Отчет по Технологии обработки больших данных.

Тема исследования: "Применение Факторного анализа для анализа текстовых произведений"

Цель исследования: Целью данного исследования является разработка и применение комбинации факторного анализа и метода опорных векторов (SVM) для атрибуции текстовых произведений на основе частоты встречаемости частей речи. Идея факторного анализа для атрибуции текстовых произведений заключается в использовании статистических методов для выявления характерных "факторов" или особенностей в текстах, которые могут служить признаками авторства. Метод опорных векторов (SVM) используется для классификации текстов на основе полученных факторов. Также проводится оценка точности предложенной методики на различных корпусах текстов.

Описание предметной области:

1. Атрибуция текстов — это область вычислительной лингвистики и обработки естественного языка (NLP), в которой используются различные методы анализа для определения автора текста, темы, стиля и других характеристик. Эта задача может быть решена с помощью различных методов, включая статистические и машинного обучения. Проблема атрибуции беспокоит многих филологов, юристов, криминалистов, историков и других специалистов уже сотни лет.

2. Частотность частей речи (POS — Parts of Speech) — это важный аспект анализа текста. Части речи включают существительные, глаголы, прилагательные, наречия и другие. Анализ частотности POS может выявить уникальные стилистические черты, присущие различным авторам или жанрам текстов. Например, авторы научных статей могут чаще использовать существительные и прилагательные, в то время как авторы художественных произведений могут использовать больше глаголов и наречий.

3. Факторный анализ — это статистический метод, который используется для описания вариативности среди наблюдаемых, коррелирующих переменных через меньшее число некоррелирующих переменных, называемых факторами. В контексте атрибуции текстов, факторный анализ может использоваться для:
- Идентификации скрытых структур в данных: Например, выявление общих стилевых характеристик среди текстов различных авторов.
- Редукции размерности: Уменьшение количества переменных (например, частот POS), сохраняя основную информацию для упрощения последующего анализа.

4. Метод опорных векторов (SVM) — это алгоритм машинного обучения, который используется для задач классификации и регрессии. В атрибуции текстов SVM применяется для:
- Классификации текстов: Разделение текстов на разные категории (например, по авторству) на основе частотности POS.
- Определения ключевых характеристик: SVM может выявить, какие POS наиболее важны для определения авторства.

5. Применение факторного анализа и SVM для атрибуции текстов (для нашей задачи).
1. Работа с корпусом текстов. Корпус должен включать произведения разных авторов.
2. Выбор характеристик. Определяются характеристики текста, на основе которых проводится анализ. В данном случае, основными характеристиками являются частотность частей речи, таких как существительные, прилагательные, глаголы и т. д.
3. Создание датасетов. Создаётся таблица, где строки представляют тексты, а столбцы - частотность различных частей речи в этих текстах.

4. Факторный анализ. Применяется для выделения скрытых факторов, объясняющих вариацию — в частотности частей речи между текстами.
5. Использование SVM на факторных данных. Полученные факторы используются в качестве признаков для обучения метода опорных векторов (SVM). SVM классифицирует тексты на основе их признаков, позволяя определить авторство каждого текста.
6. Интерпретация результатов. После классификации текстов с помощью SVM происходит интерпретация результатов. Оценивается, насколько хорошо модель смогла определить авторство текстов на основе частотности частей речи. Также проводится анализ вклада каждой части речи в различение между авторами, что может помочь в понимании стилистических особенностей каждого автора.

6. Примеры применения.
- Определение авторства спорных или анонимных текстов.
- Обнаружение плагиата путем сравнения стиля текста с текстами известных авторов.
- Разделение текстов по темам на основе стилевых различий.

Данные:

Причина: Ф.М. Достоевский публиковал собственные статьи анонимно или под другим именем, когда был редактором публицистических изданий. Требуется провести анализ произведений для присвоения авторства Ф.М. Достоевскому или другому автору.

Источник данных. ИС «СМАЛТ».

Для реализации методов были использованы тексты из информационной системы «Статистические методы анализа литературного текста» (ИС «СМАЛТ»), которая является базой данных литературных произведений с морфологическим и синтаксическим параметрами, объёмом до 300 текстов из публицистики 60—70 годов 19 века («Время», «Эпоха», «Гражданин» и др.).

ИС состоит из двух основных блоков: функционального блока, предназначенного для морфологического и синтаксического анализа текстов, пополнения БД литературных произведений, а также внесения исправлений; и аналитического блока, состоящего из модулей, реализующих разнообразные методики статистического анализа текстов.

Количество задействованных текстов Достоевского составляет 48 произведений, других авторов — 42. Использовались тексты Ф. М. Достоевского, В. П. Мещерского, А. Григорьева, Н. Н. Страхова и других авторов. Все тексты объединялись в общий корпус, на котором строились датасеты, а именно:

1. для целых текстов,
2. для текстов с шагом разбиения 1000 слов.
3. биграммы

Способ сбора данных. Данные из SMALT для формирования датасетов (Метрики Хетсо).

Со страницы «Метрики Хетсо» из ИС «СМАЛТ» выгружался нужный набор текстов в формате excel.

Индексы текстов Достоевского:[25, 32, 35, 36, 38, 40, 42, 43, 44, 45, 46, 75, 76, 77, 78, 86, 96, 97, 99, 121, 126, 130, 136, 145, 153, 154, 155, 156, 157, 163, 164, 165, 166, 167, 188, 195, 196, 200, 201, 202, 203, 204, 205, 206, 207, 209, 298, 299].

Индексы текстов других авторов: [11, 13, 24, 26, 37, 39, 89, 85, 90, 116, 117, 125, 146, 151, 152, 177, 180, 198, 199, 237, 238, 239, 245, 246, 247, 248, 286, 287, 288, 289, 290, 292, 293, 294, 295, 296, 306, 308, 309, 310, 311, 312].

