

# ИИ-анализатор финансовых новостей

## Зачем он нужен

Новостной поток TokenBel содержит много шума: статьи о погоде, спорте, кино и бытовых темах, которые не имеют отношения к рынку ценных бумаг и токенов. Анализатор превращает этот поток в чистые, структурированные финансовые события, готовые для новостного интерфейса.

Сервис работает в фоне, без участия пользователя. Как только новая статья сохраняется, он:

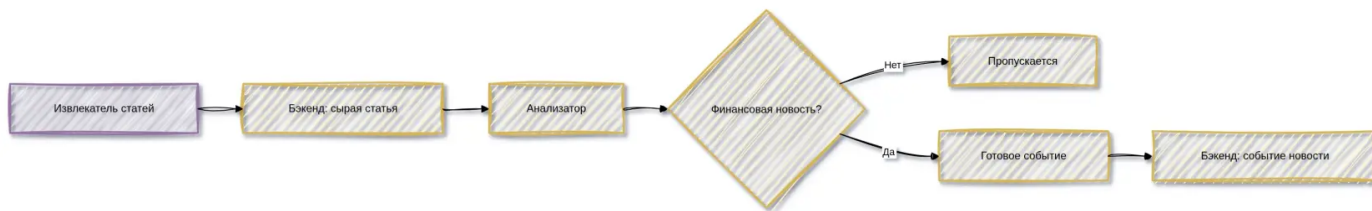
- отсеивает нерелевантные материалы ещё до обращения к модели;
- решает, относится ли статья к финансовой тематике TokenBel;
- для релевантных статей собирает готовое событие — заголовок, краткое описание, тип, влияние, теги и упомянутые организации и инструменты;
- сохраняет результат и служебный аудит качества.

В основе смыслового разбора лежит языковая модель Mistral AI. Но модель подключается не сразу: сначала работают быстрые и дешёвые проверки, чтобы не тратить ресурсы на заведомо неподходящие тексты.

---

## Где он находится в новостном процессе

Анализатор стоит между извлечением сырого текста статьи и появлением готового события в TokenBel. Он получает уведомление о новой сохранённой статье, забирает её полный текст из бэкенда, обрабатывает и записывает результат обратно.



# Как система принимает решение

Решение проходит через несколько уровней — от самого дешёвого к самому дорогому:

1. **Дешёвые детерминированные проверки** (без модели): очистить текст и быстро отсеять явный мусор.
2. **Смысловой фильтр (модель)**: достаточно ли статья релевантна финансам, чтобы тратить время на глубокий разбор.
3. **Глубокий анализ (модель, два прохода)**: сначала собрать факты и числовые значения, затем превратить их в готовое событие.

Каждый уровень может остановить обработку. Благодаря этому полный анализ достаётся только тем статьям, которые действительно его заслуживают.

## Этапы обработки статьи

### 1. Получение и проверка статьи

Система получает уведомление о новой статье и по её идентификатору забирает полный текст из бэкенда. Если текст пуст или сообщение битое, статья не уходит в обработку.

### 2. Очистка текста

Сырой текст приводится в порядок: удаляется технический мусор, лишняя разметка и пустые фрагменты. Длинные тексты укорачиваются до допустимого предела, чтобы не перегружать модель.

### 3. Быстрый фильтр шума

Перед обращением к модели система делает дешёвую проверку. Сразу отбрасываются:

- слишком короткие тексты;
- статьи с бытовыми заголовками — погода, гороскоп, рецепт, спорт (футбол, хоккей), кино и фильмы, анекдоты, анонсы мероприятий, авторынок.

Финансовые ключевые слова при этом **не обязательны**: если их нет, статья всё равно идёт дальше к модели. Фильтр намеренно мягкий, чтобы случайно не потерять важную новость с неочевидной формулировкой.

## 4. Оценка релевантности

Модель выступает в роли редактора-фильтра и отвечает на вопрос: «Стоит ли вообще разбирать эту статью?». Решение принимается по совокупности критериев. Если статья не проходит — она спокойно пропускается, без записи об ошибке.

