во многом схожи по построению и выявлению отличительных особенностей текста. В данной работе использовано упрощенное представление текста в виде 'Мешка слов' BoW (Bag of Words), то есть мы подсчитываем частотность встречаемости определенного слова в конкретной поэме относительно всех произведений. Проблема такого подхода в том, что чем больше исследуемый корпус текстов, тем больше будет размерность формируемого датасета. В таком случае стараются прибегать к урезанию тех слов, которые встречаются достаточно редко (ниже какого-то заданного порога), либо нормализации слов посредством стэмминга или лемматизации, то есть приведения слов к их начальной форме в случае лемматизации, либо к сохранению лишь главной корневой части слова - стэмминг. Однако в данном подходе мы лишаемся абсолютно всех словоформ, что может быть весьма критичным при выявлении авторских черт написания текста. Решить эту проблему может использование n-gram символов в комбинации с лемматизированным (стэммизированым) датасетом слов. Большинство слов формируются посредством суффиксов, окончаний и приставок и большинство из этих словообразующих кирпичиков укладываются в диапазон n-gram от 1-5, однако сказать наверняка, какой интервал является самым оптимальным нельзя. В следующей главе эмпирическим способом будет установлен наилучший вариант. Также в BoW представлении используем статистическую меру TF-IDF (term frequency-inverse document frequency), которая учитывает, что часто встречающиеся слова в документе не всегда являются ключевыми или информативными, поэтому оценка важности термина в документе основывается не только на его частотности в документе, но и на том, насколько часто этот термин встречается во всей коллекции документов.

Также имеет смысл проанализировать текст отдельно относительно служебных слов и местоимений. Служебные слова и местоимения являются частями речи, которые не несут смысловой нагрузки, но служат для связи слов в предложении. Они могут быть использованы для создания определенного стиля и могут отличаться в зависимости от автора или периода написания. Например, некоторые авторы могут предпочитать использовать более формальные или устаревшие формы служебных слов и местоимений, что может быть связано с их стилем письма.

## 2.3 Метрики

Ввиду несбалансированности и немногочисленности данных выбраны следующие метрики качества классификации моделей: AUC ROC, Balanced Accuracy, F1.

AUC ROC - Area Under the Receiver Operating Characteristic Curve является хорошей метрикой для несбалансированных данных, так как она учитывает долю верных положительных ответов и долю ложных положительных ответов, что позволяет оценить качество классификации в случае, когда количество объектов разных классов сильно отличается.

Balanced Accuracy (BA) - представляет собой среднее арифметическое между долей верных положительных ответов (True Positive Rate) и долей верных отрицательных ответов (True Negative Rate). В случае сильного дисбаланса классов, когда один класс составляет большую часть выборки, точность классификации может быть высокой только благодаря высокой доле верных отрицательных ответов для большего класса и низкой доле верных положительных ответов для меньшего класса.

F1 - мера является гармоническим средним между точностью и полнотой. В случае сильного дисбаланса классов, когда один класс составляет большую часть выборки, точность классификации может быть высокой только благодаря высокой доле верных отрицательных ответов для большего класса и низкой доле верных положительных ответов для меньшего класса. Полнота же может быть низкой из-за малого количества положительных объектов в выборке. F1-мера позволяет учесть оба этих аспекта и оценить качество классификации для обоих классов равномерно.

## 2.4 Методы

Далее используются перечисленные в таблице [2] методы машинного обучения, среди них присутствуют как классические методы машинного обучения, а именно logistic regression (LR), support vector machine (SVM), так и более сложные ансамблевые extra-tree classifier (ET), gradient boosting for classification (GB), random forest classifier (RF), CatBoost classifier (CB), AdaBoost classifier и нейросетевой метод multi-layer perceptron (MLP). Выбор именно таких методов позволит добиться более высоких метрик на отдельно взятых группах признаков, учитывая их линейную или сложную нелинейную структуру. Подбор гиперпараметров для каждой модели производится с применением приемов решетчатого поиска и перекрестной проверки, что значительно повышает время обучения и поиска лучшей модели среди одного класса моделей с разными гиперпараметрами, однако обеспечивает максимально возможную степень обобщающей способности модели на новых данных.

