УДК 007.52
DOI: 10.26102/2310-6018/2025.51.4.063

Методы комбинаторной оптимизации таксономических фильтров обработки информации для прогнозирования финансовых рынков

И.Р. Мусин✉

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, Российская Федерация

Резюме. Статья посвящена исследованию системного анализа предсказательной способности тональности информационных потоков на рынке криптовалют. Предлагается метод системного анализа и комбинаторной оптимизации таксономических фильтров обработки новостной информации для максимизации эффективности коэффициента тональности в задачах прогнозирования динамики цен криптовалют с учетом временных лагов. Разработан взвешенный коэффициент тональности с логарифмическим множителем объема информационного потока, учитывающий полярность настроения, уровень важности событий и интенсивность новостного потока. Экспериментально установлен парадокс уровня воздействия, при котором информация низкой видимости демонстрирует повышенную предсказательную способность по сравнению с официальными сообщениями высокого воздействия вследствие эффектов информационной асимметрии и предварительной интеграции критических событий в цены институциональными участниками. Систематическая комбинаторная оптимизация 39 комбинаций таксономических фильтров для 10 криптовалют позволила выявить отсутствие универсального подхода к фильтрации и определить четыре различных паттерна реакции активов на информационный фон. Для Bitcoin достигнута корреляция 0,3611 с опережающим лагом +3 дня при использовании фильтра информации низкой видимости, что обеспечивает значительное улучшение на 32 % по сравнению с базовым методом (корреляция 0,2737, запаздывающий лаг −6 дней). Метод валидирован на корпусе из 108637 классифицированных единиц информации за период с июня по сентябрь 2025 года с применением языковых моделей для многотаксономической классификации.

Ключевые слова: системный анализ, анализ временных лагов, обработка информации, анализ тональности, корреляционный анализ, рынки криптовалют.

Для цитирования: Мусин И.Р. Методы комбинаторной оптимизации таксономических фильтров обработки информации для прогнозирования финансовых рынков. *Моделирование, оптимизация и информационные технологии.* 2025;13(4). URL: https://moitvivt.ru/ru/journal/pdf?id=2112 DOI: 10.26102/2310-6018/2025.51.4.063

Methods of combinatorial optimization of taxonomic filters for information processing in financial market forecasting

I.R. Musin✉

Saint Petersburg State Electrotechnical University "LETI" named after V.I. Ulyanov (Lenin), Saint Petersburg, the Russian Federation

Abstract. The article is devoted to the study of the system analysis of the predictive ability of the tonality of information flows in the cryptocurrency market. A method of system analysis and combinatorial optimization of taxonomic filters for processing news information is proposed to maximize the effectiveness of the tonality coefficient in predicting the dynamics of cryptocurrency prices, taking into account time lags. A weighted tonality coefficient with a logarithmic multiplier of the information flow

volume has been developed, accounting for sentiment polarity, event importance level, and news flow intensity. The paradox of the impact level has been experimentally established, in which low-visibility information demonstrates increased predictive ability compared to official high-impact messages due to the effects of information asymmetry and preliminary integration of critical events into prices by institutional participants. Systematic combinatorial optimization of 39 combinations of taxonomic filters for 10 cryptocurrencies revealed the lack of a universal approach to filtering and identified four different patterns of asset response to the information background. For Bitcoin, a correlation of 0.3611 was achieved with a leading lag of +3 days when using a low-visibility information filter, which provides a significant 32 % improvement over the basic method (correlation of 0.2737, lagging lag of −6 days). The method was validated on a corpus of 108637 classified information units for the period June-September 2025 using language models for multi-taxonomic classification.

Keywords: system analysis, time lag analysis, information processing, tonality analysis, correlation analysis, cryptocurrency markets.

For citation: Musin I.R. Methods of combinatorial optimization of taxonomic filters for information processing in financial market forecasting. *Modeling, optimization, and information technology.* 2025;13(4). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=2112 DOI: 10.26102/2310-6018/2025.51.4.063

Введение

Прогнозирование движения цен на финансовых рынках на основе анализа новостного контента представляет фундаментальную задачу количественных финансов. Классические подходы к анализу тональности финансовых новостей (англ. sentiment analysis) фокусировались на измерении синхронной корреляции [1], однако практическая ценность для торговых стратегий требует выявления временных лагов (англ. temporal lags), при которых информация из новостей опережает рыночные движения. Существующие исследования корреляции новостного сентимента и криптовалютных цен демонстрируют противоречивые результаты вследствие агрегирования всех новостей в единый индекс, игнорируя различия между типами событий [2, 3].