Таблицы excel состоят из трех столбцов: оригинальные слова текста, слова приведенные в начальную форму, и столбец с числами. Они обозначают часть речи данного слова в диапазоне от 0 до 22.
{0: Существительное; 1: Прилагательное; 2: Числительное; 3: Местоимение; 4: Глагол; 5: Причастие; 6: Деепричастие; 7: Наречие; 8: Категория состояния; 9: Частица; 10: Предлог; 11: Союз; 12: Модальное слово; 13: Междометие; 14: Звукоподражательное слово; 15: Иностранное слово; 16: Цитата; 17: Вводное слово; 18: Старославянизм; 19: Часть фразеологизма; 20: Неязыковой символ; 21: Сокращенное слово; 22: Часть многочленного названия}
С помощью алгоритма на python получаем обработанные файлы с текстами и строим из них наши датасеты. На вход алгоритму подается массив цифр. На выход (в зависимости от задачи) выводится массив чисел — количество частей речи.

- В датасете целых текстов — это количество каждой части речи в тексте.
- В датасете с шагом разбиения 1000 — количество каждой части речи в каждом диапазоне.
- В датасете биграмм — количество каждой пары частей речи.

Датасеты.
1) В датасете целых текстов: 90 строк и 24 столбца данных, название столбцов — части речи, первый столбец — индексы текстов.
Максимальное значение - 5453, минимальное - 0.
Количество слов в датасете — 360857.
Количество числовых ячеек — 2070.

2) В датасете с шагом разбиения 1000 без перекрестного: 406 строк и 24 столбца данных, название столбцов — части речи, первый столбец — индексы текстов с номерами диапазонов.
Максимальное значение - 985, минимальное - 0.
Количество числовых ячеек — 9315.

3) В датасете биграмм: 89 строк и 355 столбцов данных, название столбцов — части речи, первый столбец — индексы текстов.
Максимальное значение - 14864878, минимальное - 0.
Количество числовых ячеек — 31240.
Из-за большого количества полученных столбцов и для решения проблемы с нулевыми значениями было принято решение исключить из датасета части речи под индексами 15, 19, 20, 23 и тексты под индексами 45, 46.

Актуальность данных. Время и срок актуальности.
Литературные тексты остаются актуальными длительное время. Обновление может проводиться редко, раз в несколько лет.

Выходной результат:

Ожидания.
1. Создание модели машинного обучения, способной точно определять авторство текстов на основе анализа частотности частей речи. Модель должна демонстрировать высокую точность и устойчивость при работе.
2. Подробный отчет, содержащий анализ частотности различных частей речи, который выявляет ключевые лингвистические признаки, характерные для каждого автора.
3. Результаты факторного анализа, показывающие, какие скрытые факторы наиболее значимо влияют на различия в стилях авторов.

4. Метрики оценки модели, такие как точность (accuracy) и т.д. Сравнение производительности модели SVM с другими методами классификации.
5. Разработка и предоставление скриптов, которые могут быть использованы для автоматической атрибуции текстов.
6. Проведение обширного тестирования модели на различных корпусах текстов для оценки ее универсальности и надежности.

Методы:

Основные методы атрибуции текстов.

1. Лингвистический анализ:
- Частотный анализ частей речи: Изучение распределения частей речи (существительных, глаголов, прилагательных и т.д.) для выявления характерных стилевых особенностей автора.
- Анализ синтаксических структур: Изучение синтаксических паттернов, таких как структура предложений, использование определенных грамматических конструкций.

2. Стилистический анализ:
- Изучение стиля: Анализ литературных произведений для выявления уникальных стилевых характеристик автора.
- Фонетические и фонологические признаки: Изучение звуковых паттернов и ритмических характеристик текста.

3. Корпусная лингвистика:
- Анализ корпусов текстов: Использование больших коллекций текстов для статистического анализа языковых паттернов.
- Аннотированные корпуса: Корпусы, где тексты размечены с указанием частей речи, синтаксических структур и других лингвистических признаков.

4. Статистические методы:
- Методы n-грамм: Анализ последовательностей из n слов или символов для выявления уникальных последовательностей, характерных для определенного автора.
- Коэффициенты сходства: Метрики, такие как косинусное сходство, коэффициент Жаккара, для сравнения текстов.

5. Обработка естественного языка (NLP):
- Токенизация и нормализация: Процессы разделения текста на отдельные токены (слова, символы) и приведение их к базовой форме.
- Теггинг частей речи: Автоматическое присвоение частям речи меток, что помогает в анализе синтаксических структур.
- Алгоритмы сравнения текстов: Включают методы на основе шинглов, хэш-функций и векторных представлений.

6. Машинное обучение:
- Наивный байесовский классификатор: Использует теорему Байеса для классификации текстов на основе вероятностных моделей.
- Метод опорных векторов (SVM): Находит гиперплоскость, оптимально разделяющую классы текстов в высокоразмерном пространстве признаков.
- Деревья решений и случайные леса: Деревья решений строят модели на основе деревьев, а случайные леса используют множество деревьев для улучшения точности и устойчивости.
- Нейронные сети: Включают многослойные перцептроны, рекуррентные нейронные сети (RNN) и трансформеры для более сложного анализа текстов.

Анализ:

Факторный анализ.
1)В датасете целых текстов:
Подготовка к ФА:
На вход: массив чисел, где первый столбец — индексы текстов, первая строка — названия столбцов.
На выход:

- Тест Бартлетта проверяет, коррелируют ли вообще наблюдаемые переменные, используя наблюдаемую корреляционную матрицу против единичной матрицы. Если тест оказался статистически незначимым, то не следует использовать факторный анализ.
  p-value: (3114.5306431112294, 0.0)
  В этом тесте Бартлетта p-value равно 0. Тест статистически значимый, что указывает на то, что наблюдаемая корреляционная матрица не является единичной матрицей.
- Тест Кайзера-Мейера-Олкина (KMO) измеряет пригодность данных для факторного анализа. Он определяет адекватность каждой наблюдаемой переменной и всей модели. KMO оценивает долю дисперсии среди всех наблюдаемых переменных. Значения KMO находятся в диапазоне от 0 до 1. Значение KMO менее 0,6 считается недостаточным.
  Общий KMO для наших данных составляет **0.909406.**
-