## 5. Извлечение фактов

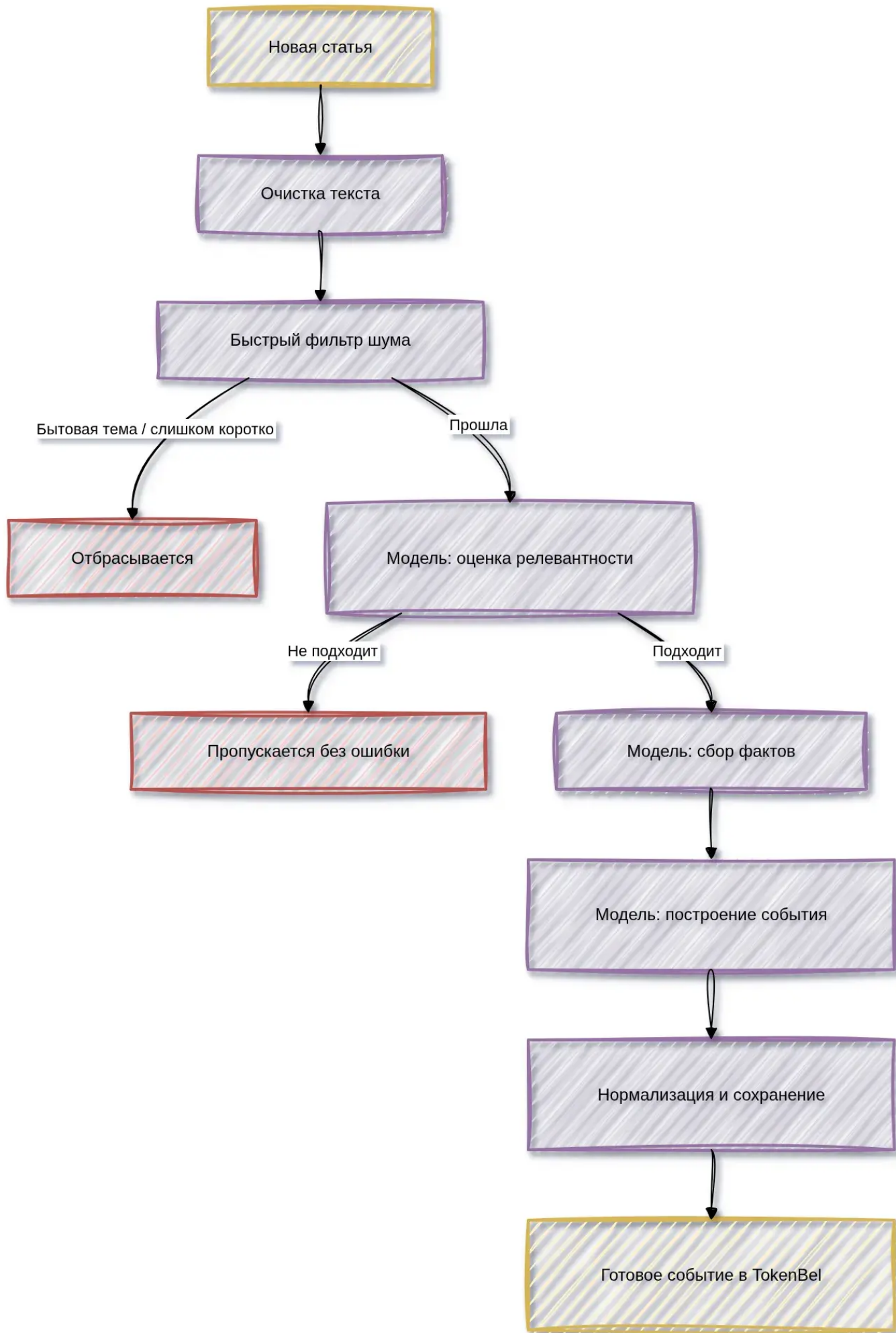
Для прошедшей статью модель отдельным проходом собирает «черновик»: факты, числовые значения, упомянутые объекты, возможные типы события и теги, а также замечания о неуверенности. Эти данные не показываются пользователю напрямую — они служат контекстом для финального шага.

## 6. Построение события

Финальный проход превращает собранные факты в готовое событие: короткий заголовок, суть в 1-2 предложениях, тип события, характер влияния, оценки важности и уверенности, теги, упомянутые сущности, ключевые факты и предупреждения.

## 7. Сохранение результата

Событие нормализуется (теги приводятся к единому виду, дубли убираются) и сохраняется в бэкенд вместе со служебным аудитом — записью о том, как модель принимала решения.



---

# Что считается релевантной финансовой новостью

Чтобы статья пошла в глубокий анализ, оценка релевантности должна одновременно:

- подтвердить, что статья действительно относится к финансовой тематике;
- выбрать режим глубокого разбора (а не поверхностного);
- набрать оценку релевантности не ниже проходного порога;
- попасть хотя бы в одну тематическую категорию.

Тематические категории охватывают широкий спектр: акции, облигации, токены, эмиссии, новости эмитентов и компаний, корпоративные действия, выплаты, дефолт-риски, налоги, регулирование, рыночная инфраструктура, вклады, драгоценные металлы и валютный рынок.

Если хотя бы одно условие не выполнено — статья считается нерелевантной и пропускается. Это нормальный, не ошибочный исход.

