Отлично, я подберу набор информативных признаков, которые можно извлечь из художественных текстов на русском языке с помощью Python, а также предложу соответствующие библиотеки, метрики качества моделей и baseline-решения. Также постараюсь найти примеры исследований или прикладных систем, где аналогичный подход работал без экспертной разметки.

Я вернусь с результатами, как только всё будет готово.

**Автоматическая оценка художественного текста: признаки, метрики, модели и примеры**

**1. Информативные признаки художественного текста (проза)**

При анализе литературного текста на русском языке можно извлечь широкий спектр признаков, отражающих стиль, содержание и эмоциональное воздействие произведения. Ниже приведены группы таких признаков с примерами и указанием инструментов для их извлечения.

**Лексические признаки**

* **Лексическое разнообразие и богатство**: измеряется соотношением уникальных слов к общему числу слов (Type-Token Ratio) ([[PDF] Lexical Features of Text Complexity: the case of Russian academic ...](https://www.shs-conferences.org/articles/shsconf/pdf/2020/16/shsconf_lltforum2020_01009.pdf#:~:text=...%20www.shs,SHS%20Web%20of)), а также более устойчивыми метриками, как MTLD или индекс Юла. Высокое разнообразие указывает на богатый словарный запас. *Инструменты*: сегментация текста (Razdel) и лемматизация (pymorphy2, Natasha) для учета разных форм слова как одного типа. Затем вычисление TTR или использование готовых функций (например, lexical\_diversity в NLTK или собственной реализации).
* **Лексическая плотность**: доля содержательных слов (существительных, глаголов, прилагательных, наречий) относительно служебных. Показывает, насколько текст насыщен смысловыми элементами. *Инструменты*: морфологический анализ (pymorphy2, Stanza) для классификации частей речи.
* **Средняя сложность лексики**: средняя длина слов (в буквах или слогах) и доля длинных слов. Это характеризует уровень сложности лексикона (например, научный стиль vs. разговорный). *Инструменты*: Razdel для токенизации, pymorphy2 для слогоделения (или подсчет гласных как приближение).
* **Редкие и оригинальные слова**: процент слов, отсутствующих в списке самых частотных слов русского языка, или имеющих низкую частоту в большом референсном корпусе. Высокое значение может указывать на оригинальность и богатство языка. *Инструменты*: готовые частотные словари (например, Национальный корпус русского языка) либо модели типа FastText для проверки редкости слов.
* **Устойчивые выражения и клише**: доля клишированных фраз или штампов. Низкая доля может свидетельствовать об оригинальности текста (). *Инструменты*: словари фразеологизмов и клише, или вычисление перплексии с помощью языковой модели (например, RuGPT-3, RuT5) – высокая перплексия означает менее предсказуемый, более оригинальный текст ().
* **Эмоционально окрашенная лексика**: частота эмоциональных слов (например, *радость, гнев, ужасный, великолепный*). Эта мера отражает эмоциональный тон текста. *Инструменты*: тональные словари (RuSentiment Lexicon) или модели анализа тональности (например, RuBERT, обученный на эмоциональных данных, модель Sentiment Analysis из DeepPavlov).

**Синтаксические признаки**

* **Длина и сложность предложений**: средняя длина предложения в словах, стандартное отклонение длины. Классическая проза часто отличается более длинными и сложными предложениями, тогда как упрощенный стиль – короткими ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels" \l ":~:text=match%20at%20L219%20richness%2C%20a,)). *Инструменты*: Razdel или Natasha для разбивки на предложения; простое вычисление длины.
* **Сложноподчиненность**: доля сложноподчиненных предложений, среднее число придаточных на предложение. Это признак синтаксической сложности и «литературности» текста ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels#:~:text=match%20at%20L219%20richness%2C%20a,)). *Инструменты*: синтаксический анализ с помощью Stanza или Natasha; подсчет союзов подчинения (*что, который, пока, хотя* и т.д.) и относительных местоимений.
* **Глубина синтаксического дерева**: средняя глубина или высота дерева зависимостей, полученного при разборе предложений. Большая глубина свидетельствует о вложенности конструкций. *Инструменты*: Stanza (модель для русского) – получить древо зависимостей и вычислить его глубину.
* **Части речи (POS) и грамматические паттерны**: распределение частей речи по тексту – например, доля существительных vs. глаголов, местоимений и т.д. Известно, что **успешные произведения** отличались повышенной долей существительных и пониженной долей глаголов ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels" \l ":~:text=lower%20proportion%20of%20verbs,focused%20less%20on%20the%20actions)), а также чаще использовали местоимения ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels" \l ":~:text=match%20at%20L279%20Ashok%20et,person%20pronouns%20separately.)). Преобладание существительных указывает на описание статических аспектов (обстановки, деталей), а глаголов – на динамику событий ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels" \l ":~:text=lower%20proportion%20of%20verbs,focused%20less%20on%20the%20actions)). *Инструменты*: морфологический разбор (pymorphy2, Natasha) для тегирования текста и подсчета распределения POS.
* **Синтаксические конструкции**: частота употребления специфических оборотов русского языка – например, причастных и деепричастных оборотов, инфинитивных конструкций, прямой речи. Высокая частота причастий и деепричастий может указывать на книжный стиль. *Инструменты*: pymorphy2 для поиска форм на PRTS (причастие) или GRND (деепричастие) в морфологической разметке; регулярные выражения для обнаружения прямой речи (например, тире в начале реплики или кавычки).
* **Пунктуация**: среднее количество запятых на предложение (показатель сложносочиненности/сложноподчиненности), количество восклицательных знаков и вопросительных (эмфатичность, эмоциональность стиля), эллипсисов “…”. Эти признаки характеризуют авторский стиль и эмоциональную экспрессию.