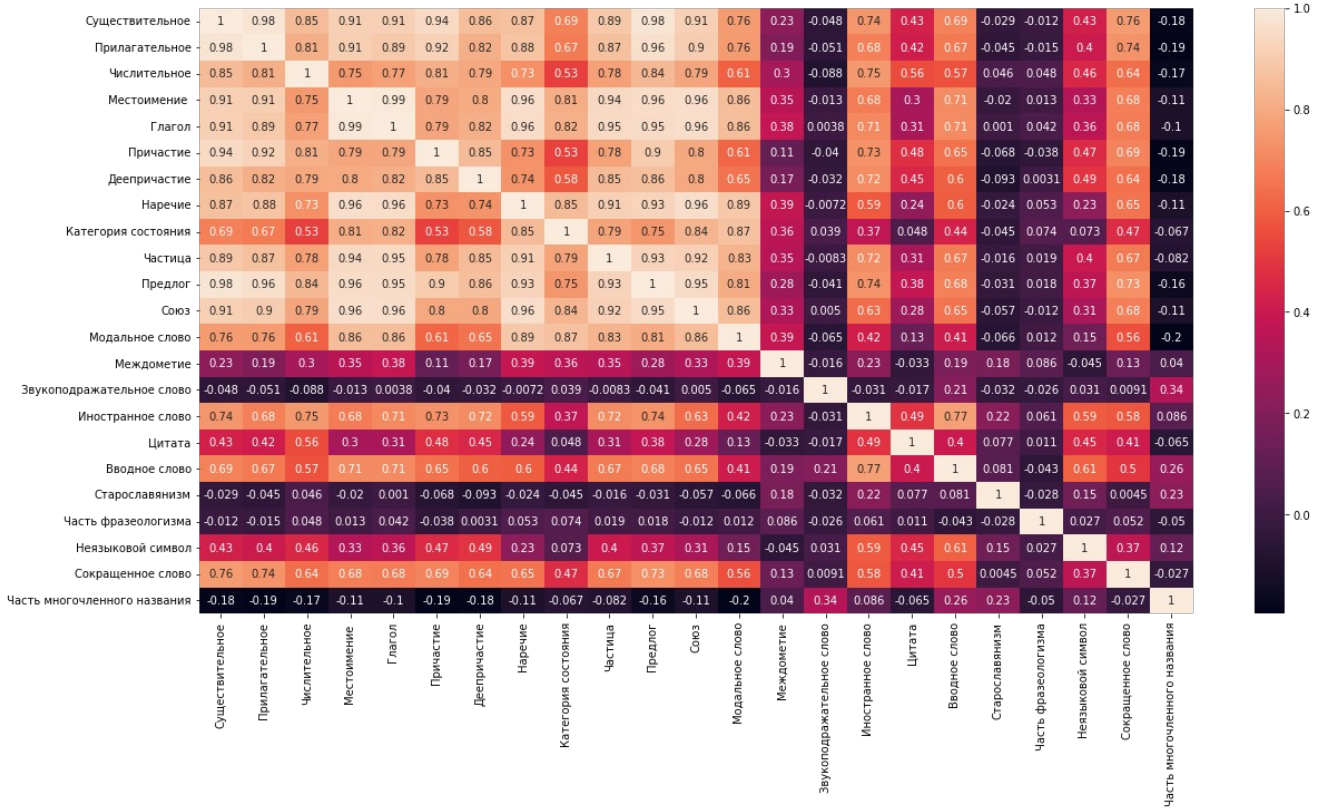


Рис. 1. Корреляционная матрица датасета целых текстов.

Матрица мультиколлинеарна - сильная положительная корреляция у первых 13 переменных. У Звукоподражательного слова, старославянизма, части фразеологизма и части многочленного названия близкие к нулю значения корреляции - между переменными нет линейной зависимости.

- Выбор количества факторов. Критерий Кайзера дал результат из 5 собственных значений, которые больше единицы. Но сделав ФА для 5 факторов обнаружилось, что 5й фактор имеет низкие факторные нагрузки и поэтому в дальнейшем будем реализовывать для 4 факторов, что подтверждается наибольшим значением объясненной дисперсии.

ФА:

На вход: массив чисел, где первый столбец — индексы текстов, первая строка — названия столбцов.

На выход:

1. Уберем доли дисперсии переменных с низким показателем(<0.3):
   Звукоподражательное слово 0.232874; Старославянизм 0.259192; Часть фразеологизма 0.017796.
   Исключим эти переменные и построим факторный анализ для 4 факторов.
2. 4 фактора объясняют 74.18% дисперсии

SS Loadings	9.528860	3.811341	0.924425	0.574714
Proportion Var	0.476443	0.190567	0.046221	0.028736
Cum Var	0.476443	0.667010	0.713231	0.741967

Таблица. 1. Дисперсия каждого фактора для уменьшенного целого датасета.

- SS Loadings показывает, какую часть общей дисперсии в данных объясняет каждый фактор. Большее значение SS Loadings означает, что фактор объясняет большую часть дисперсии исходных переменных.
- Proportion Var представляет собой долю дисперсии, объясненную каждым фактором. Значение Proportion Var близкое к 1 указывает на то, что фактор объясняет большую часть дисперсии.
- Cumulative Var представляет собой накопленную дисперсию, объясненную каждым фактором в совокупности с предыдущими факторами. Он показывает, какую часть общей дисперсии в данных объясняют первые N факторов. Нужен для определения того, сколько факторов включить в анализ для достижения определенной доли объясненной дисперсии.