Таблица 2: Список моделей

| Аббревиатуры для моделей | Название библиотеки | Имя класса |
|---|---|---|
| ET | scikit-learn | ensemble.ExtraTreesClassifier |
| RF | scikit-learn | ensemble.RandomForestClassifier |
| AB | scikit-learn | ensemble.AdaBoostClassifier |
| GB | scikit-learn | ensemble.GradientBoostingClassifier |
| CB | CatBoost | CatBoostClissifier |
| SVC | scikit-learn | svm.SVC |
| MLP | scikit-learn | neural_networks.MLPClassifier |
| LR | scikit-learn | linear_model.LogisticRegression |

Сами методы, при помощи которых будет производится классификация на поэмы Пушкина и всех остальных не даст нам ответа на поставленный вопрос, однако веса, формируемые внутри модели во время её обучения могут решить эту задачу. Интерпретация важности весов для определенного класса не является новой задачей, в scikit-learn уже имеются встроенные инструменты определения важности модели для многих моделей, включая RF, ET, GB и другие. Однако сам способ весьма наивен и не раскрывает всей нюансов, поэтому далее будет использован метод SHAP из одноименной библиотеки. SHAP (SHapley Additive exPlanations) позволяет интерпретировать прогнозы модели машинного обучения, используя алгоритм Шэпли (Shapley values) из теории игр для вычисления важности признаков. Этот алгоритм рассчитывает, насколько каждый признак влияет на прогноз модели, путем перебора всех возможных комбинаций признаков и вычисления их вклада в прогноз. Это позволяет получить точные оценки важности каждого признака.

# 3 Классификация

Замечание. Далее будет видно, что метрика F1 показывает заметно более низкие показатели чем AUC ROC или Balanced Accuracy. Это связано с тем, что метрика F1 учитывает точность (precision) и полноту (recall) модели, а так как выборка несбалансирована, то это влечёт сложность в достижении высокой F1-меры, поскольку модель сильно перекашивается в сторону большего класса, что приводит к низкой полноте и высокой точности (мало ложных положительных примеров, но много ложно отрицательных). И всё это ещё сильнее усугубляется на фоне малой выборки текстов (в вашем случае около 1031). Чтобы сделать данный эффект более мягким используем методы балансировки выборки, такие как oversampling или undersampling, однако раздуть саму выборку, то есть аугментировать наши данные, в силу особенность задачи стилометрии, не представляется возможным. Исходя из предыдущих рассуждений, в дальнейшем главной задачей будет именно максимизация F1 метрики. Каждая группа признаков разделена на тренировочный и тестовый набор в соотношении 80 к 20. В случае глубоких нейронных сетей данные делятся на тренировочный, валидационный и тестовый наборы в соотношении 70:15:15 соответственно.

## 3.1 Классификация на основе слов и их n-grams

Первая группа для классификации - n-gram'ы слов. В таблице [3] представлены модели, показавшие лучшие результаты. Наиболее хорошо для данного датасета работает логистическая регрессия и нейросетева модель.

Таблица 3: Результаты классификации на основе слов и их n-grams

| n-grams, тип | Число признаков | Метод | AUC ROC | Balanced Accuracy | F1 |
|---|---|---|---|---|---|
| 1, Vectorized | 31491 | LR | 0.80682 | 0.64844 | 0.50191 |
| | | MLP | 0.81132 | 0.68518 | 0.56465 |
| 1-2, Vectorized | 149272 | MLP | 0.81938 | 0.68931 | 0.55745 |
| | | LR | 0.80984 | 0.66243 | 0.53304 |
| 1, Lemmatized | 13414 | LR | 0.83106 | 0.70685 | 0.53980 |
| | | GB | 0.79420 | 0.66912 | 0.55829 |

Отметим, что вариант со стандартными 1-gram слов показали наилучший результаты, однако модель, обученная на лемматизированном датасете, показала результат лишь на 1 пункт ниже, но при этом размерность признакового пространства более чем в 2 раза меньше, что значительно лучше в вычислительном плане. Поэтому в дальнейшем именно эти два датасета будут использоваться в создании итоговой комбинации для финального набора данных. Набор слов, который использует автор в значительной мере должен характеризовать его стиль письма, однако в данной ситуации заметен недостаток примеров в обучающей выборке, из-за чего модели не успевают изучить характерный для данного автора набор слов. Это обстоятельство особенно заметно на метрике F1, в то время как AUC ROC стабильно имеет весьма высокие показатели.

