Богданова Анастасия

Лабораторная работа по машинному обучению №2

Вариант 1. Реферат статьи + мини-проект

Статья: Ольга Алиева. Меры расстояния для определения авторства древнегреческих текстов // Цифровые гуманитарные исследования. 2024. Т. 1. С. 8–33. — DOI: 10.31860/cgi-2024-1-8-33.

Статья *О. В. Алиевой* представляет собой экспериментальное сравнение *distance-based* методов атрибуции авторства на материале древнегреческой прозы. В работе рассматриваются методы, основанные на векторизации текстов через частотности токенов с последующим измерением расстояний между векторами.

Большинство предыдущих работ по *distance-based* стилометрии сосредоточено на исследовании текстов на современных европейских языках. Статья О.В. Алиевой демонстрирует, что методы работают и на корпусе древнегреческой прозы, что является новым для научной области.

В исследовании ставятся следующие задачи:
— определить какие меры расстояния демонстрируют наибольшую точность на отрывках разной длины при варьировании числа переменных;
— выявить различия в результатах при использовании стандартизированных и нестандартизированных значений частотности;
— сравнить точность атрибуции при работе со словоформами и трехсимвольными энграммами
— проанализировать тексты, на которых классификаторы чаще всего допускают ошибки.

В качестве материала исследования использовался корпус древнегреческой прозы объемом около 694 тыс. слов из библиотеки Perseus, включающий 57 текстов 17 авторов. Корпус является несбалансированным по авторам и объему текстов. Предварительная лингвистическая разметка (лемматизация, стемминг, морфологический или синтаксический анализ) не применялась. Тексты подвергались только токенизации и представлены в двух вариантах: словоформы и трехсимвольные энграммы.

Для сравнения были отобраны меры расстояния и сходства, ранее показавшие высокую эффективность в стилометрических исследованиях: манхэттенское и евклидово расстояния; косинусное сходство (в том числе со стандартизацией признаков по z-оценке); сходство Ружечки (minmax), эквивалентное расстоянию Танимото; канберрское расстояние; расстояние Кларка; расхождение Джеффриса и расстояние Лаббе. Для большинства методов варьировалось число наиболее частотных слов (mfw), за исключением метода Лаббе, не предполагающего отбора признаков.

Оценивание проводилось на отрывках длиной от 1000 до 7000 токенов (с шагом 500) при числе предикторов от 100 до 1000 (с шагом 100). Для каждой комбинации параметров выполнялось 10 итераций с повторной выборкой (bootstrap), что обусловлено наличием в корпусе коротких текстов. В результате для каждого метода было выполнено более 74 тысяч классификаций без стандартизации и столько же со стандартизацией, отдельно для словоформ и трехсимвольных энграмм. Точность определялась как доля корректных атрибуций от общего числа классификаций. Дополнительно анализировались методы на малых выборках с подбором оптимальных параметров для фрагментов длиной 1000–2000 токенов.

Все вычисления и визуализации были реализованы в среде R с использованием специализированных библиотек для стилометрии (Stylo) и анализа расстояний (Philentropy).

Сравнительный анализ показал, что на всех исследованных длинах отрывков и при различном числе предикторов наилучшие результаты демонстрируют расстояние Лаббе, косинусное сходство и сходство Танимото (minmax). Несколько менее эффективными оказались расхождение Джеффриса и манхэттенское расстояние. Классификация на основе трехсимвольных энграмм стабильно уступает анализу словоформ. Стандартизация признаков оказывает положительное влияние исключительно на косинусное сходство.

Анализ зависимости точности классификации от числа наиболее частотных слов (mfw) показывает, что для большинства методов увеличение числа предикторов либо приводит к росту точности, либо оказывает на нее несущественное влияние. Наиболее устойчивые результаты достигаются при использовании mfw свыше 200.

Расстояние Лаббе проявляет наибольшую эффективность на более протяженных фрагментах (от 3000 токенов), тогда как на коротких отрывках предпочтительным оказывается сходство Танимото. Эксперименты на малых выборках (1000-2000 токенов) выявили снижение точности для всех методов. В среднем по всем экспериментам наилучшие показатели демонстрирует стандартизированное косинусное сходство, которое доминирует среди лучших результатов, особенно на средних и больших отрывках. Вместе с тем автор подчеркивает, что высокие показатели точности не должны интерпретироваться как основание для безусловного подтверждения авторства: количественные методы могут служить инструментом критической проверки традиции, но не ее окончательной верификации.