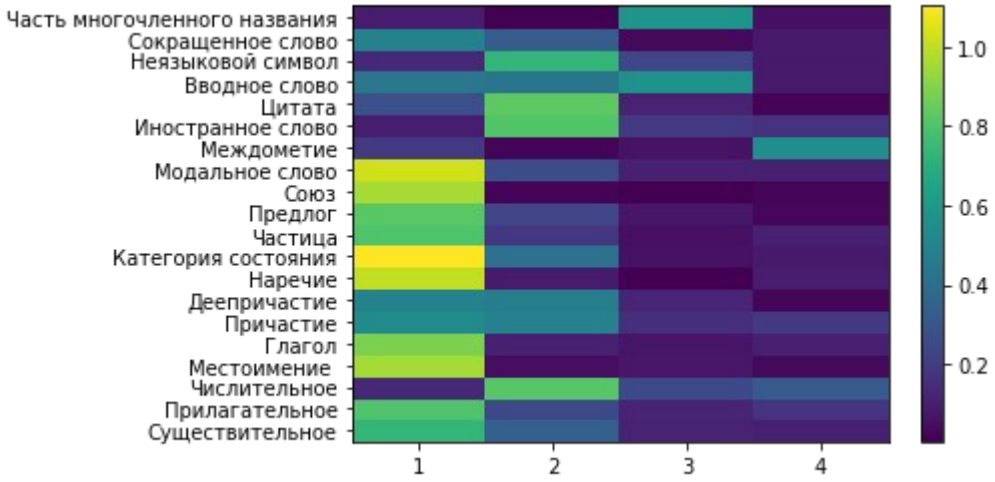
3. 

Рис. 2. Тепловая карта для 4 факторов уменьшенного целого датасета.

- Фактор 1 имеет высокие факторные нагрузки для категория состояния, модальное слово, союз, наречие, местоимение.
- Фактор 2 имеет высокие факторные нагрузки для числительное, иностранное слово, цитата, неязыковой символ.
- Фактор 3 имеет высокие факторные нагрузки для вводное слов, часть многочленного названия.
- Фактор 4 имеет высокие факторные нагрузки для междометие.

SVM:

В работе применяется линейная классификация SVM с мягким зазором и гиперпараметр C=0.01 для регулирования классификации. Разделение данных на обучающий и тестовый наборы 0.7 и 0.3 соответственно.
На вход: массив с 4 факторными нагрузками.
На выход: Точность модели с линейным ядром равна **0.875**.

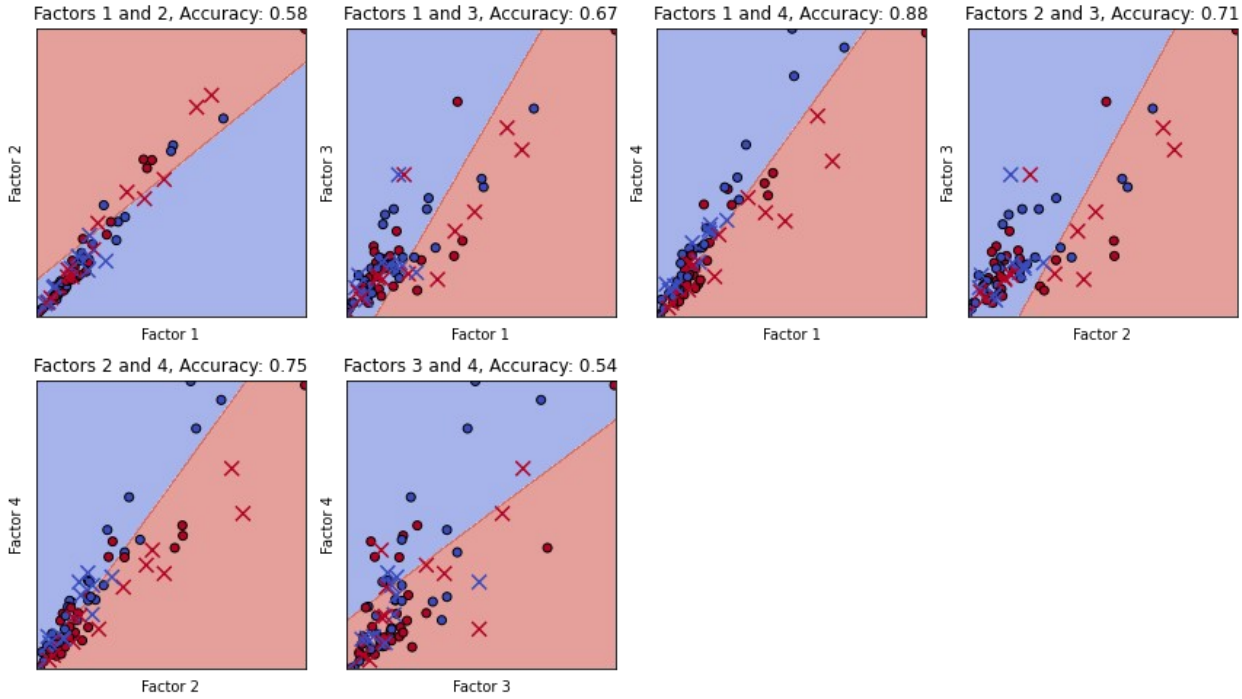


Рис. 3. График SVM для 4 факторов целого датасета с точность для каждой пары факторов(обучающие данные отображены точками, а тестовые — x).

