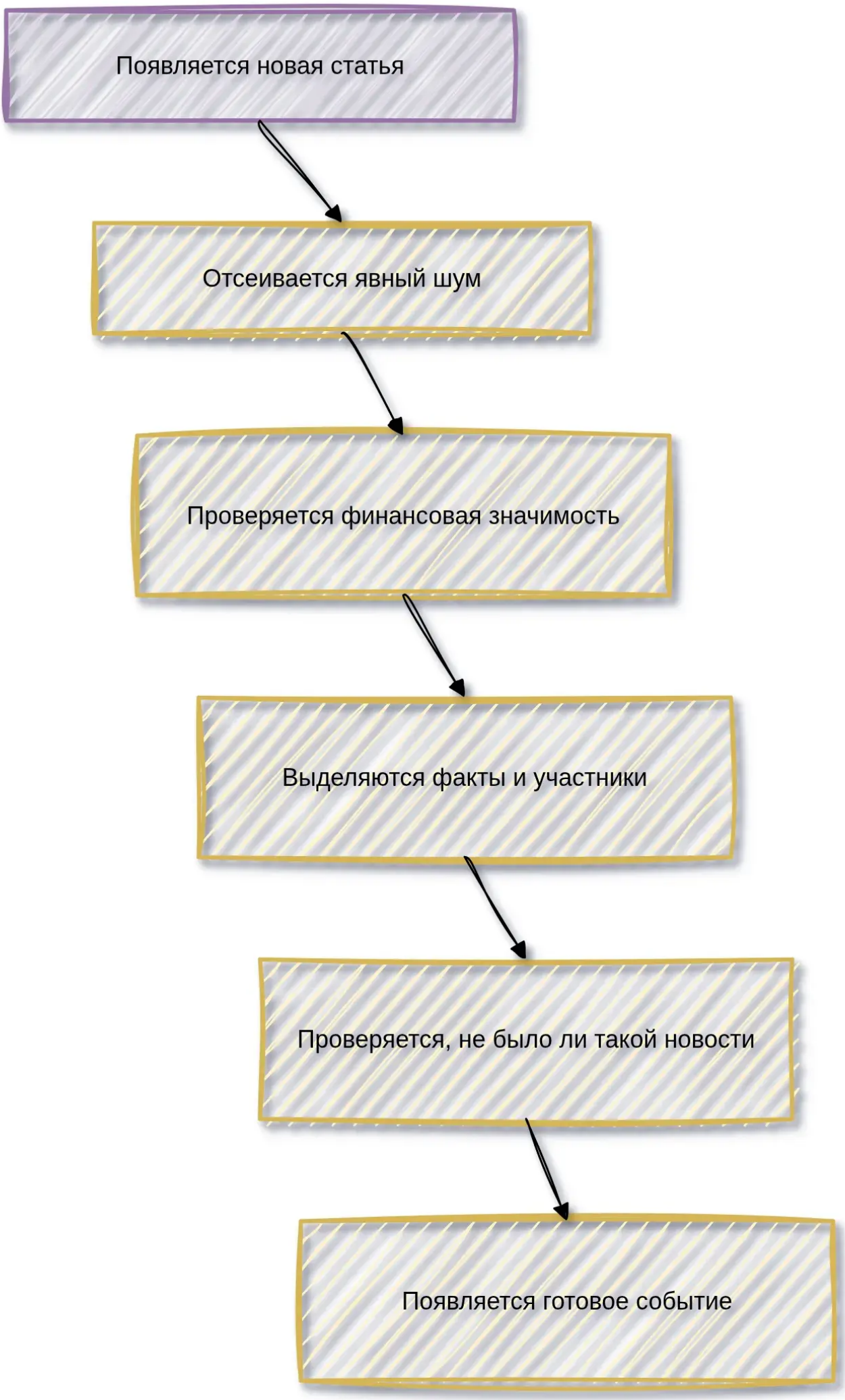
AI-анализатор финансовых новостей

Зачем он нужен

Анализатор новостей помогает превращать обычный поток статей в понятные финансовые события TokenBel. Он работает в фоне: пользователь не заполняет форму и не нажимает кнопки, а видит уже подготовленный результат в новостном разделе.

