Отчёт по лабораторной работе №1

Классификация текстов на основе стилометрических характеристик

А. А. Аладьина, В. Н. Копнин, М. В. Савостиков

Направление: автоматический анализ текстов

Корпус: Stardew Valley · 2005 реплик · 34 персонажа

01.11.2026

Аннотация

Работа посвящена стилометрической классификации реплик персонажей игры Stardew Valley. Корпус содержит 2005 реплик, принадлежащих 34 персонажам; все реплики написаны одним автором. Проверяется, различимы ли речевые маски персонажей по отдельной реплике, если исключить авторскую атрибуцию как таковую.

Построены 454 ручных стилометрических признака, символьные и частеречные n-граммы, а также несколько вариантов латентно-семантического анализа. Сравнивались логистическая регрессия, линейный метод опорных векторов, наивный байесовский классификатор, случайный лес, сверхслучайные деревья, метод ближайших соседей и дельта Барроуза. Использована повторная стратифицированная перекрёстная проверка, вложенная настройка параметров и отдельная отложенная выборка.

Лучший наблюдаемый результат — мягкое голосование трёх моделей: macro-F1 около **0,304** при перекрёстной проверке и **0,333** на отложенной выборке. Консервативная оценка при вложенном выборе конфигурации — **0,275**. Случайный уровень для 34 классов — **0,029**. Основной вклад дают лексико-тематические признаки и символьные n-граммы; после замены содержательных слов нейтральным маркером качество падает с 0,248 до 0,113. Следовательно, результат корректно интерпретировать как различимость речевых масок по лексико-тематическому профилю, а не как доказательство «чистого» идиолекта.

1. Постановка задачи

1.1. Направление исследования

Направление работы — классификация текстов по стилометрическим характеристикам. В отличие от классической авторской атрибуции, все тексты корпуса принадлежат одному автору. Исследуется другой вопрос:

> различаются ли речевые маски 34 персонажей по стилометрическим признакам отдельных реплик, если автор у них общий?

Такая постановка позволяет проверить, насколько автор намеренно создаёт разные голоса персонажей, и какие признаки несут этот сигнал.

1.2. Гипотезы

Таблица 1. Гипотезы исследования и способ их проверки

Обозначение	Гипотеза
H1	Функциональные слова, части речи и символьные n-граммы различают персонажей выше случайного уровня
H2	На коротких репликах частеречные и символьные признаки устойчивее словарного богатства и читаемости
H3	LSA по всем словам отражает тему, а LSA по частям речи и функциональным словам — стиль
H4	Качество зависит от длины реплики
H5	Ошибки группируются вокруг персонажей со сходным возрастом, полом, ролью или видом

1.3. Формальная постановка

Пусть $D=\{(x_i, y_i)\}_{i=1}^{n}$ — корпус, где $n=2005$, x_i — одна реплика, $y_i \in Y$ — имя персонажа, $|Y|=34$. Каждый объект относится ровно к одному классу.

Требуется построить классификатор

$$f: \Phi(x_i) \to Y,$$

где $\Phi(x_i)$ — вектор признаков реплики. Качество оценивается макро-усреднённой F-мерой:

$$macroF1 = \frac{1}{|Y|} \sum_{c \in Y} \frac{2 P_c R_c}{P_c + R_c}.$$

Макро-усреднение выбрано потому, что все 34 персонажа важны одинаково, а классы неравномерны.

2. Корпус

2.1. Источник и единица анализа

Корпус собран из диалогов игры Stardew Valley. Единица анализа — одна реплика персонажа; тексты на английском языке. Метка — имя говорящего, определённое по игровым диалогам и проверенное вручную.

Метаданные о поле, возрасте, виде и роли персонажа хранятся отдельно и используются только для анализа ошибок. В признаки классификатора они не подаются, чтобы не подменять стилометрическую задачу демографической.

2.2. Сбор и разметка

Правила формирования корпуса:

1. Из игровых диалогов извлекаются реплики персонажей.
2. Каждая реплика становится отдельным объектом.
3. Метка — имя говорящего; один объект относится ровно к одному классу.
4. Проводится ручная проверка соответствия реплики и метки.
5. Метаданные персонажей ведутся отдельной таблицей.
6. Служебные метки игрового движка (`<PLAYER>`, `<FARM>`, `<PET>`, `<SPOUSE>`, `<TIME>`, `<NOUN>`, `<NAME>`) заменяются нейтральными словами; их количество сохраняется отдельным признаком.

Корпус хранится в CSV с разделителем ; в кодировке cp1251; обязательные колонки — «Реплика» и «Персонаж». Метаданные персонажей ведутся отдельным XLSX-файлом. Такое разделение позволяет менять признаки, не затрагивая разметку, и проверять корпус независимо от моделей.

2.3. Очистка и шум

Выполнялась минимальная нормализация:

- Unicode-очистка: разные варианты кавычек, тире, многоточий и пробелов приводятся к единому виду;
- замена плейсхолдеров нейтральными словами;
- токенизация и разбиение на предложения средствами NLTK;
- частеречная разметка Penn Treebank с переводом в Universal POS.

Не удалялись:

- пунктуация;
- регистр;
- стоп-слова;
- короткие реплики, оставшиеся после ручной фильтрации.

Пунктуация, регистр и служебные слова рассматриваются как стилометрический сигнал, а не как шум.

Шумом считались:

- точные дубликаты;
- технические артефакты;

- общие короткие формулы;
- плейсхолдеры игрового движка.