Наилучший результат показывает пара факторов 1 и 4, их точность составляет **0.88.**
При реализации пары факторов 1 и 4 участвуют произведения Достоевского под номерами [32, 35, 44, 46, 76, 86, 96, 97, 99, 130] и других авторов [24, 39, 89, 85, 90, 116, 125, 151, 177, 180, 247]

2)В датасете с шагом разбиения 1000 без перекрестного:
Подготовка к ФА:
На вход: массив чисел, где первый столбец — индексы текстов, первая строка — названия столбцов.
На выход:

- Тест Бартлетта **p-value: (4731.873815, 0.0)**
- Тест Кайзера-Мейера-Олкина (KMO): **0.883.**

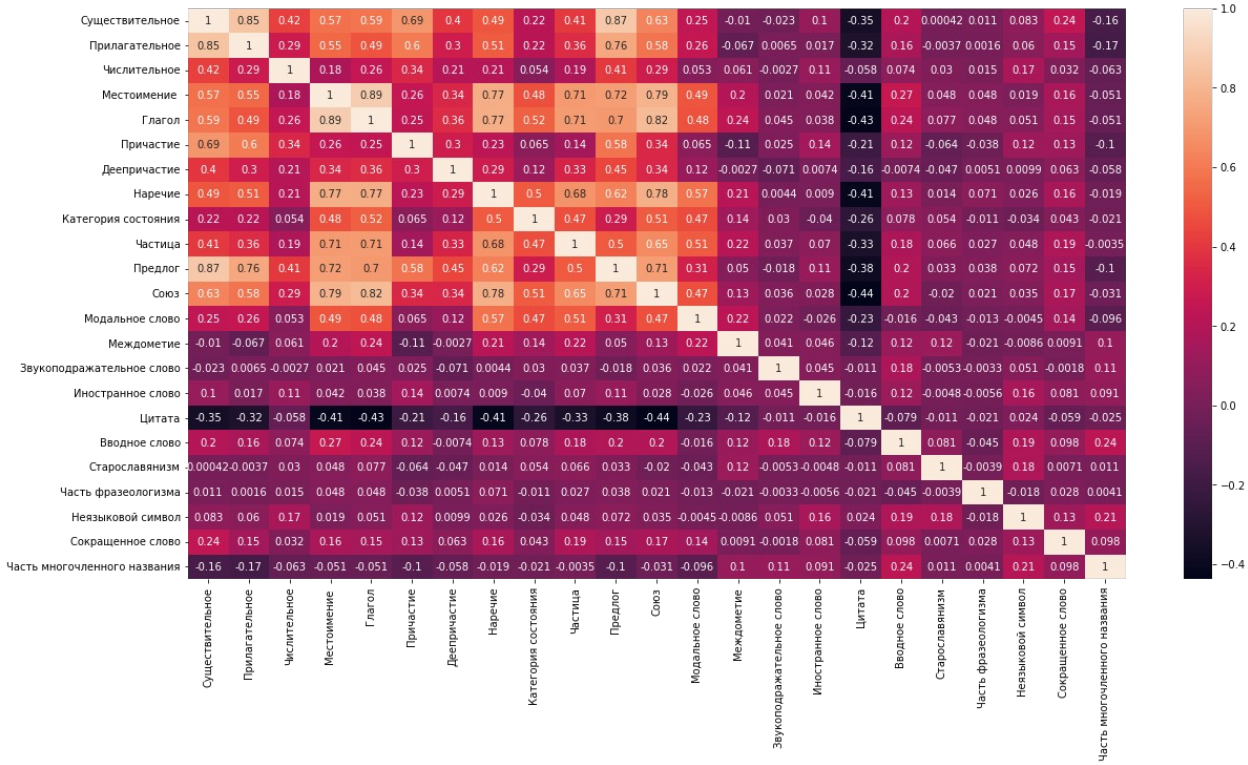
- 

Рис. 4. Корреляционная матрица.

Положительная корреляция. У части переменных близкое к нулю значение корреляции.

- Критерий Кайзера дал результат из 6 собственных значений, которые больше единицы. В дальнейшем мы продолжим использовать 6 факторов, так как при проведении ФА с другим количеством факторов показатель объясненной дисперсии был ниже.

ФА:

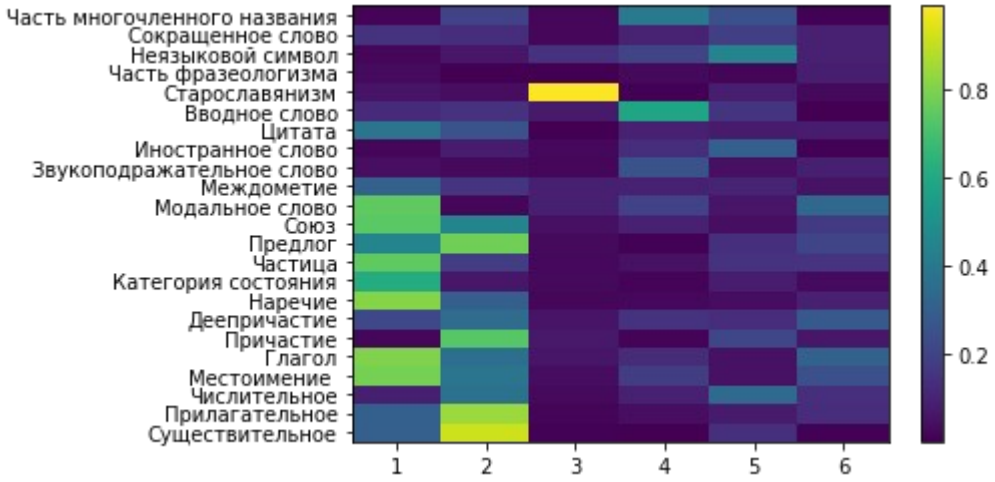
На вход: массив чисел, где первый столбец — индексы текстов, первая строка — названия столбцов.

На выход:

1. 6 факторов объясняют 49.66% дисперсии

SS Loadings	4.6851	3.7147	1.0417	0.7812	0.6879	0.5154
Proportion Var	0.2037	0.1615	0.0452	0.0339	0.0299	0.0224
Cumulative Var	0.2037	0.3652	0.4105	0.4444	0.4743	0.4967

Таблица. 2. Дисперсия каждого фактора для датасета.



2.

Рис. 5. Тепловая карта для 6 факторов общего датасета.

- Фактор 1 имеет высокие факторные нагрузки для местоимения, глагола, наречия, частицы, союза, модального слова.
- Фактор 2 имеет высокие факторные нагрузки для существительного, прилагательного, причастия, предлога.
- Фактор 3 имеет высокие факторные нагрузки для старославянизма.
- Факторы 4 имеет высокие факторные нагрузки для вводного слова.
- Факторы 5, 6 не имеют высоких нагрузок ни для одной переменной и их нелегко интерпретировать.