Задача анализатора — отделить финансово значимые материалы от шума, кратко объяснить событие, выделить связанные организации и инструменты, убрать дубли и сохранить аккуратную запись для дальнейшего просмотра.


Появляется новая статья
Отсеивается явный шум
Проверяется финансовая значимость
Выделяются факты и участники
Проверяется, не было ли такой новости
Появляется готовое событие

Место в новостном процессе

Анализатор включается после того, как новая статья уже найдена и сохранена как исходный материал. Он не ищет новости самостоятельно, а работает с теми статьями, которые поступили в новостной поток.

Для пользователя это означает простую цепочку:

- сначала появляется исходная публикация;
- затем система решает, относится ли она к тематике TokenBel;
- если относится, из неё формируется финансовое событие;
- если похожее событие уже есть, новая статья может быть присоединена к нему как дополнительный источник;
- если статья не подходит, она не попадает в готовый новостной поток.

Первичная очистка статьи

Перед смысловой проверкой статья приводится к более чистому виду. Из текста убираются лишние фрагменты, повторы, рекламные и навигационные вставки, пустые строки и другие элементы, которые мешают понять суть.

Этот этап нужен не для изменения смысла, а для того, чтобы дальнейшая проверка работала с содержанием статьи, а не с техническим или редакционным шумом вокруг неё.

Что может остановить обработку сразу

Статья может быть отброшена ещё до глубокого анализа, если она явно не годится для новостного события:

- текст слишком короткий и не содержит достаточной информации;
- заголовок явно относится к бытовым темам вроде погоды, гороскопов, спорта, кино, рецептов, афиш или авторынка.

Отсутствие очевидных финансовых слов само по себе не останавливает обработку. Если статья может быть важной, но сформулирована неочевидно, она проходит дальше.

Проверка финансовой значимости

После первичной очистки система оценивает, подходит ли статья для TokenBel. Статья должна быть связана с инвестиционными или сберегательными инструментами,

эмитентами, торговыми площадками, рынком, выплатами, регулированием или другими темами, которые помогают пользователю понимать финансовые события.

К глубокому разбору проходят материалы, где одновременно видно:

- статья действительно относится к финансовой тематике TokenBel;
- в ней есть достаточно конкретики для отдельного события;
- тема попадает хотя бы в одну подходящую категорию;
- степень уверенности достаточно высокая.

Если статья не проходит эту проверку, она спокойно пропускается. Это нормальный исход, а не ошибка.

Какие темы считаются подходящими

К подходящим темам относятся не только акции, облигации и токены. Анализатор также учитывает:

- эмиссии, размещения, изменения условий выпусков и документов;
- новости эмитентов и важные корпоративные события;
- купоны, дивиденды, погашения и другие выплаты;
- дефолтные риски, реструктуризации и ухудшение обязательств;
- налоги и регулирование, если они затрагивают инструменты или участников рынка;
- торговую инфраструктуру, площадки, правила торгов и расчётов;
- банковские вклады, ставки, сроки и доходность;
- драгоценные металлы как сберегательные или инвестиционные активы;
- валютный рынок Беларуси, курсы, динамику BYN и рыночные обзоры с конкретными фактами.

Важное правило: слово «токен» в белорусском инвестиционном контексте не считается криптовалютой автоматически. Если речь идёт о токенизированном инструменте, размещении, декларации или White paper, такая новость может быть релевантной для TokenBel.

Извлечение фактов

Для релевантной статьи система выделяет фактическую основу события. На этом этапе собираются:

- главный факт;
- дополнительные факты с оценкой уверенности;
- суммы, проценты, даты, сроки, цены, ставки и идентификаторы;
- упомянутые организации, инструменты, площадки, регуляторы, банки, валюты, рынки и другие объекты;

- возможные типы события и теги;
- замечания о неясностях, если в статье не хватает данных или формулировки неоднозначны.

Эти материалы не показываются пользователю как отдельная черновая карточка. Они помогают собрать более точное и осторожное итоговое событие.