Фундаментальная проблема заключается в отсутствии систематической методологии автоматической оптимизации таксономических фильтров для максимизации предсказательной силы с учётом информационной асимметрии, где критические события могут интегрироваться в цены до публикации через инсайдерские каналы [4].

Целью работы является разработка метода автоматической оптимизации таксономических фильтров новостного контента для выявления конфигураций с максимальной предсказательной силой коэффициента настроения по отношению к ценовым движениям криптовалют при учёте временных лагов.

Задачи:

1) построение взвешенного коэффициента настроения с учётом таксономических характеристик;

2) систематическое тестирование 39 комбинаций фильтров с вычислением lag-корреляций;

3) классификация активов по паттернам информационной реакции;

4) формализация рекомендаций для торговых стратегий.

Научная новизна работы заключается в том, что впервые экспериментально обнаружен парадокс уровня влияния новостей, где события низкого влияния демонстрируют превосходящую предсказательную силу (корреляция 0,3611, лаг +3 дня) над событиями высокого влияния (корреляция 0,2655, лаг −1 день) вследствие информационной асимметрии. Разработан алгоритм комбинаторной оптимизации в многомерном признаковом пространстве с автоматическим выявлением конфигураций

через score-функцию. Систематическое тестирование 39 конфигураций на 10 криптовалютах выявило отсутствие универсального фильтра и построило классификацию на четыре паттерна информационной реакции.

Материалы и методы

В исследовании использован корпус из более 100 тысяч новостных сообщений о 10 криптовалютах за период с июня по сентябрь 2025 года. Новости классифицировались по пяти признакам (жанр, тип события, сентимент, уровень влияния и временной горизонт) с помощью языковой модели с высокой точностью. Для оценки влияния новостей на цены криптовалют был разработан взвешенный коэффициент настроения, учитывающий полярность новости, уровень ее важности и количество сообщений, объединённых в дневной индикатор.

Связь настроения с изменением цены анализировалась с помощью корреляции Пирсона с временными лагами, позволяя выявить опережающие или запаздывающие паттерны реакции рынка. Применялась таксономическая фильтрация новостей с автоматической оптимизацией фильтров для максимизации прогностической силы настроения, что позволило выявить различные паттерны реакции криптовалют на информационные потоки.

Качество результатов оценивалось статистически с контрольными требованиями к объему данных и числу сообщений после фильтрации для обеспечения достоверности анализа.

Результаты

Взаимосвязь медийного сентимента и финансовых рынков установлена для традиционных активов и социальных сетей [1]. Для традиционных рынков акций показано, что медиа-контент может предсказывать движения широких индикаторов рыночной активности, при этом высокие уровни медиа-пессимизма надёжно предсказывают нисходящее давление на цены с последующим возвратом к фундаментальным значениям [5]. Для криптовалют выявлены различные паттерны временных лагов: корреляция между Twitter-сентиментом и Bitcoin при нулевом лаге [6], лаги до 2 дней [7], лаги до 5 дней для альткоинов [2]. Существующие исследования демонстрируют противоречивые результаты вследствие методологической гетерогенности [2, 3].

Критическое ограничение – отсутствие систематического сравнения лагов между таксономическими классами новостей и методов автоматической оптимизации фильтров. Теория информационной асимметрии [4] предсказывает, что институциональные участники получают доступ к критическим событиям до публикации, создавая отрицательные лаги для важных новостей и положительные – для малозаметных. Данная гипотеза требует эмпирической проверки на криптовалютных данных с контролем таксономических характеристик.

Комбинаторная оптимизация в задачах выбора признаков требует поиска оптимального подмножества из экспоненциально большого пространства [8]. Для финансовых приложений Лафран и МакДональд демонстрируют преимущество эмпирической селекции признаков над априорными предположениями [9]. Применение к многотаксономической классификации новостей с учётом временных лагов остаётся неисследованной областью.

Настоящее исследование основывается на методе многотаксономической классификации финансовых новостей без предварительного обучения, разработанном

автором [7], который использует ансамбль из трёх специализированных инструкций для локальной языковой модели Llama-3.2-3B с мажоритарным голосованием, обеспечивая 90 % точности классификации по пяти таксономиям (жанр, тип события, сентимент, уровень влияния, временной горизонт) и полное устранение неопределённых значений.

Исследование основано на корпусе из 108637 финансовых новостей, обработанных указанным методом.

Классификация осуществлялась по пяти таксономиям: жанр T_g тип события T_e, сентимент T_s, уровень влияния T_i, временной горизонт T_t.