## 3.2 Классификация на основе символов и их n-grams

В данной работе использовались n-gram'ы слов в диапозоне от 1 до 5, так как морфологические элементы слова, такие как приставка, суффиксы, окончания, имеют важную роль в определении авторского стиля, а окончания имеют особенную роль в написании поэтических произведений. Результаты обучения моделей на разных n-gram'ах символов представлены в таблице [4], из неё видно, что, во-первых, по сравнению с n-gram'ами слов метрики улучшились, а во-вторых,

метрики, несмотря на разные n-gram'ы, весьма схожи, однако и здесь можно выделить фаворита - 2-3-gram'ы имеют наивысший показатель метрики F1, которая, как мы ранее отмечали, для нас является наиболее важной. В данной группе признаков весьма успешно использовались ансамблевые модели на основе деревьев принятия решений, так лучшими моделями оказались AdaBoost и GradientBostting.

Таблица 4: Результаты классификации на основе символов и их n-grams

| n-grams | Число признаков | Метод | AUC ROC | Balanced Accuracy | F1 |
|---|---|---|---|---|---|
| 1 | 71 | GB | 0.84405 | 0.73384 | 0.63702 |
| | | MLP | 0.83968 | 0.72260 | 0.61975 |
| 1-2 | 1365 | MLP | 0.85780 | 0.78345 | 0.63820 |
| | | LR | 0.85030 | 0.72953 | 0.61194 |
| 2 | 1294 | LR | 0.83981 | 0.73334 | 0.62224 |
| | | GB | 0.82731 | 0.74659 | 0.64829 |
| 2-3 | 13058 | AB | 0.80994 | 0.77820 | 0.68750 |
| | | GB | 0.84543 | 0.75221 | 0.65000 |
| 3 | 11764 | MLP | 0.86754 | 0.72441 | 0.61215 |
| | | AB | 0.82506 | 0.73616 | 0.63386 |
| 3-4 | 75267 | AB | 0.85986 | 0.72872 | 0.62904 |
| | | MLP | 0.87092 | 0.70654 | 0.61831 |
| 2-4 | 76561 | LR | 0.84868 | 0.73103 | 0.65538 |
| | | MLP | 0.84984 | 0.75375 | 0.64389 |
| 2-5 | 271616 | LR | 0.84783 | 0.73283 | 0.63649 |
| | | MLP | 0.83304 | 0.72494 | 0.62849 |

Можно сделать некоторые выводы. Увеличение правой границы n-gram не ведёт к увеличению обобщающей способности моделей, таким образом n-gram'ы выше 4 приводят лишь у ухудшению показателей метрик. Расширение диапазона n-gram также не всегда является гарантом лучшей обучаемости модели, таким образом модели, обученные на наиболее широком диапазоне 1-5, а также 1-4, показали результаты значительно худшие, чем те, что представлены в таблице. Большинство признаков на основе 1-gram, которые имели наиболее важное значение, повторяют пунктуационные признаки и потому не имеют новой важной информации для итогового датасета, поэтому в дальнейшем было принято решение не использовать одиночные символьные признаки.

## 3.3 Классификация на основе знаков препинания

Пространство признаков данного датасета является самым маленьким из представленных, всего 11 признаков, включающих основные наиболее используемые знаки препинания. Данное обстоятельство позволило особенно тонко настроить гиперпараметры меделей, вследствие чего, даже несмотря на малое количество информации содержащаяся в данных, метрики качества весьма высокие.

Таблица 5: Результаты классификации на основе пунктуации

| Группа признаков | Метод | AUC ROC | Balanced Accuracy | F1 |
|---|---|---|---|---|
| Частоты знаков препинания | ET | 0.77658 | 0.72622 | 0.59615 |
| Частоты знаков препинания | MLP | 0.81881 | 0.76165 | 0.65346 |

Как показано в таблице [5] лучше всего себя показал перцептрон (полносвязная нейронная сеть), следом за ним с большим отрывом идут экстремальные леса и все остальные модели. Наиболее важным среди всех знаков оказался восклицательный знак. Классификация на основе знаков препинания может быть весьма полезна, однако, эта группа признаков не является всеобъемлющей, так как знаки препинания не всегда являются надежными индикаторами смысла и тонуса текста, и могут способствовать запутыванию модели и последующему её переобучению. Поэтому надёжнее всего использовать данный датасет как вспомогательный в комбинации с другими.