**Семантические и эмоциональные признаки**

* **Тематические признаки**: распределение тем или семантических полей. Можно получить вектор тематической принадлежности текста с помощью LDA-тем (Latent Dirichlet Allocation) или современных топик-моделей. Например, выявить, о чем текст (война, любовь, технология и пр.) и измерить разнообразие тем. *Инструменты*: Gensim (LDA для русского текста после лемматизации), RuBERT (для embedding-представления абзацев и кластеризации по темам).
* **Семантическая связность (когерентность)**: степень связности между предложениями и абзацами. Формально измеряется когезией: например, средняя косинусная схожесть между соседними предложениями в пространстве эмбеддингов. Высокая связность означает логическую последовательность мыслей. *Инструменты*: модель RuBERT или RuSimCSE (Sentence-BERT для русского) – получение эмбеддингов предложений и вычисление сходства; либо построение **Entity Grid** – матрицы упоминаний сущностей – для оценки связности (требует разрешения кореференций, например библиотека **pullenti** или Natasha для русского).
* **Эмоциональность и тон**: суммарная тональность текста (позитивная, негативная, нейтральная) и **эмоциональный профиль** (распределение по категориям эмоций: радость, грусть, гнев, страх и т.д.). Художественные тексты могут оцениваться по тому, вызывает ли язык эмоциональный отклик. *Инструменты*: модели сентимента (DeepPavlov RuBERT-тональность) для общей оценки тональности; словари NRC или RuSentiment для подсчета по категориям эмоций. Например, подсчет частоты слов радости, печали и пр. на основе переведенного NRC Lexicon.
* **Эмоциональные арки**: изменение эмоциональной тональности по ходу сюжета. Исследования показали, что *«эмоциональный поток»* в литературном тексте связан с оценкой читателей ([(PDF) Letting Emotions Flow: Success Prediction by Modeling the Flow of Emotions in Books](https://www.researchgate.net/publication/325445735_Letting_Emotions_Flow_Success_Prediction_by_Modeling_the_Flow_of_Emotions_in_Books#:~:text=with%20human%20interpretations%20of%20a%E6%98%80%E6%98%80ective,)). Можно разбить текст на главы или равные сегменты и оценить тональность каждого, выявляя динамику (например, чередование напряжения и разрядки). Важны метрики вроде амплитуды изменений или количество смен полярности. *Инструменты*: вышеупомянутые модели сентимента применяются к каждому сегменту; построение графика «эмоциональной дуги» произведения.
* **Семантическая оригинальность**: наличие нестандартных метафор, образов, оригинальных описаний. Автоматически выявить метафоры сложно, но можно использовать модели типа **DeepPavlov** для метафор (есть датасеты по распознаванию метафор на русском) или косвенно — искать слова, необычно сочетающиеся друг с другом (например, вычислять PMI пар слов: низкий PMI по общему корпусу может указывать на уникальное авторское сочетание). Также можно проверить текст на совпадение с известными цитатами или клише. *Инструменты*: подборка списков известных цитат/клишированных выражений; модели распределенных семантических представлений (word2vec/fastText) для оценки семантической близости слов в контексте (метафоры часто представляют аномальные сочетания).

**Стилистические признаки**

* **Читаемость и сложность текста**: количественные индексы читаемости – например, индекс Флеша для русского языка (адаптированная формула учитывает число слов, предложений и слогов), индекс сложности Fog Index. Более сложный для чтения текст может коррелировать с «высокой» литературой, тогда как более простой – с массовой литературой ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels" \l ":~:text=literary%20works%2C%20Ashok%20et%20al,)). Интересно, что исследования на английском показали **обратную** зависимость: великие книги часто менее удобочитаемы из-за высокой концептуальной сложности ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels" \l ":~:text=literary%20works%2C%20Ashok%20et%20al,)). *Инструменты*: библиотека textstat (есть поддержки русского для некоторых метрик), либо собственная реализация формул читаемости.
* **Формальность стиля**: степень официальности или разговорности. Можно измерять отношением книжной лексики к разговорной, наличием жаргонизмов, сленга или диалектных форм. Также помогают известные признаки формальности: доля местоимений 1-го и 2-го лица (разговорная речь обильна «я/ты», формальная избегает личных местоимений), доля наречий vs. прилагательных и т.п. *Инструменты*: списки разговорной лексики/жаргона, морфологический анализ для местоимений и частот.
* **Стилистические фигуры**: использование тропов и фигур речи – метафор, сравнений, эпитетов, аллитераций. Автоматическое распознавание затруднено, но возможны прокси: напр. метафоры – поиск нетипичных сочетаний прилагательных и существительных (семантически далеких), эпитеты – частое употребление прилагательных с оценочной семантикой, аллитерация – повторение букв/звуков (можно проверить повторяемость согласных в соседних словах). *Инструменты*: Natasha для выделения прилагательных + существительных, затем проверка семантической близости через эмбеддинги; поиск повторяющихся символов с использованием регулярных выражений.
* **Коннотации и подтекст**: наличие скрытого смысла, иронии, сарказма. Эти признаки трудноформализуемые, но можно учитывать, например, **тональную двусмысленность** (когда позитивные слова используются в негативном контексте и наоборот) или наличие слов-маркеров иронии (в русском: *ну конечно*, \* прям таки\*, смайлики в сетевых текстах). *Инструменты*: правила или небольшие классификаторы, обученные определять сарказм (на русском есть датасеты твитов с сарказмом, можно адаптировать RuBERT).