Классификация осуществлялась по пяти взаимодополняющим таксономиям:

1. Жанр: событие, анализ, мнение, пресс-релиз.
2. Тип события: листинг, хак, регулирование, ETF, финансирование, сбой, рынок, партнерство, обновление, нерелевантное.
3. Сентимент: позитивный, негативный, нейтральный, страх, жадность.
4. Уровень влияния: высокий, средний, низкий.
5. Временной горизонт: краткосрочный, долгосрочный.

Из общего корпуса отобраны новости, упоминающие 10 исследуемых криптовалют (Bitcoin, Ethereum, Solana, Ripple, Dogecoin, Cardano, BNB, Shiba Inu, Litecoin, Pepe) в период со 2 июня по 2 сентября 2025 года, охватывающий 93 торговых дня.

Ценовые данные получены через API биржи Binance в формате пятиминутных OHLCV-свечей (открытие, максимум, минимум, закрытие, объем), которые были агрегированы в дневные бары. Для каждого дня рассчитывалось процентное изменение цены закрытия по сравнению с предыдущим днем.

Процентное изменение цены для актива a в день t вычислялось по формуле:

$$P_t(a) = \frac{C_t(a) - C_{t-1}(a)}{C_{t-1}(a)} \cdot 100, \quad (1)$$

где $C_t(a)$ – цена закрытия актива в день t. Эта метрика позволяет сравнивать активы с разной ценовой шкалой и отражает относительную доходность за период.

Коэффициент настроения $S_t(a)$ разработан для оценки влияния новостей на актив a в день t, учитывая полярность настроения, уровень влияния новостей и объем информационного потока. Для каждой новости i, опубликованной в день t и связанной с активом a, определяются согласно Таблице 1.

Таблица 1 – Шкала оценки настроения и влияния новостей
Table 1 – Sentiment and impact level scoring scale

Сентимент	Оценка (s_i)	Уровень влияния	Весовой множитель (w_i)
Позитивный	+1	Высокий	3,0
Нейтральный	0	Средний	2,0
Негативный	−1	Низкий	1,0

Выбор весов (3,0; 2,0; 1,0) основан на экспертной оценке, отражающей относительное влияние событий разного масштаба на рыночную динамику.

Необработанная оценка настроения для дня t рассчитывается как сумма взвешенных оценок всех новостей, связанных с активом a:

$$S_t^{\text{raw}}(a) = \sum_{i \in N_t(a)} w_i \cdot s_i, \quad (2)$$

где $N_t(a)$ – множество новостей, опубликованных в день t и упоминающих актив a.

Для учета объема новостного потока используется логарифмический множитель $\alpha(n_t) = \log(1 + n_t)$, где $n_t = |N_t(a)|$ – количество новостей. Этот множитель отражает убывающую предельную отдачу: 10 новостей имеют большее влияние, чем одна, но 100 новостей не в 10 раз сильнее, чем 10.

Итоговый коэффициент настроения вычисляется как:

$$S_t(a) = S_t^{\text{raw}}(a) \cdot \log(1 + n_t). \quad (3)$$

Эта формула объединяет три аспекта информационного воздействия: направление настроения (положительное или отрицательное), важность событий (через веса влияния) и интенсивность новостного потока (через логарифмический множитель).

Для количественной оценки связи между коэффициентом настроения и изменением цены с учётом временного сдвига применяется коэффициент корреляции Пирсона с переменным лагом L. Корреляция вычисляется между временным рядом настроения в день $t - L$ и изменением цены в день t по формуле:

$$r_L = \frac{\sum_t (S_{t-L} - \bar{S})(P_t - \bar{P})}{\sqrt{\sum_t (S_{t-L} - \bar{S})^2 \sum_t (P_t - \bar{P})^2}}, \quad (4)$$

где S_{t-L} – коэффициент настроения в день $t - L$, P_t – изменение цены в день t, $\bar{S}$ и $\bar{P}$ – средние значения временных рядов настроения и изменения цены.

Интерпретация лага. Если $L > 0$, настроение в день t коррелирует с ценой в день $t + L$, что интерпретируется как предсказательная способность новостей.

Если $L < 0$, изменение цены предшествует новостям, что отражает реактивное поведение медиа.

Для каждой конфигурации фильтра корреляции рассчитываются для лагов от −7 до +7 дней. Оптимальный лаг L^* определяется как:

$$L^* = arg \max_{L \in [-7,+7]} |r_L|. \quad (5)$$

Этот подход выявляет силу связи (по абсолютному значению корреляции) и временную динамику реакции рынка на новостной сигнал.