Корпус прошёл ручную фильтрацию: короткие и общие фразы удалены. Точные правила этой фильтрации не сохранились, поэтому исходное распределение восстановить нельзя. Это ограничение зафиксировано в выводах.

2.4. Характеристики корпуса

Таблица 2. Характеристики корпуса

Показатель	Значение
Реплик	2005
Персонажей	34
Словоупотреблений	38 207
Средняя длина реплики	19,1 слова
Медиана длины	17 слов
Минимум / максимум	3 / 78 слов
Классы	15–139 реплик
Точные дубли реплик	0
Почти-дубли (косинусная близость ≥ 0,8)	13 пар, все внутри одного класса
Реплик с плейсхолдерами	172

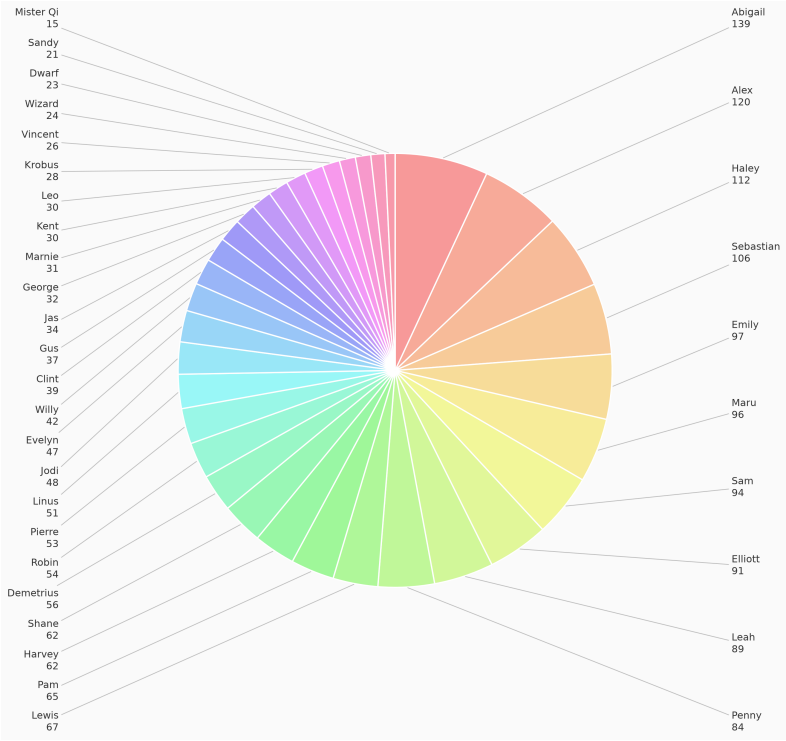


Рисунок 1. Распределение числа реплик по персонажам

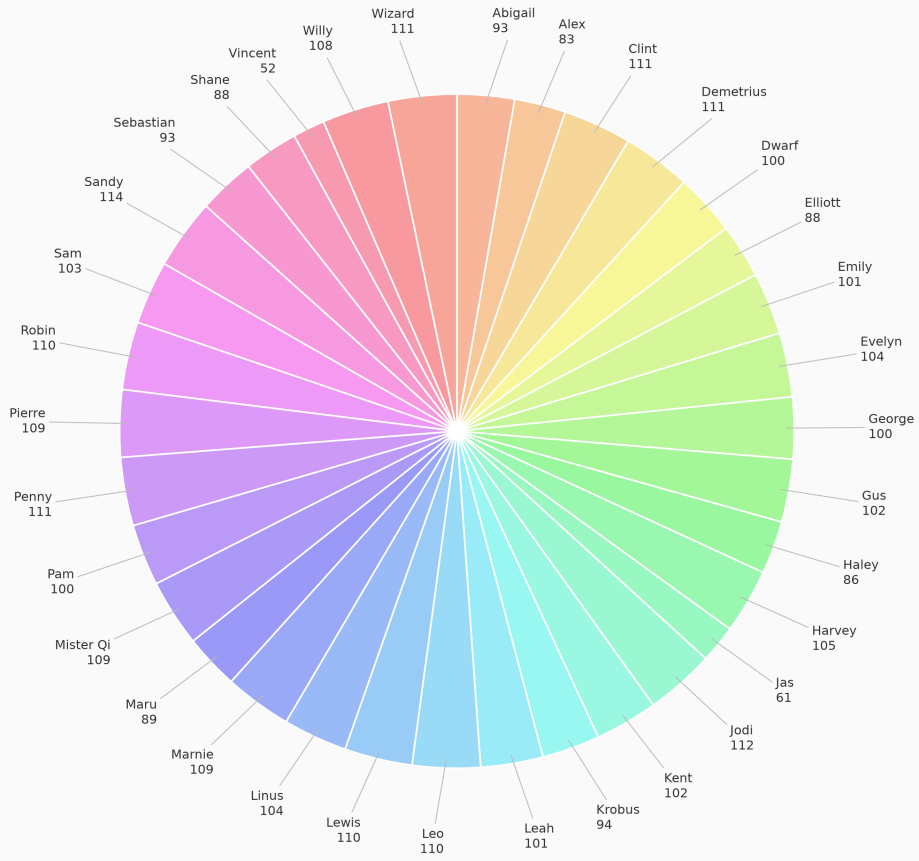


Рисунок 2. Медианная длина предложения в символах по персонажам



Рисунок 3. Персонажи корпуса

Классы неравномерны: максимум — Abigail (139 реплик), минимум — Mister Qi (15 реплик). Разница почти десятикратная. Поэтому доля правильных ответов может вводить в заблуждение: модель, всегда выбирающая частый класс, получает заметную accuracy, но низкую macro-F1.