SVM:

В работе применяется линейная классификация SVM с мягким зазором и гиперпараметр C=0.01 для регулирования классификации. Разделение данных на обучающий и тестовый наборы 0.7 и 0.3 соответственно.

На вход: массив с 6 факторными нагрузками.

На выход: Точность модели с линейным ядром равна **0.8641**.

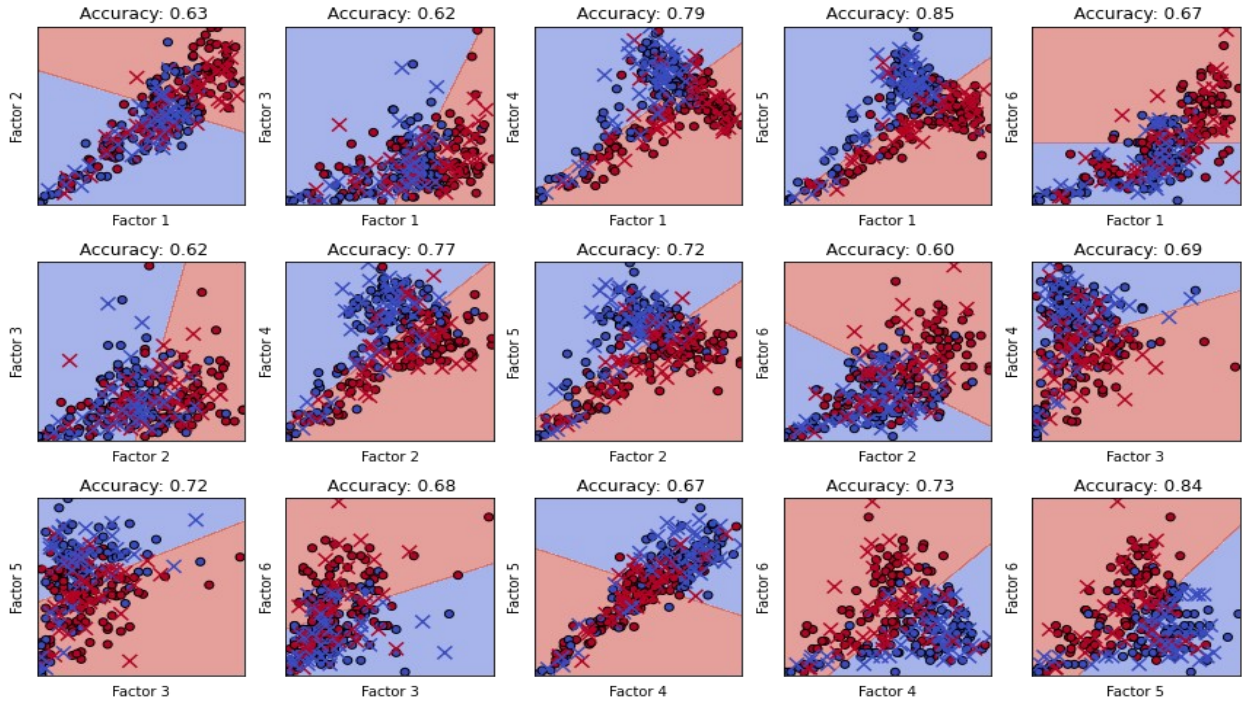


Рис. 6. График SVM для 6 факторов датасета с точность для каждой пары факторов(обучающие данные отображены точками, а тестовые — x).

Наилучший результат показывает пара факторов 1 и 5, их точность составляет **0.85.**
При реализации пары факторов 1 и 5 участвуют произведения Достоевского под номерами и индексом фрагмента [25,35_1, 38_2, 38_3, 40_1, 40_2, 42_3, 42_8, 42_9, 42_10, 42_12, 42_13, 43_1, 43_3, 44_1, 75_4, 75_5, 75_9, 75_10, 75_11, 76_2,76_5, 77_2, 77_5, 77_6, 77_9, 77_12, 77_15, 78_1, 78_3, 86_1, 96_1, 96_2, 97_1, 99_1, 99_2] и других авторов [177_4, 177_6, 177_7, 177_8, 198_1, 198_2, 198_4, 198_5, 198_6, 198_7, 198_11, 237_2, 238_1, 238_2, 239_1, 239_2, 239_3, 245_1, 245_7, 245_8, 245_12, 245_15, 245_16, 245_18, 245_19,246_1, 246_2, 246_5, 247_2, 248_2, 286_2, 288, 292_1]

2)В датасете биграмм:
Построим массив данных из датасета выбрав первые 25 дисперсий в порядке убывания.
Дисперсия покрывают 83.97% от всех данных датасета.
Названия столбцов для первых 25 дисперсий:
['17_17', '1_1', '17_1', '1_17', '4_1', '1_4', '1_5', '5_1', '17_4', '11_1', '4_17', '1_11', '1_2', '1_12', '12_1', '2_1', '17_5', '17_2', '4_4', '17_11', '5_17', '11_17', '2_17', '1_10', '17_12']

Подготовка к ФА:
На вход: массив чисел, где первый столбец — индексы текстов, первая строка — названия столбцов.
На выход:

- Тест Бартлетта **p-value: (11540.577379522332, 0.0)**
- Тест Кайзера-Мейера-Олкина (KMO): **0.7780.**

