ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ ISSN 2079-3316
научная статья искусственный интеллект, интеллектуальные системы, нейронные сети

УДК 004.832:336.761
10.25209/2079-3316-2025-16-1-83-130

Мультимодальное предсказание цен акций на примере российского рынка ценных бумаг

Касымхан Юсуфович **Хубиев**[1]✉, Михаил Евгеньевич **Семенов**[2]
[1,2] Университет «Сириус», «Сириус», Россия
[1]✉*kasymkhankhubievnis@gmail.com*

Аннотация. Классические методы прогнозирования цен активов в основном опираются на числовые данные, такие как временные ряды цен, объемы торгов, распределение лимитированных ордеров и индикаторы технического анализа. Однако новостной фон играет существенную роль в формировании цен, что делает актуальным развитие мультимодальных подходов, объединяющих текстовые и числовые данные для повышения точности предсказаний.

В данной работе решается задача прогнозирования цен финансовых активов с использованием мультимодального подхода, объединяющего временные ряды цен и текстовую модальность новостного потока. Для исследований был собран уникальный набор данных, включающий временные ряды для 176 акций российских компаний, торгуемых на Московской бирже, и 79555 русскоязычных финансовых новостей.

Для обработки текстовых данных использовались предобученные модели RuBERT и Vikhr-Qwen2.5-0.5b-Instruct (большая языковая модель), временные ряды и векторизованная текстовая модальность обрабатывались рекуррентной нейронной сетью LSTM. В ходе экспериментов сравнивались модели с одной модальностью и двумя модальностями, а также различные методы агрегации векторных представлений текстов.

Качество прогнозов оценивалось по двум ключевым метрикам: точности (accuracy) предсказания направления изменения цены (рост/снижение) и средней абсолютной процентной ошибке (MAPE) отклонения предсказанной цены от истинной. Эксперименты показали, что добавление текстовой модальности позволяет уменьшить значение MAPE на 55%.

Полученный мультимодальный набор данных представляет ценность для дальнейшей адаптации языковых моделей в финансовой сфере. Перспективные направления исследований включают оптимизацию параметров текстовой модальности, таких как временное окно, тональность и хронологический порядок новостных сообщений.

Ключевые слова и фразы: мультимодальная предсказательная модель, количественные финансы, машинное обучение

Благодарности: Результаты получены при финансовой поддержке исследования, реализуемого в рамках государственной программы федеральной территории «Сириус» «Научно-технологическое развитие федеральной территории „Сириус“» (Соглашение №18-03 от 10.09.2024)

Для цитирования: Хубиев К. Ю., Семенов М. Е. *Мультимодальное предсказание цен акций на примере российского рынка ценных бумаг* // Программные системы: теория и приложения. 2025. **Т. 16**. № 1(64). С. 83–130. https://psta.psiras.ru/read/psta2025_1_83-130.pdf

Введение

Построение прогноза цены актива является важной задачей для участников финансового рынка для стратегического планирования, оптимального управления инвестиционным портфелем и учета рисков. Существует множество попыток применения методов машинного обучения для построения таких прогнозов.

В связи с ростом популярности моделей глубокого обучения исследователи сместили свой фокус в сторону применения нейронных сетей [1–3]. При этом проблема учета новостного потока, как важного фактора влияния на поведение рынка, переосмысляется с бурным развитием генеративных моделей искусственного интеллекта и больших языковых моделей (ChatGPT, FinGPT, GigaChat, LLama и другие). В финансовой экономике большие языковые модели применяются достаточно редко и их потенциал еще не раскрыт.

Исследователи предпринимают попытки применения моделей обработки естественного языка для улучшения качества прогноза цен активов и стратегий для управления инвестиционным портфелем.

В статье [4] описывается использование оценки тональности новостей в качестве дополнительного параметра. А вторы использовали модель FinBert, обученную на финансовых данных, для оценки тональности (положительная, негативная или нейтральная) новостей. В работе использовались временные ряды свечных данных индекса американского рынка ценных бумаг Standart and Poor's 500 (S&P500). Для предсказания цены использовалась модель машинного обучения – случайный лес. Результатом исследования стал вывод – учёт тональности новостного потока улучшает качество предсказания.

В статье [5] авторы обозначили цель создать мультимодальную модель искусственного интеллекта, способную предоставлять обоснованный и точный прогноз по временным рядам. В ходе работы была реализована модель, которая генерирует прогноз месячной или недельной доходности актива, сопровождаемый текстовым пояснением языковой модели по введенному пользователем запросу. В статье [6] предложен подход к настройке инструкций для интерпретации числовых значений и трактовки финансового контекста.

В исследовании Куликовой с коллегами [7] изучен эффект от учета разделения новостей на группы по тематическому признаку. Авторы продемонстрировали, что в большинстве случаев целесообразно использовать одну тематическую группу новостей для рассмотренных моделей глубокого обучения (темпоральная сверточная сеть, D-Linear, трансформатор, и трансформатор темпорального слияния), а также определили вероятности улучшения прогнозов для рассмотренных 20 тематических групп.