2.5. Соответствие формальным требованиям

Задание предусматривает не менее 90 текстов объёмом не менее 30 предложений и примерно равные классы. В данном корпусе единица анализа — реплика; она короче формального «текста» и содержит в среднем 19 слов. По согласованию с преподавателем эксперимент выполнен на уровне реплик: это позволяет использовать все 2005 объектов и корректно поставить задачу для коротких игровых реплик. Классы не равны по объёму; для компенсации использованы веса классов и macro-F1.

Это ограничение важно учитывать при сравнении с работами, где документы содержат десятки предложений.

3. Стилометрические признаки

3.1. Ручные признаки: 454 признака в 11 группах

Таблица 3. Группы ручных стилометрических признаков

Группа	Примеры
Длина	число слов, символов, предложений; длина предложения
Лексика	длины слов, апострофы, заглавные буквы, цифры
Словарное богатство	TTR (доля уникальных слов), Guiraud, Herdan, Yule’s K, Simpson’s D, доля гапаксов и дислегомен
Символы	частоты букв, регистр, цифры, пунктуация
Пунктуация	запятые, точки, многоточия, вопросы, восклицания, повторы
Функциональные слова	артикли, предлоги, союзы, местоимения, вспомогательные глаголы
Местоимения	1-е, 2-е, 3-е лицо; отношение «я/ты»
Части речи	Universal POS, Penn Treebank, время, модальность, герундий
Синтаксис	клаузы (элементарные предложения), подчинительные союзы, пассивные конструкции, вопросы
Читаемость	Flesch, Flesch—Kincaid, Gunning Fog, SMOG, Coleman—Liau, ARI
Разговорные маркеры	междометия, смех, осторожные формулировки (хеджи), интенсификаторы, вежливость, сокращения

Словарное богатство и читаемость помечены как предварительные: на репликах длиной 10–20 слов они нестабильны. Основные группы для коротких текстов — функциональные слова, части речи, пунктуация, символьные n-граммы и разговорные маркеры.

Обоснование выбора

- Функциональные слова — классическое ядро стилометрии: закрытый класс мало зависит от темы.
- Части речи и синтаксические прокси отражают структуру фразы.
- Символьные n-граммы устойчивы к теме и фиксируют морфологию, пунктуацию и орфографию.
- Разговорные маркеры важны для игровых реплик: междометия, многоточия, смех и вежливость различают персонажей.
- LSA сжимает разреженное пространство и позволяет сравнить тематический и стилистический вклады.

3.2. Символьные и частеречные n-граммы

- Символьные n-граммы: TF-IDF по символьным n-граммам в границах слов (`char_wb`), диапазон 2–5, `min_df=3`, до 5000–10 000 признаков, L2-нормализация.
- Частеречные n-граммы: TF-IDF по последовательностям Universal POS, диапазон 1–3, до 500 признаков.

TF-IDF — вес признака, который растёт с частотой внутри реплики и падает с частотой в корпусе: частые общие n-граммы получают меньший вес, редкие характерные — больший.

Символьные n-граммы фиксируют морфологию и оформление; n-граммы частей речи — синтаксические шаблоны. Оба блока устойчивее к теме, чем словарные n-граммы.

3.3. Латентно-семантический анализ

LSA реализован как:

```
TF-IDF -> усечённое сингулярное разложение (TruncatedSVD) -> L2-нормализация
```

Варианты:

Таблица 4. Варианты латентно-семантического анализа

Вариант	Что отражает
word	лексику и тему
word2	лексику и устойчивые сочетания
content	содержание после удаления стоп-слов
fw	функциональные слова и синтаксис
char	морфологию и пунктуацию
pos	частеречные последовательности

Число компонент подбиралось в диапазоне 50–400 по объяснённой дисперсии и внешней метрике. Для `content`-LSA качество росло до 400 компонент. Ключевое ограничение: LSA по содержательным словам может отражать тему, а не только стиль. Поэтому варианты сравнивались между собой, а не выбирались заранее.

4. Классификаторы

В работе использовано шесть семейств ML-классификаторов и классическая дельта Барроуза — всего семь подходов; требование «не менее трёх» выполнено с запасом.

4.1. Логистическая регрессия

Линейная модель: для каждого класса вычисляется оценка $s_c(x) = w_c^{\top} x + b_c$, затем вероятности через softmax. Обучение минимизирует перекрёстную энтропию с L2-регуляризацией. Параметр C управляет силой регуляризации; `class_weight="balanced"` повышает вес редких классов.

Сильные стороны: устойчивость при большом числе признаков, вероятности, интерпретируемые коэффициенты. Ограничение: только линейные границы.

4.2. Линейный метод опорных векторов

Ищет линейную границу с максимальным зазором; параметр C штрафует нарушения зазора. Не выдаёт вероятностей: для тройки лучших используется решающая функция. Силён на текстовых признаках, но требует настройки C.

4.3. Случайный лес

Ансамбль решающих деревьев: каждое дерево обучается на подвыборке, разбиения используют случайное подмножество признаков. Класс определяется голосованием. Лес ловит нелинейные взаимодействия, но склонен к переобучению на 2005 объектах; ограничивались глубина и размер листьев.

4.4. Сверхслучайные деревья (ExtraTrees)

Похожи на случайный лес, но пороги разбиения выбираются случайно. Это уменьшает дисперсию и ускоряет обучение. В работе ExtraTrees с расширенным символьным словарём показали лучший результат среди одиночных моделей.