## 3.4 Классификация на основе частей речи и синтаксические зависимости

Группы признаков "части речи"и синтаксические зависимости определяют синтаксический уровень анализа текста, поэтому логичным шагом будет попробовать объединить их в целях выявления более глубоких закономерностей в предложениях. Тем не менее, прежде обучим наши модели на этих данных раздельно и получим соответствующие метрики.

Таблица 6: Результаты классификации на основе частей речи и синтаксических зависимостей

| Группа признаков | Число признаков | Метод | AUC ROC | Balanced Accuracy | F1 |
|---|---|---|---|---|---|
| Части речи | 18 | GB | 0.63251 | 0.57128 | 0.42333 |
| Синтаксические связи | 40 | MLP | 0.71335 | 0.65256 | 0.51598 |
| Части речи и синтаксические связи | 58 | SVC | 0.78120 | 0.69842 | 0.56670 |

Показатели качества на группе синтаксических признаков показывают посредственный, хоть и не ужасный результат (см. таблица), что позволяет предположить, что в комбинации с другими группами признаков, они могут улучшиться. Модели, обученные на группе признаков "части речи показывают неутешительный результат, который на несколько пунктов ниже, чем у предыдущего датасета. Лучше всего справился градиентный бустинг, показав статистику F1 в пределах 0.42333. Данные результаты показывают несостоятельность данной группы признаков в задаче выявления особенностей написания текстов. Однако в сочетании с признаками по синтаксическим зависимостям метрики моделей сильно возрастают, таким образом они хорошо дополняют друг друга, показывая прирост в 5 процентов для F1 по сравнению с результатами, которые показали модели, натренированные чисто на синтаксических зависимостях, 0.51598 на MLP и 0.56680 на SVC соответственно. Данный случай показывает, что некоторые признаки могут "хорошо сыграть"лишь в некотором сочетании с другими, возможно, также не столь очевидными признаками.

## 3.5 Классификация на основе слов служебных частей речи

Данный датасет во многом будет совпадать с группой признаков по словам, и по сути, является её подмножеством. Однако использование служебных слов, как и звуков, в меньшей степени является предметом авторской воли, а значит имеет смысл проанализировать их отдельно от остальных слов. Учет только служебных слов позволяет констатировать сведения о стилистических особенностях текстов конкретного автора без учета смыслового содержания произведений.

При достаточной точности классификации этот подход будет предпочтительнее предыдущего варианта, основанного на всех словах. В таблице [7] представ-

Таблица 7: Результаты классификации на основе слов служебных частей речи

| Группа признаков | Число признаков | Метод | AUC ROC | Balanced Accuracy | F1 |
|---|---|---|---|---|---|
| Предлоги, союзы, частицы, междометия и местоимения | 264 | LR | 0.76483 | 0.66150 | 0.53093 |
| Предлоги, союзы, частицы, междометия и местоимения | 264 | GB | 0.70760 | 0.61152 | 0.49307 |

лены результаты классификации, заметим, что наилучший результат показала простейшая логистическая регрессия, однако, результаты оказались хуже, чем при классификации по всем словами и глобально они не являются достаточно значимыми, чтобы учитывать данный набор признаков в итоговом датасете. Такие результаты также можно объяснить небольшим размером поэтических произведений.

## 3.6 Итоговый датасет

Далее представлен итоговый датасет с комбинацией признаков, показавший лучшие метрики на выбранных моделях. Датасет состоит из 26541 признаков, а именно: лемматизированные 1-gram-tf-idf-слов, 2-3-gram символов, признаки частей речи и синтаксических зависимостей, а также частоты знаков препинания. В таблице [8] представлены результаты тестирования пары моделей, натренированных на данном наборе данных. Заметим, что датасет, составленный из комбинации признаков, в значительной мере повышает обобщающую способность модели и соответственно её метрики качества.

