СРАВНИТЕЛЬНЫЙ АНАЛИЗ СТИЛОМЕТРИЧЕСКИХ СИСТЕМ

Представлены три итерации одного исследования, посвященного ***диахроническому анализу авторского стиля*** А. С. Суворина на материале его писем за 1904-1908 гг. Основной исследовательский вопрос: ***фиксируются ли измеримые изменения в идиостиле автора с течением времени***? Идентифицированы три различные системы, каждая из которых отражает определенный этап в развитии методов компьютерной лингвистики.

СИСТЕМА 1: КЛАССИЧЕСКАЯ СТИЛОМЕТРИЯ И МЕТОД DELTA

Это отправная точка исследования, использующая канонические подходы к стилометрии.

Общее описание:

Система реализована в виде Jupyter Notebook. Цель – применить стандартные стилометрические методики, в первую очередь, основанные на частотности лексических единиц, для кластеризации текстов по годам.

Методика исследования:

- **Сбор данных:** тексты писем А.С. Суворина за 1904-1908 гг. были собраны из печатного источника *[Суворин А. С. Русско-японская война и русская революция. Маленькие письма (1904–1908). М.: Алгоритм, 2005. 720 с.]* и оцифрованы.
- **Предобработка:** включает в себя токенизацию, лемматизацию и удаление стоп-слов.
- **Извлечение признаков:** основным признаком является вектор **частотностей наиболее употребимых слов**.
- **Анализ:** ключевым методом является **Burrows' Delta** [1] – мера, вычисляющая расстояние между векторами частотностей слов двух текстов. На основе этих расстояний строится дендрограмма (иерархическая кластеризация), которая визуально показывает, какие тексты (годы) ближе друг к другу по стилю.

Оценка подхода:

- **Преимущества:** высокая интерпретируемость, простота реализации, воспроизводимость. Является «золотым стандартом» в классической атрибуции.
- **Недостатки:** поверхностный анализ (только лексика), высокая чувствительность к теме, игнорирование порядка слов (подход «мешок слов»).

СИСТЕМА 2: ГИБРИДНЫЙ ПОДХОД С WORD2VEC + GRAPH И АНСАМБЛЕВЫМ МАШИННЫМ ОБУЧЕНИЕМ

Эта система представляет собой шаг вперед, переходя от простого подсчета частот к семантическому анализу и машинному обучению.

Общее описание:

Система нацелена на построение полноценного классификатора. Вместо одного типа признаков используется множество (лексические, синтаксические, семантические, графовые), которые подаются на вход ансамблю классификаторов.

Методика исследования:

- **Семантическое моделирование:** обучение модели **Word2Vec** [2] для получения векторных представлений (эмбеддингов) слов.
- **Инжиниринг признаков:** Создание комплексного набора признаков: *лексическое разнообразие, частотность частей речи, усредненные векторы текстов на основе Word2Vec, графовые метрики семантической когерентности* [3], где узлы – слова, а ребра – их семантическая близость.
- **Машинное обучение:** использование ансамблевых методов (RandomForestClassifier, GradientBoostingClassifier) *[RandomForest, GradientBoosting,* ExtraTrees (72,7%), VotingClassifier, StackingClassifier, LogisticRegression (50,7%)] и их объединение в VotingClassifier (71%), который объединяет предсказания нескольких моделей для повышения итогового качества.
- **Валидация:** применяется стратифицированная кросс-валидация (StratifiedKFold) для получения надежной оценки точности, а также статистические тесты для оценки значимости различий. Применяется стратифицированная кросс-валидация (StratifiedKFold) *[Bootstrap-анализ для доверительных интервалов. Статистические тесты (t-test, Mann-Whitney, KS-test) и расчет размеров эффектов (Cohen's d)]* для получения надежной оценки точности модели.

Оценка подхода:

- **Преимущества:** глубина анализа (учет семантики и синтаксиса), мощность ансамблевых методов, гибкость в добавлении новых признаков.
- **Недостатки:** сложность и трудоемкость ручного инжиниринга признаков, низкая интерпретируемость («черный ящик»).

СИСТЕМА 3: ГИБРИДНАЯ МОДЕЛЬ НА ОСНОВЕ BERT И ГРАДИЕНТНОГО БУСТИНГА (ВЕРСИЯ V11)

Это финальная система, которая достигла пиковой точности **87.0%**. Она основана на современной архитектуре, сочетающей трансформерные эмбеддинги и классическое машинное обучения.

Общее описание:

Ключевой особенностью успешной архитектуры стал отказ от нестабильного дообучения (fine-tuning) нейронных сетей в пользу более надежного подхода, основанного на извлечении признаков и использовании алгоритма градиентного бустинга **LightGBM** [4].