Создание события

После выделения фактов формируется готовое новостное событие. Оно описывает не саму статью, а финансовый смысл того, что произошло.

В событие входят:

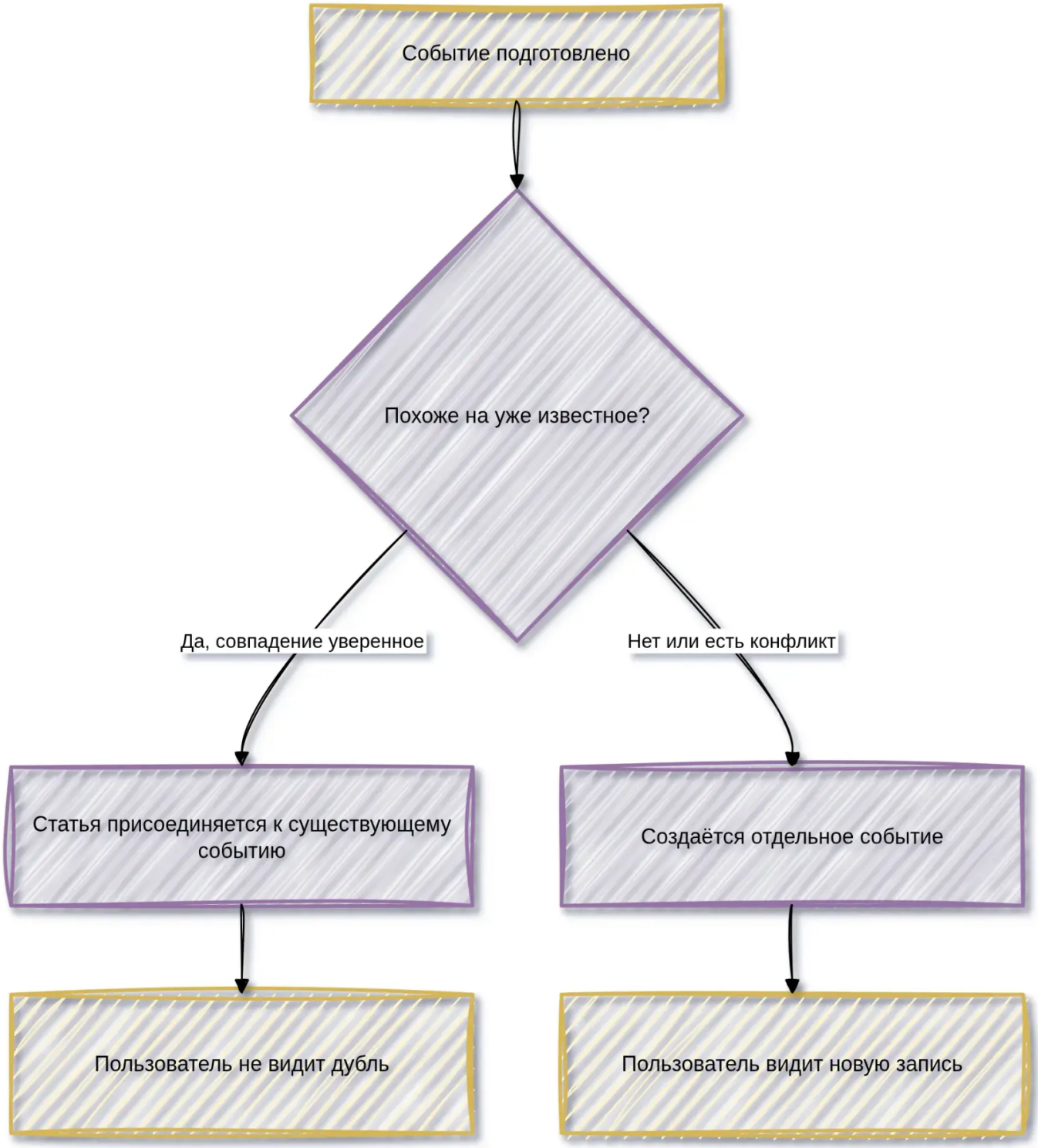
- короткий заголовок;
- подробное, но сжатое описание;
- тип события — например выплата, эмиссия, изменение регулирования, событие по токену, валютный рынок или депозитная новость;
- характер влияния — позитивный, негативный, нейтральный или смешанный;
- оценки важности и уверенности;
- теги;
- упомянутые сущности;
- ключевые факты;
- предупреждения, если часть информации неполная или требует осторожного чтения.

