

Исследование методов прогнозирования падения цены акций на основе цепочек новостей (на примере авиакомпаний)

Русалеев Владимир

27 декабря 2025 г.

Мотивация

- ▶ Финансовые рынки чувствительны к новостному фону
- ▶ Отдельные новости редко объясняют движение цен
- ▶ Важен накопительный и контекстный эффект
- ▶ Практическая цель — раннее выявление рискованных ситуаций

Постановка задачи

Цель:

- ▶ Исследовать возможность предсказывать снижение цены акции тех или иных компаний на основе анализа цепочек новостных сообщений
- ▶ Составить и обучить модель, способную предсказать падение цены акции более чем на 2–3% в течение 3 торговых дней после окончания новостного окна

Формализация:

- ▶ Объект — цепочка новостей
- ▶ Тип задачи — бинарная классификация(есть падение/нет)

Особенности задачи

- ▶ Высокая шумность новостных данных
- ▶ Редкость целевого события (5–6% среди всех новостей)
- ▶ Возможное отсутствие прямой причинно-следственной связи между движением цены и новостными заголовками

Выбор метрик качества

1. ROC-AUC — основная метрика, показывающая способность модели различать объекты между классами
2. Recall (полнота) для класса падения — ключевая метрика, равная доле корректно предсказанных падений.
3. Precision (точность) для класса падения — дополнительная метрика, показывающая долю произошедших падений среди всех предсказанных.

Данные

Новостные данные:

- ▶ Заголовки новостей
- ▶ Тикеры компаний
- ▶ Новости как конкретных компаний, так и геополитические, влияющие на всю сферу

Рыночные данные(yfinance):

- ▶ Дневные цены акций
- ▶ Объем торгов в эту дату, его отношение к среднему объему за неделю

Инженерия признаков

1. Количественные: число новостей в окне, burst-активность(всплески)
2. Тональные: средняя тональность, количество негативно окрашенных заголовков
3. Временные: день недели, месяц
4. Тематические и географические

Шаг 1: Одиночные новости

Подход:

- ▶ Каждая новость — отдельный объект
- ▶ Используются простые признаки для отдельных новостей

Результат:

- ▶ ROC-AUC: 0.4–0.5
- ▶ Практически отсутствует предсказательная сила

Вывод: одиночные новости не отражают рыночный контекст

Шаг 2: Цепочки новостей без эмбедингов

Идея:

- ▶ Агрегировать новости в скользящих окнах
- ▶ Размер окна: 14-45 дней
- ▶ Подсчет признаков уже для цепочек

Результат:

- ▶ ROC-AUC: ~ 0.65
- ▶ Recall: ~ 0.25

Вывод: контекст важен, но необходимы еще улучшения для повышения значения метрик

Шаг 3: Добавление текстовых эмбеддингов

Подход:

- ▶ Использование sentence-transformers моделей для заголовков
- ▶ Используются одиночные новости

Эффект:

- ▶ Рост ROC-AUC в сравнении с первым шагом (~ 0.63)
- ▶ Улучшение recall (~ 0.30)

Шаг 4: Добавление текстовых эмбеддингов к новостным цепочкам

Подход:

- ▶ Комбинируем идеи из предыдущих шагов
- ▶ Агрегация эмбеддингов внутри окна (std, max, mean)

Эффект:

- ▶ Существенный рост ROC-AUC (~ 0.8)
- ▶ Улучшение recall (~ 0.50)

Сравнение моделей для составления эмбеддингов

Модель	ROC-AUC
FinBERT	~ 0.81
BAAI/bge-base-en-v1.5	~ 0.825
intfloat/e5-large-v2	~ 0.833
Qwen3-Embedding-0.6B	~ 0.82
mixedbread-ai/mxbai-embed-large-v1	~ 0.825
BAAI/bge-m3	~ 0.845
all-MiniLM-L6-v2	0.84–0.845

Выбран all-MiniLM-L6-v2 как наиболее устойчивый и быстрый

Снижение размерности

PCA:

- ▶ 8–16 компонент — оптимально
- ▶ 24+ компонент → ухудшение качества

UMAP:

- ▶ Потеря глобальной структуры
- ▶ Нестабильные результаты

Сравнение моделей классификации

Модель	ROC-AUC	Recall
Logistic Regression	~ 0.8	~ 0.40
XGBoost	~ 0.82	~ 0.43
Random Forest	~ 0.82	~ 0.5
LightGBM	~ 0.83	~ 0.55
CatBoost	0.84—0.85	0.55—0.58

Ансамбли:

- ▶ Усреднение — стабильный рост качества
- ▶ Stacking — снижение recall

Дисбаланс:

- ▶ SMOTE ухудшал результаты
- ▶ Оптимизация порога оказалась эффективнее

Финальные результаты

- ▶ ROC-AUC: 0.83–0.85
- ▶ Recall: 0.57–0.63
- ▶ Precision: 0.42–0.48

Пример предсказания модели: новостная цепочка

Компания: Delta Air Lines

Новостное окно: 07.09.2014 – 07.10.2014

Предсказание модели:

- ▶ Вероятность падения: **0.595**
- ▶ Фактическое падение: **−3.16%** (в течение 3 торговых дней)

Новости в окне:

1. **2014-09-12** — Credit Suisse initiates coverage on airline stocks
2. **2014-09-20** — Wolfe Research downgrades Delta Air Lines to Peer Perform (*негатив*)
3. **2014-09-30** — Option alert: Delta Airlines Dec \$44 Call
4. **2014-10-01** — 8-K: Q3 capacity growth $\tilde{3}\%$, operating margin 15–16%
5. **2014-10-02** — Geopolitical: mortar fired from Gaza (*негатив*)
6. **2014-10-05** — Update: Q3 free cash flow \$900M

Выводы

1. Цепочки новостей информативнее одиночных сообщений
2. Эмбединги — основной источник сигнала
3. Качество представления данных важнее выбора модели
4. Новости дают полезный, но ограниченный сигнал

Дальнейшая работа

- ▶ Расширение на другие сектора
- ▶ Учёт ожиданий рынка
- ▶ Добавление макроэкономических факторов

Репозиторий

- ▶ Ссылка на GitHub:

`https://github.com/vladimir-rusaleev/project_sber_news_analysis`