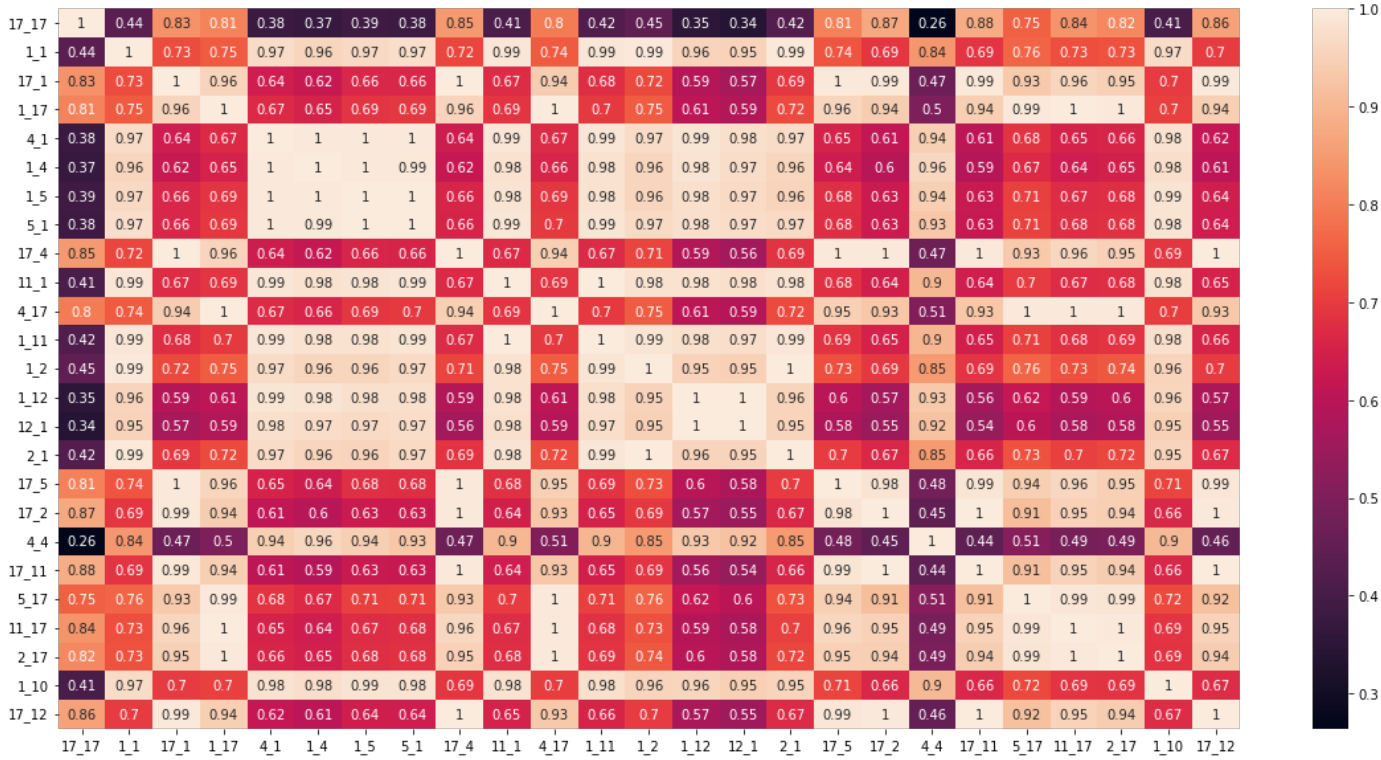


Рис. 7. Корреляционная матрица.

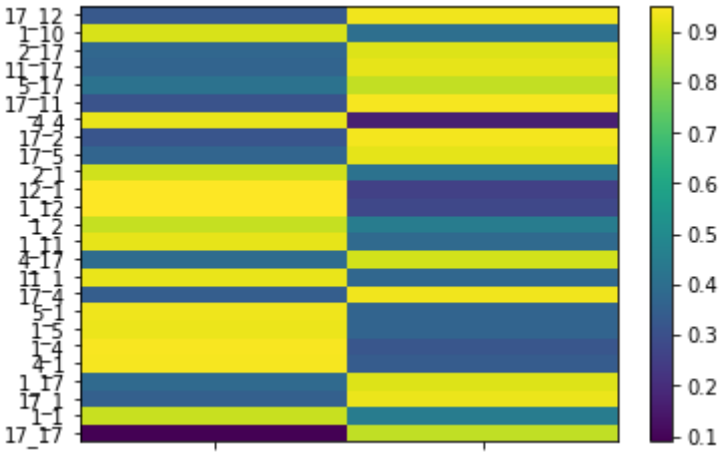
  Сильная положительная корреляция.
- Критерий Кайзера дал результат из 2 собственных значений, которые больше единицы.

ФА:
На вход: массив чисел, где первый столбец — индексы текстов, первая строка — названия столбцов.
На выход:

1. 2 фактора объясняют 96.23% дисперсии.

SS Loadings	12.404289	11.655821
Proportion Var	0.496172	0.466233
Cumulative Var	0.496172	0.962404

Таблица. 3. Дисперсия каждого фактора для датасета.



2.

Рис. 8. Тепловая карта для 2 факторов общего датасета.

- Фактор 1 имеет высокие факторные нагрузки для 1_10, 4_4, 2_1, 12_1, 1_12, 1_2, 1_11, 11_1, 5_1, 1_5, 1_4, 4_1, 1_1.
- Фактор 2 имеет высокие факторные нагрузки для '17_17', '17_1', '1_17', '17_4', '4_17', '17_5', '17_2', '17_11', '5_17', '11_17', '2_17', '17_12'.

SVM:

В работе применяется линейная классификация SVM с мягким зазором и гиперпараметр C=0.01 для регулирования классификации. Разделение данных на обучающий и тестовый наборы 0.7 и 0.3 соответственно.

На вход: массив с 2 факторными нагрузками.