Текст события должен быть нейтральным: без советов, рекламных формулировок, эмоциональных оценок и выводов, которых нет в статье.

Проверка на дубликаты

Перед сохранением нового события система проверяет, не было ли уже очень похожей новости. Это важно, потому что один и тот же факт может появиться в нескольких источниках или повториться в близких формулировках.

Проверка учитывает не только заголовок. Сравниваются смысл новости, тип события, участники, даты, факты, значения и близость по времени.



Если совпадение уверенное и не содержит противоречий, новая статья не создаёт дубль. Она связывается с уже существующим событием как дополнительный источник. Если совпадение слабое, спорное или конфликтное, система не объединяет новости автоматически и создаёт отдельное событие.

Редакторская вычитка

Когда событие не оказалось дублем, его текст проходит дополнительную редакторскую вычитку. Этот этап улучшает только читаемость: заголовок, описание, ключевые факты и предупреждения.

Структурные данные при этом не меняются. Тип события, характер влияния, оценки, теги и список упомянутых сущностей остаются теми, что были выбраны на этапе создания события. Это защищает результат от случайного изменения смысла при стилистической правке.

Что видит пользователь

Пользователь видит не весь внутренний путь статьи, а итоговую новостную запись. Она помогает быстро понять:

- что произошло;
- с каким инструментом, компанией, рынком или регулятором это связано;
- какой тип события перед пользователем;
- какие факты важнее всего;
- есть ли предупреждения о неполных или неоднозначных данных;
- не является ли статья повтором уже известного события.

Если статья была только дополнительным источником к уже существующему событию, пользователь получает более чистый поток без лишних дублей.

Что происходит с неподходящими и проблемными статьями

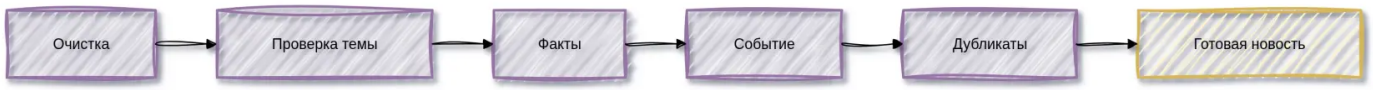
Не каждая статья становится событием. Возможны разные исходы:

- **Явно бытовая или слишком короткая статья** — не идёт в глубокий разбор.
- **Недостаточная финансовая значимость** — статья пропускается без ошибки.
- **Похожая уже известная новость** — статья присоединяется к существующему событию, если совпадение уверенное.
- **Временный сбой** — обработка повторяется позже.
- **Постоянная проблема** — например пустой текст, недоступная статья или некорректный результат проверки — фиксируется в служебной истории, чтобы не повторяться бесконечно.

Главная идея: временным проблемам даётся повторная попытка, а нерелевантные или неисправимые случаи аккуратно завершаются.

Как связаны этапы

Каждый этап сужает поток и добавляет точность. Сначала убирается очевидный шум, затем проверяется смысловая релевантность, потом выделяются факты, создаётся событие, проверяются дубликаты и улучшается читаемость текста.



Такой порядок помогает не тратить глубокий анализ на неподходящие статьи и одновременно не терять короткие, но важные сообщения о токенах, выплатах, эмиссиях, вкладах, драгоценных металлах или валютном рынке.

Техническая схема слоёв

Внутри анализатор устроен как последовательность слоёв. Каждый слой получает результат предыдущего, делает только свою часть работы и передаёт дальше уже более структурированные данные. Это важно для управляемости: ранние слои дешёво убирают мусор, средние слои извлекают смысл, а поздние слои отвечают за дубликаты, читаемость и сохранение.

Слой не должен делать всё сразу. У каждого слоя есть свои «права»: что он может менять, где может остановить обработку и какие решения ему запрещено принимать.