4.5. Наивный байесовский классификатор

MultinomialNB использует теорему Байеса и предположение об условной независимости признаков; сглаживание Лапласа задаётся параметром α. Быстрый baseline для частотных признаков; требует неотрицательных значений, поэтому LSA-компоненты для него исключались.

4.6. Метод ближайших соседей

Класс определяется по k ближайшим объектам; использовалось косинусное расстояние и веса, обратные расстоянию. Простая локальная модель; чувствительна к размерности и масштабу признаков.

4.7. Дельта Барроуза

Классический стилометрический метод: выбираются наиболее частотные слова, частоты z-нормируются по обучающей выборке, документ относится к классу с минимальным расстоянием до центроида. Использовались манхэттенское и косинусное расстояния. Дельта — независимая проверка, не заменяющая ML-классификаторы.

4.8. Ансамбли

- **Мягкое голосование** усредняет вероятности нескольких моделей.
- **Объединение через мета-классификатор (стекинг)** обучает второй уровень на предсказаниях базовых моделей. В наших данных оно переобучилось и оказалось хуже мягкого голосования.

5. Схема верификации

5.1. Перекрёстная проверка

Основная схема:

```
RepeatedStratifiedKFold(n_splits=5, n_repeats=3–5)
```

При 5 фолдах модель обучается на 4 фолдах и проверяется на 1; доли классов сохраняются внутри фолдов. Повторность позволяет оценить разброс.

5.2. Вложенная проверка

Внешние фолды используются для оценки качества, внутренние — для настройки гиперпараметров. Дополнительно выполнен вложенный выбор конфигурации признаков и

модели: в каждом внешнем фолде конфигурация выбиралась только по внутренним фолдам.

5.3. Контроль утечек

- TF-IDF, SVD, масштабирование и отбор признаков обучаются только на обучающих фолдах.
- Тестовая выборка не участвует в выборе гиперпараметров или признаков.
- Контроль перестановкой меток дал macro-F1 0,0270–0,0282 — случайный уровень.
- Все 13 почти-дубликатов находятся внутри одного класса; межклассовых почти-дублей не обнаружено, поэтому прямой утечки метки нет. При более строгой оценке такие пары всё равно следует держать в одном фолде.

5.4. Метрики

Таблица 5. Метрики качества и их назначение

Метрика	Что показывает
Macro-F1	Основная метрика: средняя F-мера по 34 классам
Доля правильных ответов	Доля совпавших ответов; обманчива при дисбалансе
Сбалансированная доля	Средняя полнота по классам
Top-3	Истинный класс попал в тройку лучших
MCC (коэффициент корреляции Мэтьюза)	Корреляция истинных и предсказанных меток
Каппа Коэна	Согласие с поправкой на случай
Log loss / Brier	Качество вероятностей
Матрица ошибок	Какие персонажи перепутываются

Дополнительно использованы бутстрэп-интервалы, критерий Мак-Немара для парных сравнений, критерий Уилкоксона и поправка Холма—Бонферрони.

6. Результаты

6.1. Сравнение моделей

Таблица 6. Сравнение конфигураций классификаторов

Конфигурация	Macro-F1	Доля правильных ответов	Top-3
Мягкое голосование трёх моделей	0,3036 ± 0,0189	0,2808	0,4823
ExtraTrees, char_max_features=10000	0,2757 ± 0,0219	0,2633	0,4554
Вложенный выбор конфигурации	0,2746 ± 0,0233	0,2603	—
Случайный лес, char_max_features=10000	0,2583 ± 0,0066	0,2391	0,4197
ExtraTrees, полный набор	0,2499 ± 0,0209	0,2374	0,4234
Логистическая регрессия, lean + content-LSA	0,2466 ± 0,0191	0,2349	0,4030
Логистическая регрессия, stylometry_pos_lsa	0,1785 ± 0,0060	0,1679	0,3322
Только ручные признаки	0,1051 ± 0,0038	0,1012	0,2338
Случайный уровень	0,029	0,029	0,088

На отложенной выборке 80/20 мягкое голосование дало **0,3333**; ExtraTrees на полном наборе — 0,2793; случайный лес с `char_max_features=10000` — 0,2554; логистическая регрессия с content-LSA — 0,2459.

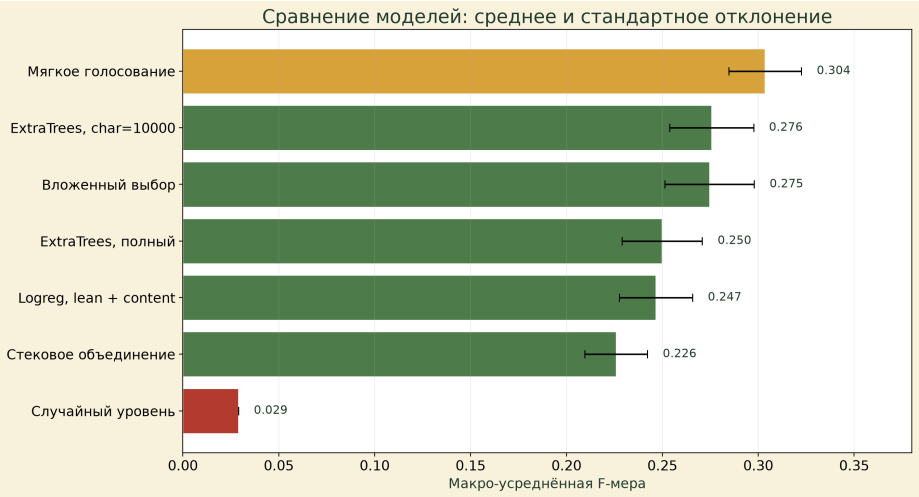


Рисунок 4. Сравнение моделей

6.2. Вклад признаков

Таблица 7. Вклад групп признаков

Условие	Macro-F1
Только длина	0,035
Только ручные стилометрические признаки	0,105
Ручные + символьные и частеречные n-граммы	0,182
То же + content-LSA	0,249
ExtraTrees с расширенным символьным словарём	0,276
Мягкое голосование	0,304