Таблица 8: Результаты классификации на основе итогового датасета

| Метод | AUC ROC | Balanced Accuracy | F1 |
|---|---|---|---|
| MLP | 0.89091 | 0.77484 | 0.72790 |
| GB | 0.86917 | 0.79153 | 0.71068 |

Также в целях повышения качества классификации был использован метод ансамблирования, который хорошо подходит, в случае обучения на разных группах признаков - стекинг. В данном случае, мы берём модели, которые показали наилучшие метрики на определенной группе признаков и которые являются базовыми, то есть будут служит основой для обучения итоговой мета-модели. В таблице [9] отображены результаты эксперимента, которые показывают небольшой прирост метрик качества. В поле методы отображена используемая мета-модель.

Таблица 9: Результаты классификации на основе стекинга

| Метод | AUC ROC | Balanced Accuracy | F1 |
|---|---|---|---|
| LR | 0.89697 | 0.81668 | 0.73413 |
| RF | 0.84612 | 0.80782 | 0.71175 |

Однако в данном случае становится затруднительно интерпретировать веса модели и анализировать важность признаков для позитивного и негативного классов. Поэтому данный эксперимент является сугубо статистическим.

# 4 Результаты

Теперь интерпретируем результаты, полученные на моделях, обученных на основе итогового датасета, используя метод SHAP. Обратим внимание на рис. 2, на данном графике каждая точка представляет отдельный объект из тестового набора данных, а каждый признак представлен горизонтальной линией на графике, в данном случае представлено 25 признаков, вносящих наибольших вклад в определение класса. Вертикальное положение точки на графике показывает значение SHAP для данного объекта, а цвет точки указывает на значение соответствующего признака для этого объекта. Если точка красная, то значение признака для этого объекта выше, чем среднее значение признака в тестовом наборе, а если точка синяя, то значение признака для этого объекта ниже, чем среднее значение признака в тестовом наборе. Положительные значения SHAP для данного объекта означают, что данный признак вносит положительный вклад в принятие решения моделью, а отрицательные значения SHAP означают отрицательный вклад. Чем больше значение SHAP для данного признака и объекта, тем больший вклад этот признак вносит в принятие решения моделью.

Для удобства введём следующие обозначения:

- word_ - словесные признаки
- letter_ - символьные признаки
- punct_ - пунктуационные признаки
- speech_ - признаки по частям речи
- syntax_ - признаки по синтаксическим зависимостям
- _space_ - пробел (промежуток между словами)
- _ENTER_ - переход на новую строку ("/n")

В названиях признаков по n-граммам символов пробелы заменяются на служебное слово _space_. Значительная доля всех важных признаков относится к символьной группе признаков, их можно разделить на две категории. Те что, захватывают пунктуационные признаки косвенно или напрямую, например, признак троеточие '...' является 3-gram'ой или признак 'т,', который вероятно показывает предпочтение авторов после глагола в настоящем времени (думает, делает, мечтает) использовать сложные грамматические обороты и на те, что обозначают те самые грамматические кирпичики: приставки, окончания и т.д.

Интересно отметить, что среди первых 25 значимых признаков есть только 4 признака по словам и самым значимым из них является 'Свобода'. И это неудивительно, ведь тема свободы является одной из наиболее значимых и постоянных тем в творчестве Александра Сергеевича Пушкина.

В длинном списке важных признаков нашлось лишь одно место для синтактической уровня анализа текста, а именно признак, указывающий на частоту употребления предлогов из речевой группы признаков, он оказался на 14 позиции.

Доминирующими чертами авторского стиля оказались признаки, характеризующие предпочтения авторов в знаках препинания, согласно которым Пушкин, в отличие от других авторов, предпочитал использование точек и многоточий, а не восклицательных знаков, отличающих других авторов. Учитывая итог по словам, по которым Пушкин тоже не был склонен к употреблению междометий, можно отметить, что другие поэты пушкинской эпохи были гораздо эмоциональнее и писали с более выраженным восторженным слогом.

При анализе важных признаков, описывающих отношения, следует отметить, что утверждение о том, что пушкинскому творчеству не свойственно широкое употребление служебных слов, здесь подтверждается значениями признаков 'speech_PREP', а также 'letter_o_space', 'word_вновь' и 'letter_space_к_space_'.