Методика исследования:

1. **Постановка задачи: бинарная классификация.** Первоначальная задача по разделению текстов на 4 класса (по годам) была упрощена до бинарной: **«ранний период» (1904-1905)** против **«позднего периода» (1906-1907)**. Это позволило модели сфокусироваться на более выраженных стилистических изменениях.
2. **Извлечение семантических признаков:** используется предобученная модель **DeepPavlov/rubert-base-cased** [5] в режиме **экстрактора признаков**. Для получения векторного представления текста применяется стратегия **Mean Pooling** (усреднение эмбеддингов всех токенов с учетом маски внимания), что дает более богатое семантическое представление по сравнению со стандартным [CLS] токеном.
3. **Извлечение стилометрических признаков:** параллельно генерируется более 2500 признаков, включая лексические, морфологические (с помощью pymorphy2), N-граммы символов, а также **N-граммы частей речи** для захвата синтаксических конструкций.
4. **Отбор и объединение:** из стилометрического набора с помощью фильтрового метода SelectKBest (на основе ANOVA F-test) отбираются **200 наиболее информативных** признаков. Они объединяются с 768 признаками от BERT, формируя итоговый вектор размерностью 968.
5. **Классификация:** для обучения используется классификатор **LightGBM** с параметрами, подобранными в ходе предварительных экспериментов для достижения баланса между сложностью и регуляризацией (n_estimators=600, learning_rate=0.02, num_leaves=41, reg_alpha=0.1, reg_lambda=0.1).
6. **Оценка:** для получения надежной оценки применяется **ансамблирование**: вся процедура 5-блочной стратифицированной кросс-валидации повторяется 3 раза с разным случайным состоянием (random_state) для усреднения результатов и минимизации влияния случайности разделения данных.

Оценка подхода:

- **Преимущества:**
  - **Высочайшая точность (SOTA):** комбинация глубоких семантических представлений и мощного классификатора обеспечивает лучшие результаты.
  - **Стабильность и надежность:** отказ от дообучения BERT и использование ансамблирования устранили проблему нестабильности и переобучения.
  - **Автоматизация и гибкость:** модель автоматически извлекает семантические признаки, которые эффективно дополняются классической стилометрией.
- **Недостатки:**
  - **Сложность интерпретации:** хотя **LightGBM** позволяет оценить важность признаков, логика его работы остается сложной для понимания.
  - **Вычислительные требования:** извлечение BERT-эмбеддингов остается

ресурсоемкой задачей.

ЭВОЛЮЦИЯ ПОДХОДОВ И ПРИНЯТЫЕ РЕШЕНИЯ

Путь к успешной архитектуре (v11) был итеративным и позволяет сделать важные выводы о работе с современными NLP-моделями на ***малых данных***.

1. Отказ от дообучения (Fine-tuning) чистого BERT

Первоначальные эксперименты были сфокусированы на дообучении нейросетевых классификаторов на основе BERT. Это современный и мощный подход, однако в данном проекте он столкнулся с непреодолимыми трудностями, связанными с малым объемом выборки (278 документов).

- **v2 (частичная заморозка):** модель, в которой были заморожены нижние 6 из 12 слоев BERT, показала лучший результат среди нейросетевых подходов (**48.2%**). Это был компромисс, позволяющий адаптировать верхние слои под задачу, не давая модели «забыть» базовые знания о языке. Однако результат был нестабилен и сильно зависел от случайного разделения данных.
- **v3 (полная разморозка):** попытка дообучить все слои BERT привела к классическому **переобучению**. Слишком мощная модель с миллионами параметров идеально «запомнила» обучающие примеры, но потеряла способность к обобщению, что привело к падению точности на валидационных данных (лучшая точность упала до 36.4%).
- **v4 (усиленная регуляризация):** увеличение dropout и другие методы регуляризации, призванные бороться с переобучением, привели к обратному эффекту – **недообучению**. Модель стала слишком «скованной», не смогла извлечь из данных полезный сигнал, и точность снова снизилась (до 29.1%).

Стало очевидно, что попытка дообучить сложную нейросеть на малом датасете – это тупиковый путь, ведущий либо к переобучению, либо к недообучению. Система достигла потолка для данной архитектуры.

2. Переход к гибридной модели с классическим ML

Осознав проблемы с дообучением, мы кардинально сменили стратегию, что и привело к успеху:

1. **BERT как экстрактор признаков:** вместо дообучения мы стали использовать BERT только для одной задачи – преобразования текста в качественный семантический вектор (эмбеддинг). Это позволило «заморозить» самую мощную, но нестабильную часть системы.
2. **Гибридизация:** семантические признаки от BERT были объединены с большим набором классических стилометрических и синтаксических признаков.
3. **Замена классификатора:** вместо нейронной сети был использован

LightGBM – алгоритм, который идеально подходит для работы с готовыми табличными данными и гораздо более устойчив на малых выборках.

Эта новая архитектура (v6-v9) сразу дала стабильный результат, который превысил пиковую точность нейросетевых подходов, достигнув **67.3%**.

3. Финальный прорыв: упрощение задачи

Даже с новой архитектурой мы уперлись в потолок, ограниченный сложностью задачи. Решающим шагом стало упрощение 4-классовой задачи до **бинарной классификации** (1904-1905 гг. против 1906-1907 гг.). Это не только логично с исторической точки зрения (события 1905 года как водораздел), но и позволило модели сфокусироваться на более четком и сильном сигнале в данных.

Именно это изменение в сочетании со стабильной архитектурой и ансамблированием позволило преодолеть целевой порог и достичь итоговой точности **87.0%**.

Источники

[1] Burrows, J. F. (2002). ‘Delta’: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing, 17*(3), 267-287.

[2] Mikolov, T., et al. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[3] Tripto N. I., Ali M. E. (2023). The Word2vec graph model for author attribution and genre detection in literary analysis. *arXiv preprint arXiv:2310.16972*.

[4] Ke, G., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems, 30*.

[5] Kuratov, Y., et al. (2019). Adaptation of deep bidirectional transformers for Russian language. *arXiv preprint arXiv:1905.07213*.