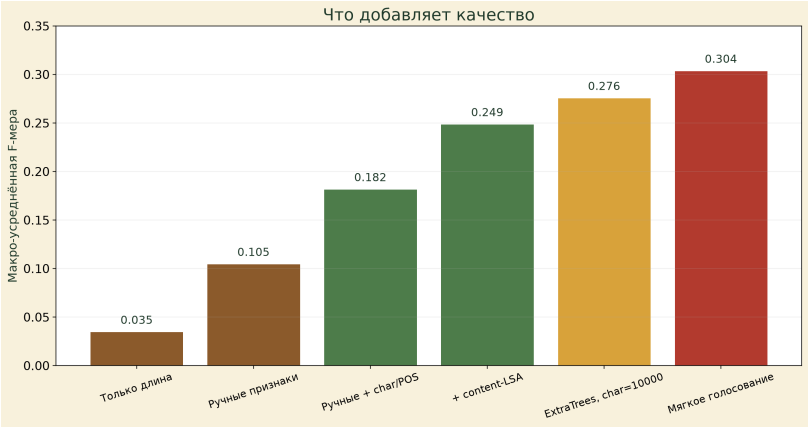


Рисунок 5. Вклад признаков

Основной прирост дают символьные n-граммы, лексика и `content`-LSA. Функциональные слова и POS-LSA сами по себе слабы (0,06–0,07): это важный отрицательный результат.

6.3. Стиль или содержание

Содержательные слова (существительные, глаголы, прилагательные, наречия, числительные) заменялись нейтральным маркером `CONTENT`; служебные слова, пунктуация и частеречные теги сохранялись.

Таблица 8. Контроль содержания: исходный и маскированный текст

Условие	Macro-F1
Исходный текст, стилевой набор, ExtraTrees	0,2483
Маскированный текст, стилевой набор, ExtraTrees	0,1129
Исходный текст, стилевой набор, LogReg	0,1467
Маскированный текст, стилевой набор, LogReg	0,1067

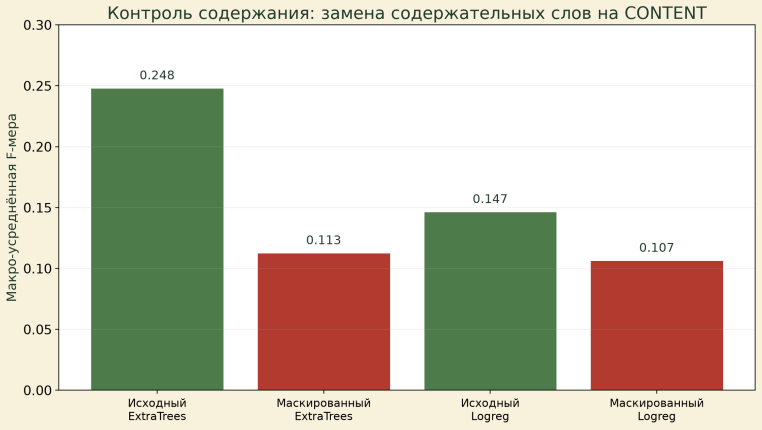


Рисунок 6. Контроль содержания

Падение более чем вдвое показывает, что значительная часть качества связана с лексическим содержанием. Остаток 0,113 всё ещё в 3,9 раза выше случайного уровня: некоторый стилевой сигнал сохраняется, но он слабее тематического.

6.4. Матрицы ошибок

Строки матрицы — истинный персонаж, столбцы — предсказание модели, диагональ — правильные ответы, вне диагонали — ошибки. Полная матрица 34×34 на доске нечитаема; поэтому в основном тексте показаны 12 самых частых персонажей и столбец «прочие», а также 18 самых частых пар ошибок.