Статья демонстрирует конкретные показатели точности, дает рекомендации по выбору метрик и числа признаков, предлагает эмпирическую проверку спорных текстов (Исократ, Плутарх, Платон). Анализ матриц ошибок показал, что ни один метод не обеспечивает стопроцентной точности; характер ошибок различается в зависимости от метрики и, вероятно, связан с жанровой близостью текстов и редакторскими вмешательствами, что требует дополнительного исследования. Полученные результаты интерпретируются с учетом как статистической значимости, так и ограничений стилометрического подхода применительно к древнегреческим текстам.

Эксперимент, проведенный автором статьи, отличается высокой степенью воспроизводимости и масштабностью перебора параметров, а использование bootstrap-процедур позволяет частично компенсировать несбалансированность корпуса, однако результаты трудно считать полностью универсальными. В широком контексте машинного обучения статья не создает новый алгоритм, а лишь последовательно применяет существующие методы. Она дает методологические рекомендации, демонстрирующие преимущества разных метрик и условия их эффективности, но универсальные закономерности для других языков или жанров остаются открытыми.

Статья убедительно демонстрирует, что количественные методы эффективны как инструмент отрицательной атрибуции, однако вопрос о возможности статистического подтверждения авторства остается открытым и требует интеграции количественных подходов с традиционными филологическими методами.

Мини-проект.

Воспроизведение эксперимента О.В. Алиевой на материалах корпуса, собираемого для магистерской диссертации «Атрибуция текстов русской нелегальной печати на примере газеты «Искра» (1900–1906 гг)»

Цель эксперимента: оценить устойчивость distance-based методов атрибуции авторства и их метрик при изменении языка, жанра и длины текстов. Дополнительно сравнить результаты с базовыми методами машинного обучения, сопоставить результаты. Выполнить эксперименты в Python вместо R.

Материал исследования: мини-корпус русской революционной публицистики начала XX века, включающий 84 текста пяти авторов (Ленин, Плеханов, Мартов, Парвус, Троцкий) + 7 dubia-текстов из газеты «Искра». Корпус характеризуется жанровой однородностью, ограниченным числом тем, возможными редакторскими правками и небольшой длиной текстов. Представленность авторов и длина текстов несбалансированы. Предварительная лингвистическая разметка не применялась, тексты подвергались только базовой очистке и токенизации.

Методы:
1. Distance-based подходы: использовались косинусное сходство и расстояние Танимото (minmax), применялись TF-IDF матрицы с различными числами признаков (300, 500, 1000). Длинные тексты разрезались на chunks по 1000 слов. Оценка результатов: анализ accuracy на полном корпусе, анализ предсказанных авторов для dubia-текстов, проверка устойчивости решений при изменении числа признаков, сравнение поведения метрик, анализ margin как характеристики уверенности самой метрики (margin — разница между первой и второй максимальной вероятностью).
2. Машинное обучение: использовались базовые модели классификации Naive Bayes, Logistic Regression, обученные на тех же TF-IDF признаках, что и в

distance-bace экспериментах. Для оценки качества моделей применялась кросс-валидация с сохранением пропорций авторов (StratifiedKFold) и рассчитывались метрики precision, recall, f1-score. Для dubia-текстов вычислялась вероятность принадлежности к каждому автору, оценивалась уверенность модели через margin.

В целом результаты указывают на то, что distance-based подходы сохраняют работоспособность на новом материале, однако их выводы существенно зависят от структуры корпуса. В условиях малого и тематически однородного корпуса метрики склонны отражать сходство жанра и редакторской правки, а не устойчивые авторские особенности. Машинное обучение показало более категоричные предсказания для авторов с большим объемом текстов в корпусе. Авторы, слабо представленные в корпусе моделями выделяются слабо. Logistic Regression в целом демонстрирует более устойчивые результаты и большую согласованность между предсказанными и фактическими авторами.

Общие выводы: distance-based методы и ML-модели дают взаимодополняющую информацию. Для небольшого и несбалансированного корпуса оба подхода ограничены, особенно для авторов с малым числом текстов. Margin и показатели уверенности помогают интерпретировать надежность предсказаний, но не заменяют экспертной проверки. Оба подхода дают основу для дальнейшей атрибуции dubia-текстов и формируют методологию для расширения корпуса текстов.