---

## Что сохраняется в результате

Для релевантной статьи в бэкенд записывается готовое событие новости:

- **заголовок события** — короткая формулировка;
  - **краткое описание** — суть новости в 1-2 предложениях;
  - **тип события** — одна из финансовых категорий (выплата, эмиссия, корпоративное действие и т. д.);
  - **характер влияния** — позитивное, негативное, нейтральное или смешанное;
  - **оценка важности** и **оценка уверенности** модели;
  - **теги** — нормализованные и без дублей;
  - **упомянутые сущности** — эмитенты, облигации, акции, токены, платформы, регуляторы с подсказкой типа и оценкой уверенности;
  - **ключевые факты** и **предупреждения** модели;
  - **служебный аудит** — как модель работала: версии промптов, ответы по этапам и флаги починки ответа. Аудит не виден пользователю, но нужен для проверки качества и отладки.
-

# Что происходит с нерелевантными или проблемными статьями

Поведение зависит от причины:

- **Бытовая тема или слишком короткий текст** — статья сразу отбрасывается на быстром фильтре и не уходит к модели.
- **Не прошла оценку релевантности** — статья спокойно пропускается. Это нормальный исход, не сбой: запись об ошибке не создаётся.
- **Временно недоступен бэкенд или модель** (сбой сети, таймаут, перегрузка, внутренняя ошибка сервера) — система попыбует обработать статью позже, автоматически.
- **Статья пустая, битая или явно не подлежит обработке** (например, бэкенд вернул ошибку «не найдено» или конфликт) — повторять бессмысленно, статья снимается с обработки, а проблема записывается в историю обработки для разбора.
- **Модель вернула некорректный ответ** — делается одна попытка исправить ответ; если и она не помогает, проблема фиксируется и статья снимается с обработки.

Основная логика простая: временным сбоям система даёт второй шанс, а постоянные проблемы фиксирует, чтобы они не застревали в бесконечных повторах.

---

## Почему анализ разделён на несколько шагов

Новости — это в основном шум. Если отправлять каждую статью сразу в глубокий анализ, модель будет тратить время и ресурсы на тексты, которые не стоят внимания.

Поэтому ранние проверки максимально дешёвые и быстрые, а дорогие вызовы модели подключаются только там, где нужен смысловой разбор. Модель тоже используется по ролям: сначала как редактор-фильтр («стоит ли разбирать?»), затем как исследователь («какие факты здесь есть?»), и только потом как автор финального события.

Разделение на этапы делает систему одновременно дешёвой, быстрой и точной: дешёвые отсечения закрывают основной поток, а глубокий анализ достаётся только релевантным материалам.

---

# Краткий глоссарий

- **Сырая статья** — исходный текст новости, сохранённый извлекателем до какой-либо обработки.
  - **Событие новости** — структурированная запись о финансовом событии: заголовок, описание, тип, влияние, теги и упомянутые сущности.
  - **Быстрый фильтр шума** — детерминированная проверка до модели: отбрасывает слишком короткие тексты и явные бытовые темы.
  - **Оценка релевантности** — решение модели о том, достаточно ли статья относится к финансам TokenBel для глубокого разбора.
  - **Тематическая категория** — область, к которой относится статья (акции, облигации, токены, выплаты, регулирование и т. д.).
  - **Тип события** — характер финального события (выплата, эмиссия, корпоративное действие, дефолт-риск и т. д.).
  - **Упомянутая сущность** — конкретный объект из текста (эмитент, облигация, акция, токен, платформа, регулятор) с подсказкой типа и оценкой уверенности.
  - **Служебный аудит** — сохраняемая вместе с результатом запись о работе модели: версии промптов, ответы по этапам и флаги починки ответа. Нужен для проверки качества и отладки.
- 

## Техническая справка

### Модель

Смысловой разбор ведёт языковая модель **Mistral AI** (`mistral-small-latest`). Запросы идут с `temperature: 0` и жёстким форматом ответа `json_object`, чтобы ответы были стабильными и предсказуемыми. Если модель вернула невалидный JSON, делается одна повторная попытка с явным запросом «верни строгий JSON»; дальнейший брак уже не чинится.

### Слои конвейера

**Layer 0 — очистка (детерминированная, без модели).** Сырой текст приводится в порядок: убирается технический мусор и пустые фрагменты, слишком длинные тексты укорачиваются до допустимого предела.

**Layer 0.5 — быстрый фильтр (детерминированный, без модели).** Жёстко отбрасывает заведомо нерелевантное:

- слишком короткие тексты;
- статьи с бытовыми заголовками — погода, гороскоп, рецепт, спорт, кино, анекдот, анонсы мероприятий, авторынок.

Финансовые ключевые слова — **негейтирующая** подсказка: при их отсутствии статья всё равно идёт дальше к модели, чтобы фильтр не потерял важную новость с неочевидной формулировкой.

**Layer 1 — оценка релевантности (модель).** Решает, стоит ли разбирать статью. Чтобы пройти дальше, должны одновременно выполняться условия:

- статья признана релевантной;
- выбран режим глубокого разбора;
- оценка релевантности не ниже порога;
- указана хотя бы одна тематическая категория.

**Layer 2.1 — извлечение фактов (модель, только аудит).** Собирает факты, числовые значения, упомянутые объекты и замечания о неуверенности. Не задаёт финальные поля события — служит контекстом для следующего слоя.

**Layer 2.2 — построение события (модель).** Превращает факты в готовое событие: заголовок, описание, тип, характер влияния, оценки важности и уверенности, теги, упомянутые сущности, ключевые факты и предупреждения.

---

Revision #2

Created 2026-06-15 22:30:17 UTC by Admin

Updated 2026-06-15 22:47:56 UTC by Admin