На выход: Точность модели с линейным ядром равна **0.75**

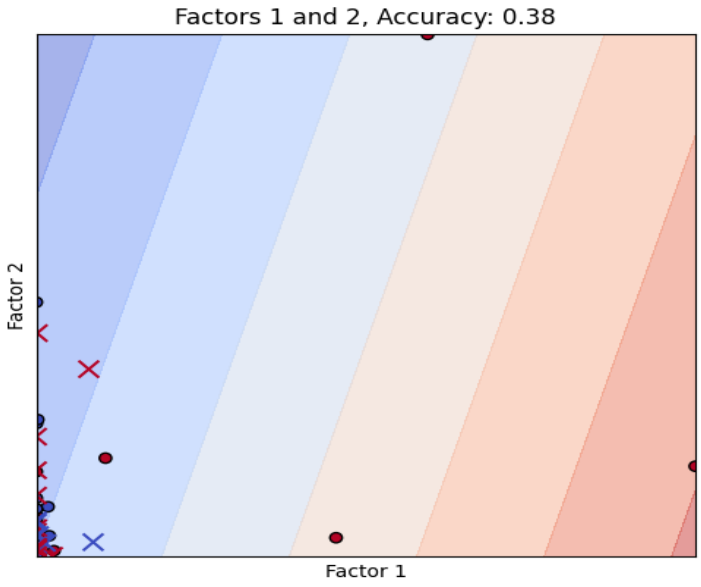


Рис. 9. График SVM для 2 факторов датасета с точность для каждой пары факторов(обучающие данные отображены точками, а тестовые — x).

Пара факторов 1 и 2 показывает точность **0.38.**
При реализации пары факторов 1 и 2 участвуют произведения Достоевского под номерами и индексом фрагмента [32, 76, 97, 99, 126, 130] и других авторов [177, 180].

Поскольку SVM не дает высокой точности классификации, то построим точечную диаграмму 2-х факторов для подбора другого классификатора.

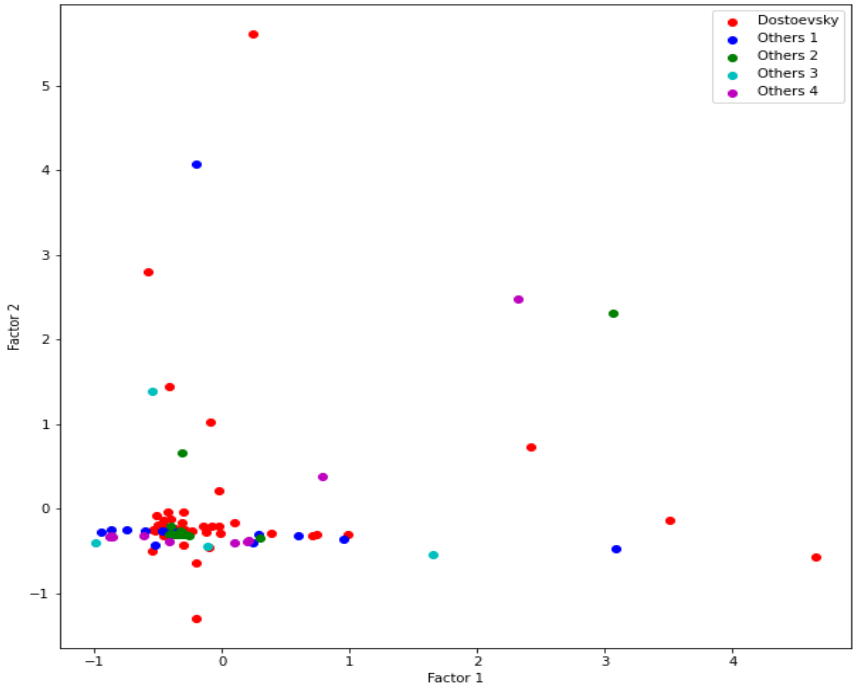


Рис. 10. Точечная диаграмма для 2 факторов датасета биграмм с вращением.

Описание применимости и выводы:

Применимость.

- Анализ авторства: Определение авторства анонимных или спорных литературных произведений.
- Сравнительный анализ стилей: Сравнение стилистических особенностей различных авторов или различных периодов творчества одного автора.
-Изучение эволюции стиля: Исследование изменений стиля писателя на протяжении его карьеры.

Выводы.

Общие показатели факторного анализа значительно ниже у датасета с текстами разбитыми на части, чем с целыми. Полученные данные факторного анализа и SVM на наших двух датасетах не могут быть использованы для атрибуции, поэтому выдвинуто предположение о создании других наборов признаков для апробации факторного анализа в рамках атрибуции тестовых произведений.

1. Целые тексты авторов:
    1. В общем датасете наилучшие показатели в факторном анализе у факторов 1 и 4. Объясняется 74.18% дисперсии при условии, что ФА строится для 4 факторов и датасет уменьшен.
    2. Фактор 1 имеет высокие факторные нагрузки для модального слова, союза, категории состояния, наречия, местоимения.
       Фактор 4 имеет высокие факторные нагрузки для междометия.
    3. SVM при этих двух факторов дает лучший результат с точностью 0.88

2. Тексты с шагом разбиения 1000:
    1. Общий датасет дает лучший показатель дисперсии при 6 факторах— 49.66% и самая высокая факторная нагрузка у 3го фактора(старославянизм). Но общие показатели ФА значительно ниже у датасета с текстами разбитыми на части, чем с целыми.
    2. SVM для 4 факторов дает максимальный результат с точностью 0.79 при 1 и 4. Для 6 факторов наилучший результат 0.85 при 1 и 5.

Данные полученные на датасете биграмм после факторного анализа классифицированы SVM с низкой точность, поэтому есть предложение построить точечную диаграмму 2 факторов и подобрать другой классификатор.

3. Биграммы
    1. Есть явное противопоставление параметров 17 и 1 (1 — существительное, 17 — старославянизм).
    2. Пара факторов 1 и 2 показывает точность 0.38 в SVM.

Возможные пути улучшения:

1. Расширение корпуса данных: Увеличение объема, разнообразия текстовых данных, разнообразное построение датасетов(перекрестный шаг разбиения, другие диапазоны шагов разбиения, три-граммы, создание «общего» текста, и т. д).
2. Использование других методов машинного обучения: Исследование других методов классификации, таких как нейронные сети или ансамблевые методы.
3. Дополнительные признаки: Включение в анализ других текстовых признаков, таких как длины предложений, структура абзацев, синтаксические структуры, лексическое разнообразие и т.д.