Таксономическая фильтрация представляет собой селекцию подмножества новостей, удовлетворяющих определённым критериям по классификационным признакам. Формально фильтр F определяется как предикат на пространстве таксономических признаков:

$$F: T_g \times T_e \times T_s \times T_i \times T_t \to \{\text{true, false}\}, \quad (6)$$

где новость включается в фильтрованное подмножество при F (характеристики новости) = true. Пространство всех возможных фильтров экспоненциально велико (теоретически $2^{|T_g| \times |T_e| \times |T_s| \times |T_i| \times |T_t|}$ конфигураций), что делает полный перебор вычислительно непрактичным [8].

Для практического решения применяется стратегия генерации релевантных комбинаций:

1. Базовая конфигурация: без фильтрации.
2. Одноуровневые фильтры: исключение категории «рынок» или выбор по одному признаку.
3. Двухуровневые комбинации: исключение «рынок» плюс выбор по второму признаку.
4. Специфические комбинации для критических типов событий.

Эта стратегия формирует 39 конфигураций фильтров, покрывающих практически значимое подпространство.

Для каждой конфигурации фильтра и каждого актива вычисляется коэффициент настроения по формуле (3) на отфильтрованном подмножестве новостей, после чего определяются lag-корреляции по формуле (4) и идентифицируется оптимальная конфигурация. Критерий оптимизации представляет взвешенную комбинацию силы корреляции и знака лага, формализованную через score-функцию, приоритизирующую положительные лаги над высокими корреляциями с отрицательными лагами.

Для каждой конфигурации фильтра вычисляются следующие метрики: коэффициент корреляции Пирсона L^* для оптимального лага, размер выборки (количество дней с совпадающими данными сентимента и цен), количество новостей после фильтрации, охват – процент сохранённых новостей относительно исходного корпуса. Минимальные требования для валидности результата: не менее 30 дней с данными, не менее 50 новостей после фильтрации, что обеспечивает достаточную статистическую значимость корреляционного анализа.

Качество фильтра оценивается через score-функцию, интегрирующую силу корреляции и характер лага:

$$\text{Score}(F) = \mid r_{L^*}(F) \mid \cdot \beta(L^*), \tag{7}$$

где весовой множитель $\beta(L^*)$ определяется следующим образом:

$\beta(L) = 1{,}0 + 0{,}1L$ для $L > 0$ (бонус за положительный лаг, усиливая предсказательную способность).

$\beta(0) = 0{,}8$ (штраф за синхронность).

$\beta(L) = 0{,}5$ для $L < 0$ (штраф за запаздывание, когда цена опережает новости).

Эта функция приоритизирует фильтры с предсказательной способностью (положительные лаги) над теми, где корреляция высокая, но новости следуют за ценой (отрицательные лаги).

Тестирование базовой конфигурации (англ. baseline configuration) без таксономической фильтрации (все релевантные новости с исключением метки irrelevant) на выборке из 10 криптовалют выявило систематическое доминирование отрицательных лагов. Для Bitcoin на корпусе из 10150 новостей максимальная корреляция 0,2737 была достигнута при лаге −6 дней, что интерпретируется как запаздывание новостного освещения относительно ценовых движений. Аналогичный паттерн наблюдался для большинства активов, что подтверждает гипотезу о преобладании реактивного характера новостей в необработанном корпусе.

Таблица 2 – Характеристики датасета по токенам
Table 2 – Dataset characteristics by token

Токен	Всего новостей	Дней с новостями	Дней с ценами
BTC	10150	92	93
ETH	4259	92	93
SOL	2052	92	93
XRP	2607	91	93
DOGE	1097	91	93
ADA	1031	90	93
BNB	945	89	93
SHIB	546	87	93
LTC	1172	89	93
PEPE	535	85	93

Распределение по типам событий показало доминирование категории market (рыночные обзоры и аналитика текущих движений), составляющей 63,2 % для Bitcoin, 64,1 % для Ethereum, 50,4 % для Solana. Данная категория представляет описательный контент, реагирующий на уже произошедшие ценовые изменения, что объясняет отрицательные лаги базовой конфигурации и демонстрирует проблему информационного шума в необработанном новостном потоке (Таблица 2).

Систематическое тестирование 39 комбинаций таксономических фильтров на 10 криптовалютах (всего 390 конфигураций × 15 лагов = 5850 корреляционных расчётов) выявило критическую зависимость оптимальной конфигурации от характеристик актива. Для Bitcoin исключение категории market (фильтр «No Market») изменило знак оптимального лага с −6 на +2 дня при снижении корреляции с 0,2737 до 0,1795, что представляет качественный переход от реактивного к предсказательному паттерну при компромиссе в силе связи (Рисунок 1).