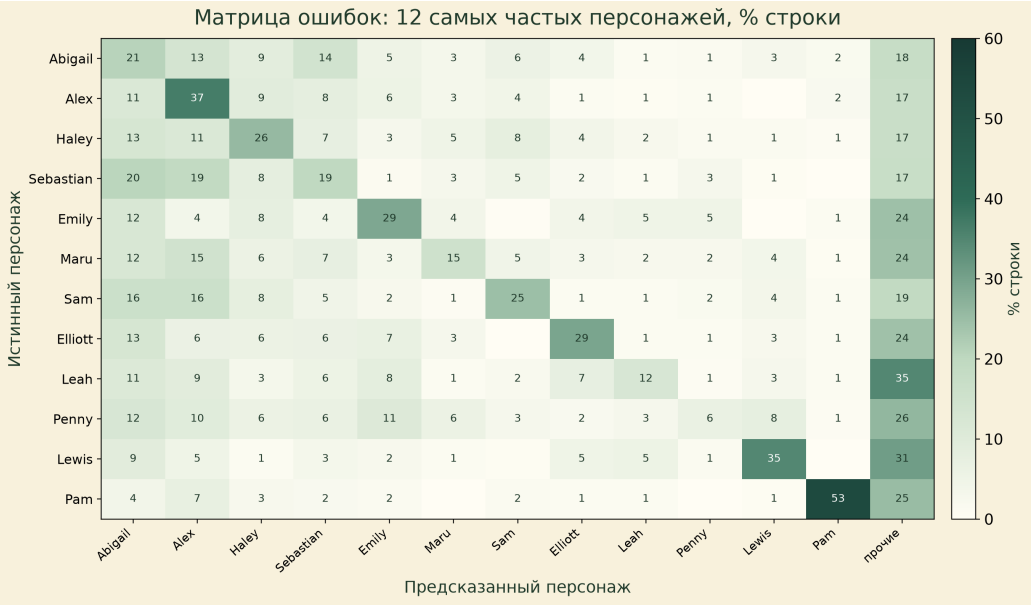


Рисунок 7. Матрица ошибок: 12 персонажей

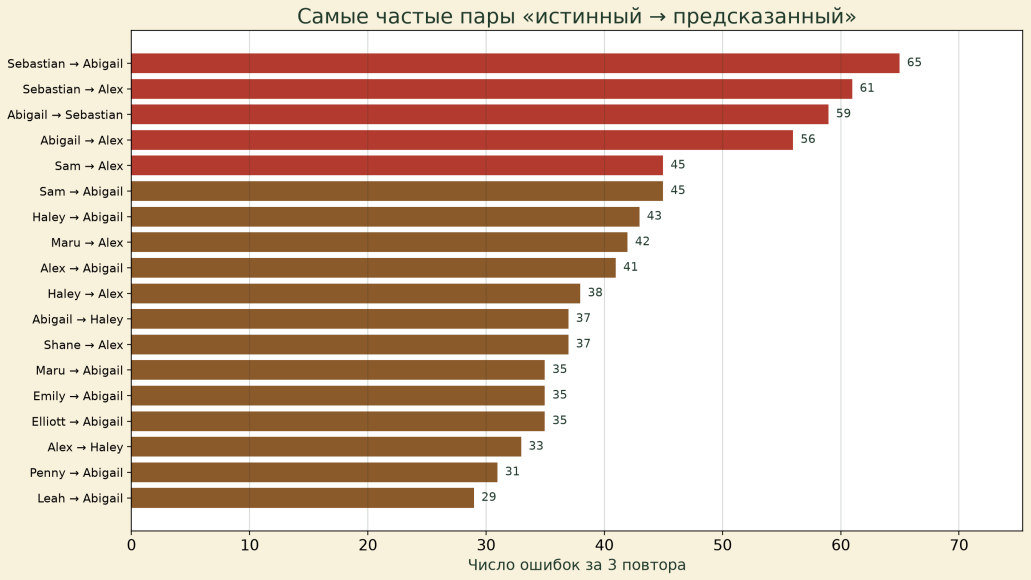


Рисунок 8. Самые частые пары ошибок

Основные пары ошибок (за три повтора): Sebastian → Abigail (65), Sebastian → Alex (61), Abigail → Sebastian (59), Abigail → Alex (56). Крупнейший клубок — молодые персонажи Abigail, Alex, Sebastian, Haley, Sam. Многие из них писались в одних и тех же обновлениях и сюжетных контекстах, поэтому сходство голосов ожидаемо; это не разные авторы.

6.5. Ошибки по персонажам



Рисунок 9. F-мера по персонажам

Лучшие результаты по F-мере: Willy, Evelyn, Pierre, Pam, Leo, Vincent. Худшие: Kent, Mister Qi, Penny, Abigail, Leah, Jodi. Редкие классы имеют широкие интервалы; у Mister Qi всего 15 реплик.

Abigail и Alex являются классами-аттракторами: модель ошибочно относит к ним чужие реплики. Abigail получает 10–18% всех предсказаний при истинной доле около 6,9%. Простая коррекция априорных вероятностей классов не улучшила macro-F1.

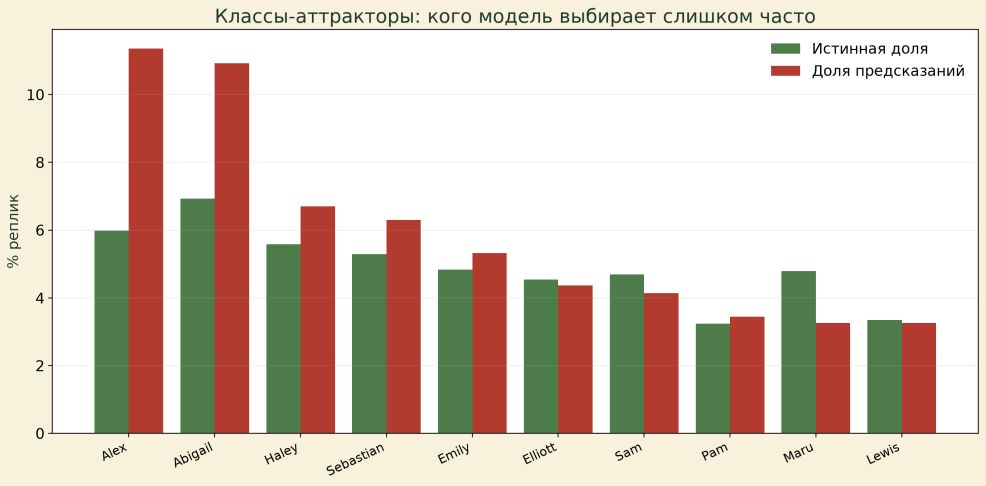


Рисунок 10. Классы-аттракторы

6.6. Длина реплики

Таблица 9. Качество классификации по длине реплики

Длина	Macro-F1
0–5 слов	0,097
6–10	0,244
11–15	0,233
16–20	0,239
21–30	0,295
31+	0,271

Зависимость немонотонная: максимум приходится на 21–30 слов, после 30 слов качество снижается. Гипотеза «чем длиннее, тем лучше» не подтверждается.