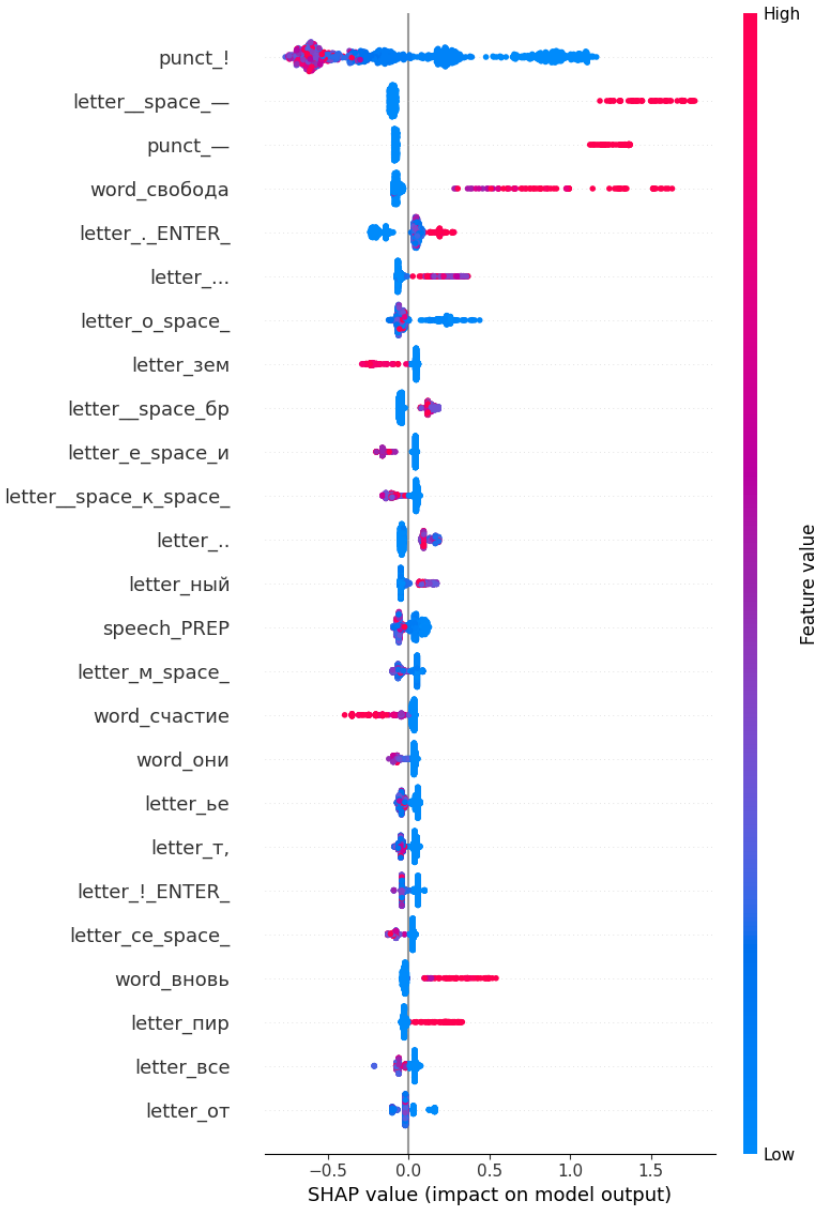



Рис. 2: Распределение влияния признаков на результаты классификации по группам признаков.

# 5 Extra

## 5.1 Плотные векторные представления и рекуррентные нейронные сети

В данной главе будет описан метод решения задачи классификации при помощи полноценной глубокой рекуррентной нейронной сети, написанной на базе библиотеки keras, используя специальные блоки GRU и LSTM. Особенность глубоких сетей заключается в огромном количестве параметров, которые необходимо настроить в процессе обучения. Чтобы обучить такое количество параметров нужны действительно большие данные, именно поэтому для задач с небольшим набором данных популярно решение, которое основано на подходе transfer learning (перенос обучения) и реализовано в следующей главе с архитектурой Bert. Особенностью же рекуррентных сетей является формат входных данных, теперь мы представляем каждую поэму как некоторую последовательность, то есть учитываем порядок следования данных на вход нейронам. Вторая особенность реку-