**Прагматические признаки**

* **Нарративная перспектива**: от чьего лица ведется повествование. Можно измерить долю местоимений 1-го лица (*я, мы*), 2-го лица (*ты, вы*) и 3-го лица (*он, она, они*). Первый взгляд (от **я**) создает субъективность и интимность, второй – эффект обращения к читателю, третий – более объективный стиль ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels" \l ":~:text=match%20at%20L279%20Ashok%20et,person%20pronouns%20separately.)). *Инструменты*: морфологический анализ (Natasha, Stanza) для местоимений и окончаний глаголов (определять лицо и число).
* **Диалогичность**: вовлечение читателя в диалог. Признаки: прямые обращения к читателю (*дорогой читатель*, *как вы думаете*), риторические вопросы (вопросительные предложения, не требующие ответа), побудительные предложения (императив). Высокая диалогичность повышает ощущение взаимодействия. *Инструменты*: поиск шаблонов с обращениями (**регекс** по словам *читатель*, *дорогой*, *уважаемый*), распознавание императивных форм глагола (через морфоразбор), подсчет вопросительных знаков.
* **Признаки жанрового дискурса**: например, в детской литературе автор может часто напрямую поучать или спрашивать читателя, в публицистике – обращаться к аудитории, в художественной – реже. Это можно учесть как бинарные признаки (наличие/отсутствие обращений, вопросов) или количественно.
* **Реализм vs. субъективность**: измеряется степенью выраженности мнения рассказчика. Например, наличие модальных слов (*возможно, по-моему, кажется*) и конструкций оценочности. Много модальных слов = сильное присутствие голоса автора и его позиции (прагматический эффект доверительности либо, наоборот, неопределенности). *Инструменты*: список модальных слов и вводных конструкций, поиск по тексту.

**Жанровые признаки**

* **Жанровая классификация**: предположительный жанр текста (детектив, фантастика, романтика, классическая проза, и т.д.). Его можно либо определить заранее (по метаданным) и использовать как признак, либо автоматически классифицировать по тексту. Жанр влияет на восприятие: напр. детективы и триллеры обычно динамичны, фантастика наполнена уникальными сущностями, а классическая литература отличается сложным стилем. *Инструменты*: готовые классификаторы жанров (например, RuBERT fine-tuning на корпусе с жанровыми метками), либо набор ручных критериев (например, наличие специфичной лексики: *луч, корабль, магия* – фантастика; *убийство, расследование* – детектив; *любовь, счастье, слезы* – романтика).
* **Жанровые клише**: подсчет типичных для жанра элементов. Например, в фэнтези – вымышленные топонимы/имена (распознавать NER-инструментами необычные собственные имена), в историческом романе – архаизмы и даты, в детективе – диалоги допроса (много вопросительных предложений, обращений типа *господин*). Высокое соответствие жанровым шаблонам может как повышать оценку у целевой аудитории, так и снижать за предсказуемость – это зависит от контекста применения.
* **Смешение жанров**: тоже признак оригинальности – можно измерить степень, в которой текст одновременно содержит лексику разных жанров (например, любовная линия в научной фантастике). Это вычисляется через пересечение словаря жанров или через многоклассовую жанровую классификацию (распределение вероятностей по жанрам, показывающее «размытость» жанра).