6.7. Верификация и кластеризация

- Верификация «свой/чужой»: ROC-AUC 0,630–0,633; EER (равная ошибка ложного принятия и ложного отказа) 0,412–0,415; top-1 0,133–0,141. Выше случайного, но практически слабо.
- Кластеризация: ARI (скорректированный индекс Рэнда) = 0 для k = 3/5/8; NMI (нормированная взаимная информация) 0,14–0,53. Устойчивых кластеров, соответствующих персонажам, нет.

6.8. Калибровка и ансамбли

Коррекция априорных вероятностей вида $p'_{ic} \propto p_{ic}\pi_c^{-\alpha}$ проверялась по схеме «один повтор — отложенный, остальные — обучающие». Среднее изменение macro-F1 составило **−0,0022**: простой пересчёт приоров не помог.

Мягкое голосование трёх моделей (ExtraTrees на полном наборе, ExtraTrees с `char_max_features=10000`, логистическая регрессия с lean + content-LSA) дало **0,3036 ± 0,0189**. Объединение через мета-классификатор оказалось хуже (0,2259 ± 0,0163) из-за переобучения на малом числе объектов.

7. Обсуждение

1. **Персонажи статистически различимы, но сигнал умеренный.** 0,304 macro-F1 при 34 классах — примерно в 10 раз выше случайного уровня 0,029. Абсолютное значение ограничено короткими репликами, дисбалансом и общими темами.
2. **Основной вклад — лексико-тематический.** Символьные n-граммы и `content`-LSA дают основной прирост; функциональные слова и POS-LSA сами по себе слабы. После маскирования содержательных слов качество падает более чем вдвое.
3. **Наиболее устойчивая одиночная модель — ExtraTrees с `char_max_features=10000`.** Она выбрана в 14 из 15 внешних фолдов вложенной процедуры и даёт 0,2757 ± 0,0219.
4. **Мягкое голосование — лучший наблюдаемый результат**, но состав ансамбля выбирался по предыдущим сериям. Для строгого вывода нужна внешняя валидация.
5. **Ошибки — это коллапс редких классов в классы-аттракторы**, а не доказательство сходства характеров. Abigail, Alex и ещё несколько популярных классов забирают чужие реплики.
6. **Один автор и общие контексты** делают путаницу между персонажами ожидаемой. Это не означает, что автор «не старался»: превышение случайного уровня как раз говорит о намеренно созданных речевых масках.

8. Ограничения

- Один корпус, одна игра, один автор; внешняя валидация не проводилась.
- Единица анализа — реплика, а не текст из 30 предложений; это согласовано, но ограничивает сравнение с документными корпусами.
- Классы неравномерны; редкие классы имеют широкие интервалы.

- Ручная фильтрация корпуса невоспроизводима.
- Нет идентификаторов сцен и диалогов.
- Выбор признаков и ансамбля частично опирался на предыдущие серии; вложенная проверка снижает, но не устраняет оптимизм.
- Содержательные слова и тема вносят большой вклад; «чистый стиль» не доказан.
- Простая коррекция приоров не помогла; калибровка редких классов требует дальнейшей работы.

9. Выводы

1. Речевые маски 34 персонажей статистически различимы по отдельным репликам: лучшая модель — 0,304 macro-F1, консервативная оценка — 0,275 при случайном уровне 0,029.
2. Основной сигнал — лексико-тематический: символьные n-граммы и `content`-LSA дают основной прирост; функциональные слова и частеречная LSA сами по себе слабы.
3. После замены содержательных слов качество падает с 0,248 до 0,113: содержание доминирует, но стилевой сигнал сохраняется.
4. Главная проблема — редкие классы и классы-аттракторы; мягкое голосование помогает, но не решает её полностью.
5. Корректная формулировка результата: различимы речевые маски по лексико-тематическому профилю; утверждение о «чистом стиле» требует внешней валидации и тематического контроля.

10. Воспроизводимость

10.1. Структура проекта

```
lab1/
├── stylometry/          # чтение, признаки, LSA, модели, метрики
├── scripts/             # проверка окружения, запуск серий, упаковка
├── tests/               # pytest
├── configs/             # конфигурация
├── data/                # корпус и метаданные
├── report/              # этот отчёт
├── presentation/        # PPTX/PDF и LaTeX-версия
├── README.md
├── RUNBOOK.md
├── EXPERIMENTS.md
├── FEATURES.md
├── METRICS_AND_CLASSIFIERS.md
├── TUNING.md
├── FINAL_REPORT.md
├── FINAL_RESULTS.csv
├── requirements.txt
└── requirements-report.txt
```

Краткая и подробная инструкции к архиву (`START_HERE.md`, `ARCHIVE_README.md`) лежат в его корне, рядом с каталогом `lab1`.