рентных нейронных сетей заключается в наличие так называемой памяти, которая позволяет запоминать информацию о предыдущих токенах в больших последовательностях. Мы отказываемся от простейшей SimpleRNN сети, так как она не способна сохранять информацию о длинных последовательностях, что, в условиях когда поэма может достигать длины в более чем 2000 тысячи токенов, недопустимо. Однако токенизировать последовательность недостаточно, каждый токен необходимо представить как некоторый вектор. Было принято решение использовать плотные векторные представления выбранной размерности, так как они компакты и обладают полезными свойствами. Например, свойство семантической близости, которое означает, что близкие по значению слова имеют близкие векторные представления, может быть весьма полезно в определении авторского стиля. В данной работе используются следующие алгоритмы получения векторных представлений: FasText, Word2Vec, GloVe. Так как для обучения векторных представлений каждого из перечисленных выше методов необходимы огромные корпуса текстов, было принято решение использовать уже предобученные на художественных корпусах текстов модели или, по крайней мере, имеющих некоторый процент художественных произведений в тренировочном датасете. В таблице [10] представлены результаты тестирования полученных моделей, имеющих наилучшие показатели метрик.

Таблица 10: Результаты классификации на основе рекуррентной нейронной сети

| Метод | AUC ROC | Balanced Accuracy | F1 |
|---|---|---|---|
| LSTM + GloVe | 0.83475 | 0.73672 | 0.63758 |
| LSTM + FasText | 0.81306 | 0.71948 | 0.61314 |
| LSTM + Word2Vec | 0.80192 | 0.69581 | 0.57295 |

Можно заметить, что результаты на несколько пунктов хуже, чем полученные из итогового датасета и имеют точность приблизительно на уровне символьной группы признаков. Наилучший результат имеет нейронная сеть с рекуррентным блоком LSTM и векторным представлением GloVe, что весьма логично, ведь он целиком обучен на русской художественной литературе. В то же время Word2Vec имеет наихудший результат и связано это, в первую очередь, с неудобным исходным векторным представлением слов, имеющих вид 'слово_частьречи'. Тэггинг (Universal POS tags), который используют создатели данной векторной модели RusVectōrēs, весьма нетривиален, и не позволяет автоматически определить все слова к нужной части речи, из-за чего некоторые слова выпадают из общего контекста. К тому же, данная модели натренирована на корпусе текстов НКРЯ (Национальный Корпус Русского Языка), который включает не только художественные произведения. FastText имеет показатели чуть хуже GloVe, однако является намного более затратным в контексте используемой памяти. Заметим, что блок GRU во всех трёх случаях показал себя хуже LSTM и потому отсутствует в итоговой таблице.

## 5.2 BERT

В данной главе для решения поставленной задачи использована архитектура нейронной сети, уже долгое время являющаяся SOTA (state of the art) в области обработки естественного языка - сеть с механизмом внимания (attention) или трансформер. Есть множество вариаций трансформеров, в данной работе при помощи библиотеки PyTorch была реализована предобученная на русском корпусе текстов модель Bert от HuggingFace, которая насчитывает свыше 180 млн. параметров. Очевидно, целиком дообучать модели на наших данных весьма проблематично с вычислительной точки зрения, к тому же это может привести к быстрому переобучению, поэтому было принято решение, тренировать лишь достроенную часть сети, являющуюся полносвязной и имеющую значительно мень-

шее количество параметров. Такой подход называется fine-tuning. Результаты данного эксперимента представлены в таблице [11].

Таблица 11: Результаты классификации на основе трансформера

| Метод | AUC ROC | Balanced Accuracy | F1 |
|---|---|---|---|
| RuBERT | 0.86769 | 0.75877 | 0.66271 |

Можно заменить, что метрики стали немного лучше чем при использовании плотных векторных представлений и рекуррентных сетей, однако всё ещё хуже итогового датасета.

## Список литературы

[1] Lagutina, K.; Lagutina, N.; Boychuk, E.; Vorontsova, I.; Shliakhtina, E.; Belyaeva, O.; Paramonov, I.; Demidov, P.G. A Survey on Stylometric Text Features. In Proceedings of the 25th Conference of Open Innovations Association (FRUCT), Helsinki, Finland, 5–8 November 2019; pp. 184–195.

[2] Batura, T.W. Formal methods of attribution of texts and their implementation in software products. Softw. Prod. Syst. 2013, 4, 286–295. (In Russian).

[3] 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 (Switzerland). - 2022. - Vol.12. - Iss. 3. - Art.1674.