**Структурные признаки**

* **Деление на главы и части**: количество глав, средний объем главы, наличие разделов. Восприятие читателя может зависеть от удобочитаемости структуры (короткие главы делают текст динамичнее). *Инструменты*: анализ разметки текста (если главы отмечены номерами или заголовками), либо эвристики по разрывам строк/символам разделителей.
* **Соотношение диалогов и описания**: измеряется долей текста, заключенного в кавычки или после тире (прямая речь), по сравнению с общим объемом. Например, *классические романы* XIX века нередко содержат обширные описания с меньшей долей диалогов, тогда как массовая современная проза часто имеет больше диалогов. Этот признак влияет на темп повествования и вовлеченность.
* **Композиция повествования**: формальные признаки начала, развития и конца. Например, наличие пролога/эпилога, кульминационных глав (можно косвенно определить через всплеск эмоциональных или экшен-слов ближе к середине-концу текста). *Инструменты*: простейший подход – деление текста на части (начало, середина, конец) и сравнение характеристик (эмоциональный тон, упоминание главных конфликтов) между ними. Более продвинутый – попытка автоматически выделить фабульные элементы, но это исследовательская задача (есть попытки на основе схемы Проппа для сказок и др.).
* **Графические элементы**: в некоторых текстах используются письма, дневниковые записи, стихотворные вставки. Их наличие можно выявлять по маркерам (обращения в начале письма, стихотворный размер). Эти вставки влияют на восприятие (повышают интерактивность, эстетическую ценность). *Инструменты*: регулярные выражения для шаблонов писем (*Дорогой ..., ... С уважением*), проверка текста на стихотворный размер (например, библиотека **rupo** для русского стихосложения – полезно, если в тексте есть поэтические вставки).

Каждый из перечисленных признаков можно извлечь с помощью открытых NLP-библиотек для русского языка. **Natasha** предоставляет удобные обертки для разбиения текста, морфологического и синтаксического анализа. **Stanza** (русская модель StanfordNLP) даст леммы, POS-теги и зависимости. **DeepPavlov** содержит готовые модели классификации (например, тональность, NER) и даже предобученные трансформеры (RuBERT). С помощью **pymorphy2** можно быстро лемматизировать текст и получить морфологические характеристики, а **razdel** пригодится для токенизации без лишних зависимостей. При необходимости семантических эмбеддингов можно использовать **ruBERT**, **RuT5** или модели **fastText** – они помогут при расчете когерентности, тематических признаков и др. Для расчета сложных стилистических метрик (метафоры, эмоциональные арки) могут потребоваться сочетания инструментов и собственных скриптов, однако многие исследования указывают, что такие формальные показатели реально уловить количественно ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels" \l ":~:text=websites%20or%20social%20networks,)) ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels" \l ":~:text=writing%20style%20to%20predict%20the,)).

**2. Метрики качества моделей и оценки связи признаков с реакциями пользователей**

Оценивать успешность модели, предсказывающей *восприятие текста* (через прокси-рейтинги вроде пользовательских оценок или лайков), можно с помощью метрик, отражающих корреляцию и ранжирование, а также метрик из области оценки генеративных моделей и интерпретируемости. В условиях отсутствия «золотого стандарта» (ручной разметки экспертов) акцент делается на согласованности модели с наблюдаемыми пользовательскими сигналами.

**Ранговые и корреляционные метрики**

* **Спирменовская корреляция ρ**: метрика ранговой корреляции Spearman, отражающая монотонную связь между предсказанными моделью баллами и реальными пользовательскими рейтингами (например, средними оценками или коэффициентами лайков). Спирмен ρ особенно полезен, когда нас интересует порядок текстов по качеству, а не точные значения. Высокий ρ означает, что модель хорошо упорядочивает тексты аналогично читательским оценкам.
* **Кендалловская τ**: другой коэффициент ранговой корреляции (Kendall Tau), более строгий при учете перестановок в ранге. Подходит для оценки согласованности упорядочивания, особенно на данных с множеством одинаковых оценок или при сравнениях парного предпочтения.
* **NDCG (Normalized Discounted Cumulative Gain)**: метрика из информационного поиска, применимая, если мы рассматриваем задачу как ранжирование произведений по оценке. NDCG учитывает позицию релевантных (высоко оцененных) текстов в выдаче: штрафует, если высокорейтинговый текст оказался ниже в списке. Она особенно полезна, когда важно качество топ-N рекомендаций (например, насколько в топ-10 по предсказаниям модели действительно находятся лучшие по пользовательским рейтингам произведения). В контексте оценки литературных текстов NDCG может показать, насколько хорошо модель «поднимает» наверх действительно популярные тексты.
* **RMSE/MSE и MAE**: метрики среднеквадратичной или среднеабсолютной ошибки между предсказанным скалярным рейтингом и реальным. Их можно применять, если модель выдаёт конкретный числовой балл (например, предсказывает среднюю оценку 0–5). Однако на субъективных данных эти метрики труднее интерпретировать – важно не столько точное совпадение, сколько относительный порядок и тенденции. Тем не менее, снижение MSE при добавлении признаков покажет, что модель лучше подгоняется под прокси-оценки.