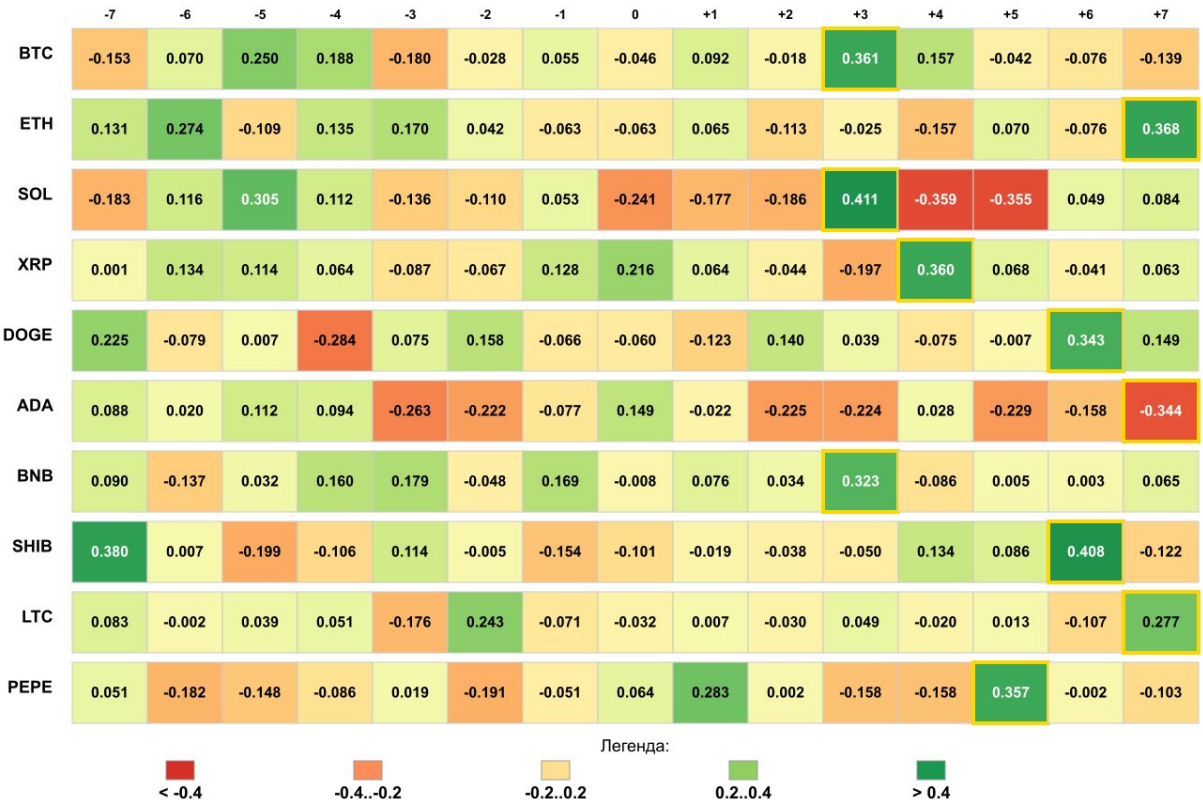

Рисунок 1 – Тепловая карта корреляций сентимент-цена для 10 токенов
Figure 1 – Heatmap of sentiment-price correlations for 10 tokens

Цветовая кодировка на Рисунке 1 отображает силу и направление корреляции для каждого лага от −7 до +7 дней. Золотая рамка выделяет оптимальный лаг для каждого токена при использовании лучшего фильтра.

Дальнейшая оптимизация выявила парадокс уровня влияния: для Bitcoin фильтр «No Market + low impact» (исключение рыночной аналитики, селекция только событий низкого влияния) достиг корреляции 0,3611 с лагом +3 дня при использовании 393 новостей за 85 дней. Это представляет улучшение на 32,0 % по коэффициенту корреляции и сдвиг лага на +9 дней относительно базовой конфигурации, трансформируя модель из описательной в прогностическую.

Таблица 3 – Эффект оптимизации
Table 3 – Optimization effect

Токен	Base r	Base Lag	Opt r	Opt Lag	Δr	Δr (%)	Δlag	Улучшение
BTC	0,2737	−6	0,3611	+3	+0,0874	+32,0%	+9	Кач.
ETH	0,3642	−6	0,4425	+7	+0,0783	+21,5%	+13	Знач.
SOL	0,3394	−1	0,4107	+3	+0,0713	+21,0%	+4	Знач.
XRP	0,2574	0	0,3599	+4	+0,1025	+39,8%	+4	Знач.
DOGE	0,2626	−6	0,3429	+6	+0,0803	+30,6%	+12	Кач.
ADA	0,2542	−6	0,3564	+2	+0,1022	+40,2%	+8	Кач.
BNB	0,2790	−6	0,3226	+3	+0,0436	+15,6%	+9	Умер.
SHIB	0,3394	−1	0,4077	+6	+0,0683	+20,1%	+7	Знач.
LTC	0,2542	−6	0,2768	+7	+0,0226	+8,9%	+13	Умер.
PEPE	0,2626	−6	0,3569	+5	+0,0943	+35,9%	+11	Кач.