Слой	Основная задача	Что разрешено	Что запрещено
Вход очереди	Получить сообщение о сохранённой сырой статье и загрузить её из backend.	Проверить версию сообщения, токены доступа и наличие статьи. Повторить обработку при временной ошибке.	Самостоятельно искать новости или менять смысл статьи.
Layer 0: Cleaner	Очистить заголовок и текст от технического шума.	Нормализовать символы, пробелы, кавычки, тире, убрать emoji, пустые и служебные строки, зафиксировать диагностику очистки.	Решать, финансовая статья или нет; создавать событие; обращаться к AI.
Layer 0.5: Prefilter	Быстро остановить очевидно неподходящие материалы до расходов на AI.	Остановить слишком короткий текст или заголовки про явно бытовые темы. Найти финансовые ключевые слова как подсказку.	Отбрасывать статью только потому, что финансовые слова не найдены: спорные случаи должны идти в Layer 1.

Слой	Основная задача	Что разрешено	Что запрещено
Layer 1: Screening	Оценить релевантность TokenBel через Mistral.	Вернуть решение: пропустить статью или отправить в глубокий разбор. Учитывать категории, уверенность и правило, что белорусский инвестиционный «токен» не равен криптовалюте автоматически.	Формировать итоговое событие, сохранять данные или объединять дубликаты.
Layer 2.1: Fact Distillation	Вытащить фактическую основу статьи.	Собрать главный факт, дополнительные факты, суммы, проценты, даты, ставки, идентификаторы, организации, инструменты, площадки и неясности.	Писать финальный заголовок события или выбирать окончательные структурные поля.
Layer 2.2: Event Builder	Собрать каноническое новостное событие.	Определить заголовок, описание, тип события, характер влияния, важность, уверенность, теги, сущности, ключевые факты и предупреждения.	Проверять дубликаты, сохранять запись или позже позволять редакторскому слою менять структурные поля.
Embedding	Подготовить смысловой вектор события для поиска дублей.	Построить стабильное представление события из фактов, сущностей, тегов и дат; отправить его в модель эмбеддингов.	Использовать сырой HTML, рекламу и служебный шум; принимать редакторские решения.
Dedupe Lookup	Проверить, не описывает ли статья уже известное событие.	Спросить backend о похожих событиях, учитывать семантическую близость, заголовок, сущности, время и конфликты. При уверенном совпадении присоединить статью как источник и остановиться.	Автоматически объединять спорные или конфликтные новости; при неудачном присоединении создать дубль.
Layer 2.3: Humanizer	Улучшить читаемость события перед сохранением.	Полировать только текст: заголовок, описание, ключевые факты и предупреждения.	Менять тип события, влияние, важность, уверенность, теги и список сущностей. Эти поля всегда остаются из Layer 2.2.

На практике это даёт несколько защитных правил:

- ранние слои могут остановить только очевидный мусор, но не должны терять потенциально важную финансовую новость;

- Layer 2.2 является источником правды для структуры события;
- Layer 2.3 улучшает язык, но не имеет права менять смысловую классификацию;
- проверка дублей происходит до финальной вычитки, чтобы не тратить редакторский слой на статью, которая будет просто присоединена к уже существующему событию;
- сохранение происходит только в конце, когда событие прошло очистку, релевантность, факты, сборку, проверку дублей и вычитку.

Краткий глоссарий

- **Сырая статья** — исходный новостной текст, который ещё не прошёл очистку и смысловой разбор.
- **Новостное событие** — готовая запись о финансовом факте: что произошло, с кем связано и почему это относится к TokenBel.
- **Релевантность** — соответствие статьи тематике инвестиционных и сберегательных инструментов TokenBel.
- **Тег** — короткая тематическая метка, которая помогает сгруппировать похожие новости.
- **Упомянутая сущность** — организация, инструмент, площадка, регулятор, валюта или другой объект, который важен для понимания события.
- **Дубликат** — статья, которая описывает уже известное событие и может быть присоединена к нему как дополнительный источник.

Revision #6
Created 2026-06-15 22:30:17 UTC by Admin
Updated 2026-06-27 00:16:44 UTC by Admin