**Метрики для субъективных суждений и генеративных моделей**

* **BLEURT**: метрика на основе трансформеров от Google. Обычно применяется для оценки качества генерируемого текста по отношению к эталону, учитывая семантику и нюансы. В нашем случае нет прямого «эталонного» текста, но BLEURT можно задействовать косвенно – например, для оценки сходства рецензий или аннотаций модели с пользовательскими отзывами. Кроме того, BLEURT обучен предсказывать человеческую оценку качества текстов, поэтому теоретически может использоваться как дополнительный сигнал качества текста, особенно если модель генерирует какие-то пояснения.
* **BERTScore**: метрика, сравнивающая два текста путем вычисления сходства между эмбеддингами слов (обычно берутся предобученные BERT). Она хорошо коррелирует с человеческими оценками при сравнении текста с образцом. В применении к нашей задаче BERTScore мог бы применяться, например, чтобы сравнить стиль оцениваемого текста со стилем **эталонных высококачественных текстов**. Если собрать корпус классической литературы как образцовой, можно измерять близость нового текста к этому корпусу (в смысле использования лексики и выражений) – некий показатель «литературности». Однако прямо метрика BERTScore скорее вспомогательная.
* **MAUVE**: относительно новая метрика для оценки сгенерированного текста, которая измеряет расстояние между распределением генерированных текстов и реальных (человеческих) текстов ([MAUVE - a Hugging Face Space by evaluate-metric](https://huggingface.co/spaces/evaluate-metric/mauve#:~:text=MAUVE%20,distribution%20of%20human%20text%2C)) ([README.md · evaluate-metric/mauve at main - Hugging Face](https://huggingface.co/spaces/evaluate-metric/mauve/blob/main/README.md#:~:text=MAUVE%20was%20found%20to%20correlate,Metric%20Card%20for%20MAUVE)). MAUVE учитывает разнообразие и правдоподобие текста. В контексте оценки качества произведений можно мыслить так: если у нас есть распределение признаков **у популярных текстов** vs. **непопулярных**, MAUVE могла бы показать, насколько наша модель выдает распределение ближе к «успешным» текстам. На практике же MAUVE упоминается как пример метрики, которую можно использовать, если мы генерируем какие-то предсказания в текстовой форме либо сравниваем стиль целой группы текстов с эталонной группой.
* **Метрики когнитивного воздействия**: существуют специализированные метрики типа **Hauge, surprise, etc.**, но они редко используются вне академических работ. Можно упомянуть, что оценка литературного качества связана с субъективным восприятием, поэтому разрабатываются и метрики вроде *перепадов эмоций* или *увлеченности*. Например, индекс увлеченности чтением (может быть вычислен через анализ отзывов: сколько раз читатели отметили «захватило»). Однако для автоматической системы эти метрики скорее являются результатом, чем инструментом оценки.

**Интерпретируемость и важность признаков (explainable metrics)**

* **SHAP (SHapley Additive exPlanations)**: метод вычисления вклада каждого признака в предсказание модели на основе концепции значений Шэпли из теории игр. SHAP позволяет объяснить, **почему** модель присвоила тексту определенный оценочный балл, выделяя, какие признаки повысили, а какие понизили оценку ([An Introduction to SHAP Values and Machine Learning Interpretability](https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability" \l ":~:text=An%20Introduction%20to%20SHAP%20Values,feature%20impacts%20the%20model%27s%20prediction)) ([An introduction to explainable AI with Shapley values](https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html" \l ":~:text=This%20tutorial%20is%20designed%20to,explanations%20of%20machine%20learning%20models)). В нашем случае, используя SHAP, можно для каждого произведения показать, что, например, «высокая доля диалогов и высокая эмоциональная вариативность повысили прогнозируемый рейтинг, а простота синтаксиса снизила». Это важно для доверия к системе. *Инструменты*: библиотека shap (Python) поддерживает многие модели (деревья, нейросети, линейные модели) и выдает как локальные объяснения для отдельных предсказаний, так и глобальную важность признаков.
* **Permutation Importance (перестановочная важность)**: простой и наглядный способ оценить значимость признака – случайным образом перемешать значения этого признака по выборке и посмотреть, насколько ухудшится качество модели ([23 Permutation Feature Importance – Interpretable Machine Learning](https://christophm.github.io/interpretable-ml-book/feature-importance.html#:~:text=23%20Permutation%20Feature%20Importance%20%E2%80%93,feature%2C%20which%20breaks%20the)). Сильное падение качества означает, что признак был важен. Этот метод помогает в отсутствии разметки понять, какие извлекаемые признаки наиболее влияют на предсказания (и, соответственно, на пользовательские реакции). *Инструменты*: для моделей sklearn есть готовая функция permutation\_importance, а для других можно реализовать вручную, пересчитав метрику качества на испорченных данных.
* **Partial Dependence Plots и ICE**: графики частичной зависимости показывают, как меняется предсказанный рейтинг при изменении одного признака, усредненно по всем объектам. Individual Conditional Expectation (ICE) – то же на уровне отдельных текстов. Эти методы позволяют выявить нелинейные эффекты: например, увидеть, что при увеличении длины текста от 0 до ~10000 знаков рейтинг растет, а дальше — плато. Такие визуализации дополняют количественные важности и делают модель прозрачнее для исследователя.
* **Каппа согласия и agreement-метрики**: если модель оценивания сравнивается с человеческими оценками (даже прокси), иногда вычисляют метрики согласия типа Cohen’s Kappa (для классификации на «хороший/плохой») или Krippendorff’s alpha для непрерывных шкал. Они показывают, насколько модель «согласна» с пользователями в рамках погрешности межчеловеческого разброса. В отсутствии абсолютной разметки можно трактовать разных пользователей или разные сайты как разные аннотаторы и стремиться к высокому соглашению.