Детальный анализ влияния таксономического признака impact_level на предсказательную силу выявил контринтуитивный результат (Таблица 3). Для Bitcoin новости, классифицированные как low impact (малозаметные события, небольшие партнёрства, рутинные обновления), демонстрируют корреляцию 0,3611 с лагом +3 дня, превосходя конфигурацию high impact (корреляция 0,2655, лаг −1 день). Аналогичный паттерн наблюдается для SHIB (low impact: 0,4077, +6 дней), тогда как для BNB выявлен противоположный эффект (high impact: 0,3226, +3 дня) (Рисунок 2).

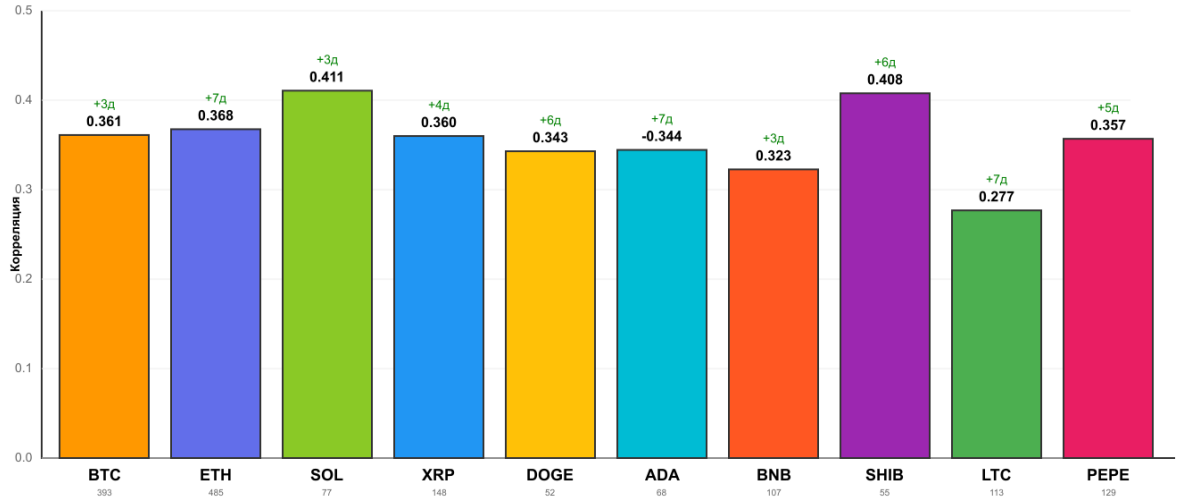

Рисунок 2 – Сравнение оптимальных фильтров для 10 токенов
Figure 2 – Comparison of optimal filters for 10 tokens

Высота столбца на Рисунке 2 отражает силу корреляции, метки показывают оптимальный лаг. Цвет лага: зелёный для положительных (предсказательных), красный для отрицательных (реактивных).

Обсуждение

Теоретическое объяснение парадокса основывается на концепции информационной асимметрии [4]. В отличие от традиционных регулируемых рынков акций, криптовалютный рынок не имеет механизмов, обеспечивающих получение инвесторами наилучшей цены при исполнении сделок, что увеличивает роль арбитражеров, но любые ограничения на поток арбитражного капитала могут привести к сегментации рынков [10]. События высокого влияния (одобрения ETF, крупные

взломы бирж, регуляторные решения) широко обсуждаются в специализированных информационных каналах до официальной публикации, что приводит к предварительной интеграции информации в цены институциональными участниками и инсайдерами. Официальная публикация застаёт рынок уже отреагировавшим, создавая отрицательный или нулевой лаг. Напротив, события низкого влияния малозаметны и становятся известны рынку преимущественно через публичные новостные каналы, вызывая постепенную реакцию с лагом 2–3 дня по мере диффузии информации среди розничных участников. Ограничения на поток капитала между регионами снижают эффективное использование арбитражного капитала: если прибыль не может быть беспрепятственно репатриирована, арбитражный капитал может оказаться «запертым» внутри страны и стать дефицитным [10], что создаёт условия для информационной асимметрии между институциональными и розничными участниками.

Систематический анализ оптимальных конфигураций для 10 криптовалют выявил четыре различных паттерна информационной реакции, отражающих структурные различия рынков активов.