10.2. Установка и запуск

```
cd lab1
python -m venv .venv
source .venv/bin/activate
python -m pip install -r requirements.txt
python scripts/download_nltk_data.py

python scripts/check_environment.py
python -m pytest -q

# аудит и признаки
python -m stylometry.cli audit
python -m stylometry.cli features

# основной эксперимент
python -m stylometry.cli run \
    --preset stylometry_pos_lsa --model logreg --repeats 5

# полная серия
bash scripts/run_all_experiments.sh

# дополнительные эксперименты
python scripts/run_final_experiments.py \
    --stages nested topic calibration ensemble extra \
    --out results/final

# пересборка отчёта (pandoc + LibreOffice)
python scripts/build_report.py
```

10.3. Программное обеспечение

- Python 3.11+;
- NumPy, pandas, SciPy, scikit-learn, matplotlib;

- NLTK (`punkt`, `punkt_tab`, `averaged_perceptron_tagger_eng`, `stopwords`);
- python-pptx и Pillow — для презентации;
- pandoc, LibreOffice, python-docx и PyMuPDF — для сборки отчёта;
- XeLaTeX и Beamer — для LaTeX-версии презентации.

Приложение A. Полная матрица ошибок

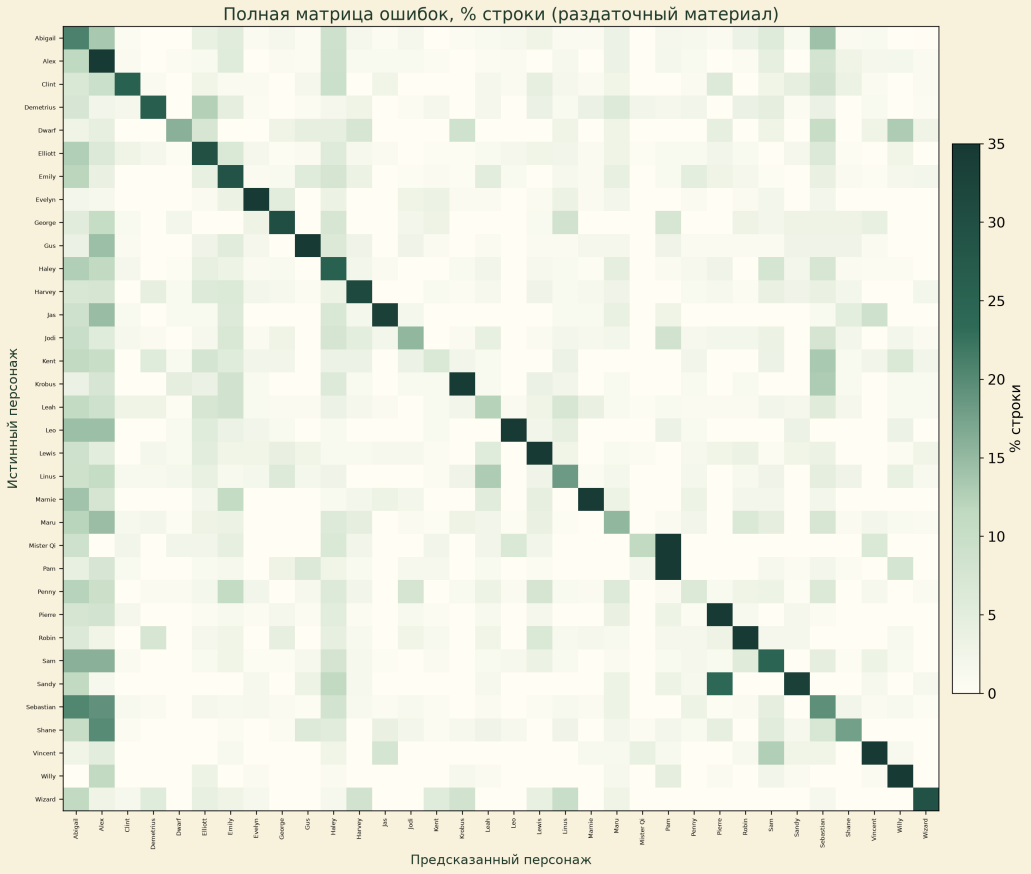


Рисунок 11. Полная матрица ошибок 34×34

Приложение B. Сводная таблица результатов

См. `FINAL_RESULTS.csv` в корне проекта.

Приложение C. Состав архива

В архив включены код, тесты, конфигурации, корпус, метаданные, документация, отчёт, презентация (PPTX и LaTeX), сводные результаты и инструкции запуска. Виртуальное окружение, кэш NLTK и промежуточные матрицы признаков в архив не включены: они восстанавливаются скриптами.

Приложение D. Источники и программные средства

1. Burrows, J. F. «Delta»: a Measure of Stylistic Difference and a Guide to Likely Authorship // Literary and Linguistic Computing. 2002. Vol. 17, no. 3. P. 267–287.
2. Eder, M., Rybicki, J., Kestemont, M. Stylometry with R: A Package for Computational Text Analysis // The R Journal. 2016. Vol. 8, no. 1. P. 107–121.
3. Mosteller, F., Wallace, D. L. Inference and Disputed Authorship: The Federalist. Reading, MA: Addison-Wesley, 1964.
4. Stamatatos, E. A Survey of Modern Authorship Attribution Methods // Journal of the American Society for Information Science and Technology. 2009. Vol. 60, no. 3. P. 538–556.
5. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830.
6. Bird, S., Klein, E., Loper, E. Natural Language Processing with Python. Sebastopol, CA: O’Reilly Media, 2009.
7. ConcernedApe. Stardew Valley. Video game. 2016.