**Оценка по прокси-сигналам вовлечённости**

* **Precision/Recall@K по вовлеченности**: Если брать метрику «лайки/просмотры» как целевую, можно сформировать бинарную задачу: топ X% текстов по этому коэффициенту считать «хорошими». Тогда оцениваем модель как классификатор: например, точность предсказания топ-10% (Precision@10%) – доля действительно популярных текстов среди тех, что модель отнесла к топу. Recall@K – покрытие популярных текстов среди рекомендаций модели.
* **ROC AUC**: тоже в случае бинаризации задачи (например, популярное vs непопулярное произведение по порогу лайков). AUC покажет способность модели отличать успешные тексты от неуспешных по всем порогам. Она удобна, когда не хотим выбирать конкретный порог, а смотрим общую дискриминацию.
* **Pairwise Accuracy**: метрика парных сравнений – случайным образом берутся две разных книги, и модель «угадывает», какая из них получила более высокий пользовательский рейтинг. Доля правильных угадываний по большому числу пар дает представление, насколько модель в среднем сохраняет относительный порядок. Это связано с ранговыми метриками, но может быть более наглядно при коммуникации результата.
* **Корреляция по группам**: можно биновать тексты по реальным оценкам (например, низкие, средние, высокие рейтинги) и проверять средние предсказания модели для этих групп. Правильная тенденция: у группы с высокими реальными рейтингами должна быть заметно выше средняя предсказанная оценка, чем у группы с низкими. Этот подход не дает одного числа, но демонстрирует, что модель реально различает уровни успеха.
* **Метрики вовлеченности с временны́м компонентом**: если доступны временные ряды просмотров/лайков (например, скорость набора популярности), то модель можно проверить на умение предсказать не только статичный рейтинг, но и динамику. Например, корреляция между предсказанным качеством и реальной скоростью роста аудитории. Однако такие данные выходят за рамки статичного текстового анализа.

В условиях отсутствия экспертной разметки перечисленные метрики позволяют косвенно оценить качество модели. Ранговые метрики (Spearman, Kendall) непосредственно измеряют попадание в порядок предпочтения читателей. Метрики вроде NDCG пригодятся, если задача – рекомендовать лучшие тексты на вершину. А методы интерпретации (SHAP, permutation importance) не столько оценивают числовое качество, сколько усиливают доверие, показывая, что модель опирается на осмысленные признаки (например, действительно учитывает связность и эмоциональность, а не случайные слова).

**3. Базовые модели для сравнения (Baseline)**

При построении своей системы оценки качества текста важно сравнивать ее с простыми базовыми моделями, чтобы убедиться, что сложные признаки действительно дают выигрыш. Возможные бейзлайны:

* **TF-IDF + линейная регрессия/классификатор**: один из самых простых подходов – представить каждый текст как вектор TF-IDF (частотность слов с уменьшением веса частых слов) и обучить линейную модель (например, линейную регрессию для прогнозирования рейтинга или логистическую регрессию для классификации «выше среднего/ниже среднего»). Такой подход игнорирует порядок слов, учитывает только частотный профиль. Тем не менее, он способен улавливать тематические корреляции (например, слова *“любовь”, “отношения”* могут коррелировать с жанром романтики и определенными оценками). В исследованиях подобные методы показывали определенную предсказательную силу: например, модель на bag-of-words + логистическая регрессия отличала бестселлеры от небестселлеров с точностью ~75% ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels#:~:text=dataset%20containing%20,relative%20success%20of%20a%20literary)).
* **Простые статистические признаки + регрессия**: еще более простой baseline – взять несколько элементарных численных характеристик текста и обучить на них модель. Например, признаки: длина текста (число слов/символов), средняя длина предложения, доля уникальных слов, средняя частота слов (по языковому корпусу). Линейная регрессия или решающее дерево на таких признаках даст точку отсчета. Этот подход проверяет, не объясняется ли вариативность рейтингов банальными факторами вроде размера текста или примитивной лексической разнообразности. Если такая модель дает, скажем, Spearman ρ = 0.3, то новая модель должна превосходить это значение.
* **Лексические шаблоны (rule-based)**: можно заложить руками несколько правил на основе интуиции и смотреть, насколько они совпадают с реальностью. Например: *“Если в тексте >50% диалогов, снизить оценку качества на 1, если средняя длина слова >6 букв, повысить на 1”* – условно, вручную настроенная формула рейтинга. Это утрированно, но как baseline может выступать даже **константная модель** (всем текстам присваивает средний рейтинг) или **слегка параметризованная формула**. Сравнение с такой моделью покажет, насколько ML-модель превосходит наивный подход.  
  *Пример шаблона*: score = 0.5 \* (уникальные\_слова/общие\_слова) + 0.5 \* (средняя\_длина\_предложения) – подобная линейная комбинация двух признаков могла бы служить эталоном, с которым соревнуется сложная модель.
* **Модель на предобученных эмбеддингах**: здесь практически не используется ручной инженерии признаков – каждый текст превращается в вектор фиксированной размерности с помощью предобученной модели. Варианты: усреднить word2vec/fastText эмбеддинги всех слов текста; или взять **[CLS]**-эмбеддинг из RuBERT для всего текста; или использовать Doc2Vec. Затем обучить простой алгоритм (например, тот же Random Forest или линейную регрессию) предсказывать рейтинг по этому вектору. Смысл baseline: проверить, не справится ли **сам по себе** языковой модель/эмбеддинг с задачей предсказания популярности. Если такой подход работает почти так же хорошо, как ваш сложный набор признаков, возможно, часть признаков дублирует информацию, уже содержащуюся в эмбеддинге.
* **“Zero-shot” на основе больших моделей**: в некоторых случаях можно попробовать использовать большие предобученные модели напрямую. Например, взять сверхбольшую модель типа **ruGPT-3** (или GPT-3 через API) и попросить её оценить текст по качеству (zero-shot learning, без обучения на ваших данных). Конечно, это скорее экспериментальный baseline: сравнить, насколько согласуется такая «оценка от GPT» с реальными рейтингами. Вероятно, она будет шумной, но если неожиданно окажется высокой – это важный ориентир.
* **Случайная модель**: как тривиальный нижний порог можно определить качество случайного ранжирования. Например, перетасовать реальные рейтинги и вычислить Spearman – он будет около 0, или NDCG ~ 0.5 для топов. Все модели должны превосходить случайную. Аналогично, **константный прогноз** (всем одно значение) даст нулевую корреляцию и определенный MSE (равный дисперсии рейтингов), это тоже порог для регрессии.