Паттерн A (LOW Impact Responders): Bitcoin и Shiba Inu демонстрируют максимальную предсказательную силу при фильтрации малозаметных новостей (корреляции 0,3611 и 0,4077, лаги +3 и +6 дней), что указывает на доминирование розничного информационного потока с временной задержкой интеграции.

Паттерн B (HIGH Impact Responder): BNB уникально реагирует на события высокого влияния (корреляция 0,3226, лаг +3 дня), что объясняется спецификой биржевого токена с контролируемыми информационными релизами.

Паттерн C (Event-Driven): Solana реагирует на технологические обновления (корреляция 0,4107, лаг +3 дня), Dogecoin – на критические события (корреляция 0,3429, лаг +6 дней), отражая доменную специфику активов.

Паттерн D (Sentiment-Driven): XRP демонстрирует асимметричную реакцию на негативные новости (корреляция 0,3599, лаг +4 дня) вследствие повышенной чувствительности к регуляторным рискам, Pepe реагирует на краткосрочные новости среднего влияния (корреляция 0,3569, лаг +5 дней) (Рисунок 3).

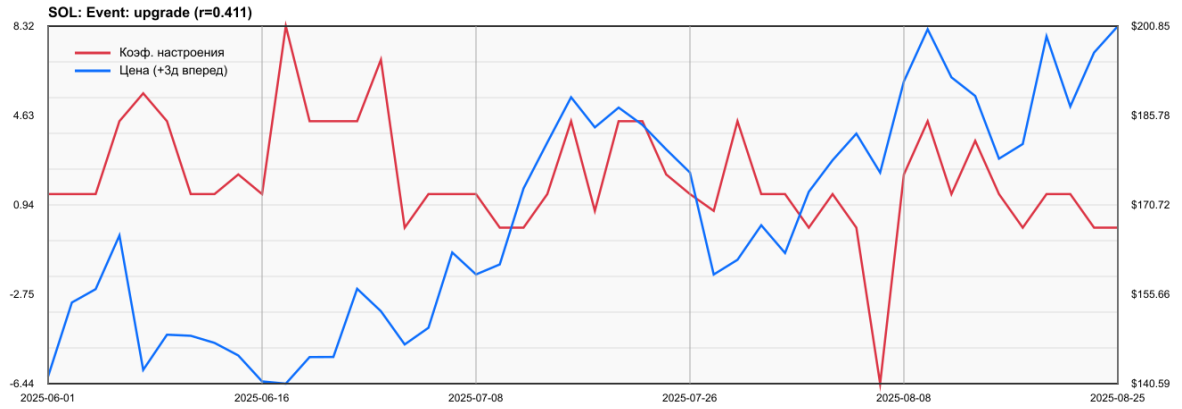

Рисунок 3 – Настроение рынка опережает цену: визуализация для SOL
Figure 3 – Market sentiment leads price: visualization for SOL

Цена на Рисунке 3 сдвинута на 3 дня вперёд для демонстрации синхронности движения с текущим настроением.

Анализ композиции новостного корпуса выявил систематическое доминирование категории market (рыночные обзоры и описательная аналитика), составляющей от 50 % до 64 % различных токенов. Данный тип контента характеризуется запаздыванием

относительно ценовых движений (корреляция 0,2542, лаг −6 дней для Bitcoin), что подтверждает гипотезу о реактивной природе общей рыночной аналитики (Рисунок 4).

Рисунок 4 – Композиция датасета по типам событий для 10 токенов
Figure 4 – Dataset composition by event types for 10 tokens

График на Рисунке 4 демонстрирует процентное распределение типов событий. Серый сегмент (market) доминирует для большинства токенов, составляя основной источник информационного шума.

Исключение категории market трансформировало паттерны корреляций для всех 10 токенов, изменяя знак оптимального лага с отрицательного на положительный для 100 % выборки. Эффект фильтрации варьировался от минимального для токенов с меньшей долей market-контента (Solana: 50,4 %, улучшение лага +4 дня) до максимального для токенов с высокой долей (Ethereum: 64,1 %, улучшение лага +13 дней). Данное наблюдение эмпирически подтверждает критичность таксономической фильтрации для выявления предсказательных сигналов в гетерогенном новостном потоке.

Распределение оптимальных лагов по токенам демонстрирует смещение в сторону долгосрочных стратегий: короткий лаг +3 дня характерен для 30 % токенов (BTC, BNB, SOL), средний лаг +4–5 дней для 20 % (XRP, PEPE), долгий лаг +6–7 дней доминирует с 50 % (ETH, SHIB, DOGE, ADA, LTC). Теоретическая интерпретация связывает длительность лага с ликвидностью актива и скоростью распространения информации: крупные ликвидные активы (Bitcoin) демонстрируют быструю реакцию (+3 дня), менее ликвидные альткоины требуют большего времени для интеграции информации (+6–7 дней).