Бейзлайны важны для валидиции: если сложная модель даёт улучшение всего на 0.01 по Spearman против TF-IDF, надо задуматься о пользе сложных признаков. В то же время, сильный baseline (например, fine-tuned BERT) ставит высокую планку, и превосходя его, вы показываете ценность интерпретируемых признаков. В литературе отмечалось, что **простейшие признаки оказываются весьма эффективными** в задачах стилометрии и качества ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels" \l ":~:text=Stamatatos%2C%202009%3B%20Ashok%20et%20al,)) ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels" \l ":~:text=writing%20style%20to%20predict%20the,)), поэтому сравнение с ними обязательно.

**4. Примеры исследований и систем без экспертной разметки**

Ниже приведены несколько известных работ и проектов, где для оценки литературных текстов применялись формализованные признаки и *прокси*-оценки (читательские рейтинги, лайки, показатели вовлечённости) – без прямой экспертной разметки:

* **Ashok et al. (EMNLP 2013)** – *“Success with Style: Using Writing Style to Predict the Success of Novels”*. В этой работе на английских текстах авторы предсказывали **успех книги** (измеряемый числом скачиваний с Project Gutenberg) по стилевым характеристикам ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels#:~:text=websites%20or%20social%20networks,)) ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels#:~:text=,in%20stylometry%2C%20genre%20discrimination%20and)). Были использованы сотни признаков: от распределения частей речи и грамматических правил до тональности и узнаваемости имен. Модель SVM смогла отличить популярные книги от непопулярных лучше случайного. Интересные выводы: *успешные романы характеризовались более высоким употреблением существительных и местоимений при более низком употреблении глаголов* ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels" \l ":~:text=lower%20proportion%20of%20verbs,focused%20less%20on%20the%20actions)), а также *более сложным языком и худшей читаемостью, что парадоксально соответствовало высокой концептуальной сложности* ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels" \l ":~:text=literary%20works%2C%20Ashok%20et%20al,)). Эта работа показала принципиальную возможность предсказать читательскую популярность через формальные стилевые маркеры.
* **Maharjan et al. (2018)** – *“Letting Emotions Flow: Success Prediction by Modeling the Flow of Emotions in Books”*. Авторы исследовали связь **эмоционального сюжетного рисунка** с оценкой книги читателями. Они моделировали эмоции персонажей или тональность повествования по главам и использовали характеристики *эмоционального потока* для предсказания рейтингов книг. Выяснилось, что определенные паттерны эмоциональных взлетов и падений коррелируют с популярностью ([(PDF) Letting Emotions Flow: Success Prediction by Modeling the Flow of Emotions in Books](https://www.researchgate.net/publication/325445735_Letting_Emotions_Flow_Success_Prediction_by_Modeling_the_Flow_of_Emotions_in_Books#:~:text=with%20human%20interpretations%20of%20a%E6%98%80%E6%98%80ective,)). Проще говоря, книги, вызывающие эмоции по нарастающей или чередующие контрастные эмоции, удерживают внимание и получают выше оценки. Эта работа подчеркнула важность динамических семантических признаков (не только “статического” среднего тонального баланса).
* **Каноничная vs массовая литература (различия)**: Несколько исследований в цифровой гуманитаристике на материале литературных канонов. Например, работа Bizzoni et al. (2022) исследовала, отличаются ли **канонизированные произведения** от остальных с точки зрения статистической непредсказуемости текста (). Используя большие языковые модели, измеряли **перплексию** и **энтропию** текста: гипотеза была, что классика менее шаблонна и более непредсказуема. Результаты показали, что действительно у признанных литературных текстов перплексия выше, а также меньше клишированных фраз (). Это пример, как без ручной разметки («канон» определили списком известных книг) метрика на основе LM-сюрприза помогает количественно уловить качество.
* **van Cranenburgh & Koolen (2020)** – эксперимент *“Literary Taste Test”*. Исследователи провели уникальный эксперимент: брали отрывки из романов, **анонимно** предъявляли читателям и собирали оценки «понравилось/не понравилось». Цель – выяснить, насколько объективно качество текста считывается без знания автора/названия. Затем пытались предсказать эти оценки с помощью автоматического анализа. Они использовали стилистические и синтаксические признаки (в т.ч. из работ van Cranenburgh et al. 2019 по синтаксической сложности) и получили умеренные, но значимые корреляции с читательскими предпочтениями. Важно, что это делалось **без экспертной разметки** – вкусы самих читателей служили обучающими сигналами.
* **Предсказание популярности фанфиков и онлайн-историй**: В прикладных системах на платформах вроде Wattpad, FanFiction.net и российских Самиздат-площадках тоже предпринимались попытки ранжировать тексты по ожидаемой популярности. Например, Frey et al. (2013) анализировали, какие лингвистические черты связаны с числом лайков/комментариев к онлайн-статьям ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels" \l ":~:text=quantifiable%20quality%20measures%20like%20number,)). Arakawa et al. (2014) изучали японские новеллы в интернете, предсказывая число отзывов по стилю ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels#:~:text=match%20at%20L531%20quantifiable%20quality,)). Для русского сегмента можно отметить прикладные проекты в сфере маркетинга контента: некоторые компании делали сервисы оценки **вовлеченности постов** по тексту (учитывая символы эмодзи, восклицания, длину поста – схожие с литературой признаки, но приложенные к соцсетям) ([Success with style: Using writing style to predict the success of novels | Request PDF](https://www.researchgate.net/publication/286941537_Success_with_style_Using_writing_style_to_predict_the_success_of_novels" \l ":~:text=quantifiable%20quality%20measures%20like%20number,)). Эти системы показывают, что подход «текстовые признаки + пользовательские реакции» успешно работает вне академических задач.
* **Статистический анализ русской классики**: Кислицына и Орлов (2018) провели масштабный статистический анализ корпуса русской художественной литературы. Хотя цель была авторское стилевое своеобразие и автоматическое распознавание авторов, побочным результатом стало подтверждение, что формальные признаки (распределения слов, длины предложений, частотные словари) устойчиво различают тексты по качественным категориян. Например, тексты разных эпох и жанров образуют кластеры в пространстве частотных признаков. Это наводит на мысль, что **восприятие качества** читателями (часто связанное с принадлежностью к определенному стилю или эпохе) тоже может фиксироваться такими признаками.
* **Проект “Textometr” (2021)**: Хотя он посвящен оценке сложности текста для изучающих язык, а не литературному качеству, заслуживает упоминания. Это онлайн-инструмент, использующий регрессионную модель и >800 текстов с известным уровнем (A1–C2), чтобы предсказать уровень сложности нового текста ([Textometr: an online tool for automated complexity level assessment of texts for Russian language learners - Laposhina - Russian Language Studies](https://journals.rudn.ru/russian-language-studies/article/view/27498/en_US" \l ":~:text=Evaluation%20of%20text%20accessibility%20seems,on%20the%20dataset%20of%20more)) ([Textometr: an online tool for automated complexity level assessment of texts for Russian language learners - Laposhina - Russian Language Studies](https://journals.rudn.ru/russian-language-studies/article/view/27498/en_US" \l ":~:text=learning%20and%20natural%20language%20processing,not%20only%20to%20teachers%20and)). Он выдает набор характеристик: от доли слов по частотным спискам до предположительного времени чтения. Этот пример показывает, как **без прямой разметки качества** (уровень языка – тоже прокси) можно построить интерпретируемую модель с полезными фичами. Для нашей задачи многие признаки пересекаются (лексические богатство, сложность синтаксиса), хотя цель иная.

**Вывод:** Современные исследования подтверждают возможность автоматически оценивать и даже предсказывать *литературное качество и успех* текста, опираясь только на открытые данные – сами тексты и показатели вовлеченности читателей. Комбинация разнообразных признаков (лексических, синтаксических, семантических, стилистических и пр.) вместе с метриками ранжирования дает воспроизводимую систему, чьи решения можно объяснить через значимые языковые особенности. Такой подход уже применялся для английских и других языков, и адаптация его на русский язык реализуема благодаря наличию мощных NLP-инструментов (Natasha, DeepPavlov, ruBERT и др.) и корпусов русской литературы. Главное – обеспечить грамотную валидацию на прокси-метриках и сравнение с простыми базовыми моделями, чтобы новая система действительно давала ценную информацию о тексте, соотносящуюся с мнением читателей.

Источники

<https://ru.wiktionary.org/wiki/Приложение:Список_частотности_по_НКРЯ/101—1000>