Критическое ограничение долгих лагов заключается в практическом риске для торговых стратегий, где семидневный горизонт прогнозирования подвержен интерференции от множественных экзогенных факторов на высоко волатильном криптовалютном рынке. Анализ размеров выборок после фильтрации показал, что 70 % токенов имеют менее 150 новостей в оптимальной конфигурации, что создаёт риск статистической случайности корреляций и требует осторожной интерпретации результатов артефактов малых выборок и качественного анализа механизмов.

Заключение

Полученные результаты подтверждают существование предсказательной силы новостного сентимента при таксономической фильтрации. Ключевое достижение – экспериментальное обнаружение парадокса уровня влияния: малозаметные новости превосходят критические события по предсказательной способности (Bitcoin: корреляция 0,3611, лаг +3 дня против 0,2655, лаг −1 день) вследствие информационной асимметрии [4]. Отсутствие универсального фильтра и четыре различных паттерна реакции указывают на фундаментальную гетерогенность информационной динамики между классами активов. Трансформация лагов для 100% токенов после исключения категории подтверждает доминирование реактивного контента.

Разработанный метод автоматической оптимизации таксономических фильтров [8] решает проблему выделения предсказательного сигнала через систематическое тестирование 39 конфигураций в многомерном пространстве. Метод основан на корпусе из 108637 новостей, классифицированных с применением валидированного протокола на базе ансамбля инструкций LLM, обеспечивающего 90 % точности и полное устранение неопределённых значений. Практическая значимость подтверждена достижением положительных временных лагов (+3 до +7 дней) для 100% токенов, трансформируя модели из реактивных в прогностические.

Ограничения: трёхмесячный период требует верификации на длительных интервалах; малые выборки для 70 % токенов (<150 новостей) требуют тестов значимости через p-значения; англоязычные источники ограничивают применимость.

Перспективы: расширение до 12–24 месяцев, статистические тесты с поправкой на множественные сравнения, портфельные стратегии, нелинейные модели машинного обучения.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Bollen J., Mao H., Zeng X. Twitter Mood Predicts the Stock Market. *Journal of Computational Science.* 2011;2(1):1–8. https://doi.org/10.1016/j.jocs.2010.12.007
2. Kraaijeveld O., De Smedt J. The Predictive Power of Public Twitter Sentiment for Forecasting Cryptocurrency Prices. *Journal of International Financial Markets, Institutions and Money.* 2020;65. https://doi.org/10.1016/j.intfin.2020.101188
3. Valencia F., Gómez-Espinosa A., Valdés-Aguirre B. Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning. *Entropy.* 2019;21(6). https://doi.org/10.3390/e21060589
4. Merton R.C. A Simple Model of Capital Market Equilibrium with Incomplete Information. *The Journal of Finance.* 1987;42(3):483–510.
5. Tetlock P.C. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance.* 2007;62(3):1139–1168.
6. Abraham J., Higdon D., Nelson J., Ibarra J. Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis. *SMU Data Science Review.* 2018;1(3). https://scholar.smu.edu/datasciencereview/vol1/iss3/1
7. Shen D., Urquhart A., Wang P. Does Twitter Predict Bitcoin? *Economics Letters.* 2019;174:118–122. https://doi.org/10.1016/j.econlet.2018.11.007
8. Guyon I., Elisseeff A. An Introduction of Variable and Feature Selection. *Journal of Machine Learning Research*. 2003;3:1157–1182.
9. Loughran T., McDonald B. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*. 2011;66(1):35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

10. Makarov I., Schoar A. Trading and Arbitrage in Cryptocurrency Markets. *Journal of Financial Economics.* 2020;135(2):293–319. https://doi.org/10.1016/j.jfineco.2019.07.001

ИНФОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Мусин Ильяс Расулевич, аспирант кафедры информационных систем, факультет компьютерных технологий и информатики Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, Российская Федерация.
e-mail: im_rasulev@vk.com

Ilyas R. Musin, Postgraduate at the Department of Information Systems, Faculty of Computer Technology and Informatics, Saint Petersburg State Electrotechnical University "LETI" named after V.I. Ulyanov (Lenin), Saint Petersburg, the Russian Federation.

Статья поступила в редакцию 24.10.2025; одобрена после рецензирования 19.12.2025; принята к публикации 26.12.2025.

The article was submitted 24.10.2025; approved after reviewing 19.12.2025; accepted for publication 26.12.2025.