Во всех выше перечисленных исследованиях модели были имплементированы с помощью мультимодального подхода для рынка ценных бумаг США, язык модальности – английский. При этом новостной поток был интегрирован в входной вектор предсказателя не напрямую, а через блок предобработки в виде дополнительного параметра, например, оценки тональности, частоты новостей по активу, класс новости.

Целью данной работы является демонстрация преимущества нового мультимодального метода над предсказаниями, построенными только на числовых данных, и представление русскоязычного набора данных финансовых новостей.

Для достижения поставленной цели сформулированы следующие основные задачи.

(*1*) Сформировать мультимодальный набор данных из временных рядов и новостных сообщений.

(*2*) Создать предсказательную модель – для использования одной и двух модальностей.

(*3*) Провести обучение предсказательной модели и анализ значений функций и метрик точности, Accuracy и MAPE.

В данной работе мы предлагаем новый мультимодальный подход для интеграции новостного потока во временной ряд числовых данных. Текст новостей отображается в векторное представление и подается в модель наряду с вектором временных рядов. Наша гипотеза заключается в следующем: мультимодальный подход позволит предсказательным моделям извлечь семантическую информацию из текста, что улучшит качество предсказания цены актива.

1. Сбор и структурирование данных

Мультимодальность подразумевает применение более одной модальности данных, что влияет на структуру данных и логику разработки предсказательной модели. Мы используем два вида модальности:

числовую – временные ряды цен на акции,
текстовую – новостной поток.

Для обучения предсказательной модели и исследования её работы был сформирован оригинальный набор данных.

Временные ряды в виде свечей с указанием цен открытия (open), закрытия (close), максимальной (high) и минимальной цен (low) мы получили с помощью программного интерфейса Algopack (API) Московской биржи (MOEX). Для проведения численного эксперимента мы выбрали временные ряды акций с 7 июля 2022 года по 30 августа 2024 года для 176

компаний. В указанный временной интервал российский фондовый рынок продемонстрировал фазы быстрого роста и падения, индекс IMOEX вырос за этот период с 2213,81 до 2650,32 пунктов (+19, 72%).

Мы собрали 79555 новостных сообщений из различных источников, в том числе – сетевое издание «РБК» (1823), «БКС Экспресс» (11331) и «БКС Теханализ» (9670), сайт инвестиционной компании «Финам» (20647), сайт сообщества трейдеров «SmartLab.ru» (30857), а также телеграм-канал «РДВ» (5227).

Выбор указанных источников аргументирован несколькими причинами. Во-первых, на этих источниках опубликованы новости за необходимый временной интервал. Во-вторых, институциональное различие источников, стиль изложения с разной степенью экспертности – позволит сформировать более объективное освещение событий, связанных с используемыми временными рядами.

Новостные сообщения были токенизированы с использованием двух моделей RuBert [8] и Vikhr-Qwen2.5-0.5b-Instruct [9] (далее Qwen). Под словом в контексте токенезированного текста подразумевается токен – элемент векторного пространства в виде индекса словаря токенезатора.

Описательные характеристики (среднее, отклонение, минимальное, максимальное количество слов, а также квартили) набора данных приведены в таблицах 1 и 2 (токены). Необходимо отметить, что токенизация

Таблица 1. Статистические характеристики набора данных после токенизации, RuBert

Источник	Mean	Std	Min	Max	Q25	Q50	Q75
РДВ	134	88	8	512	65	123	187
Финам	221	135	18	512	116	178	284
БКС Экспресс	20	10	4	82	13	17	26
БКС Теханализ	502	37	29	512	512	512	512
РБК	43	7	16	75	39	44	48
SmartLab	21	8	5	82	15	19	25

может привести к увеличению количества слов в тексте (например, за счет разделения слова на составные части).

В таблице 3 приведены примеры того, как изменяется фраза после токенизации. Например, слово «открывает» разделяется на три составных элемента: «от», «##к», «##рывает», где префикс «##» означает, что токен является продолжением предыдущего токена.

Таблица 2. Статистические характеристики набора данных после и токенизации, Qwen

Источник	Mean	Std	Min	Max	Q25	Q50	Q75
РДВ	215	157	3	1324	92	187	304
Финам	453	405	35	5732	211	319	501
БКС Экспресс	36	19	5	163	23	32	47
БКС Теханализ	1493	310	40	2221	1448	1545	1665
РБК	75	12	28	105	68	77	83
SmartLab	33	12	7	120	25	31	39

Таблица 3. Примеры оригинального и токенизированного текста

Оригинальный текст	Токенизированный текст
Доллар снова ниже 69 рублей	До ##лла ##р снова ниже 69 рублей
Москвич банкрот?	Москви ##ч банк ##рот ?
НПО Наука Отчет РСБУ	Н ##П, ##О Наука От ##чет Р ##С ##Б ##У
Т-банк это желтый банк	Т - банк это же ##лт ##ый банк

Особенности новостных сообщений. На ресурсе «БКС Теханализ» новостные статьи часто большие по объему, что накладывает ограничение на использование токенизаторов. В частности, из таблиц 1 и 2 видно, модель RuBert на больших текстах усекает токенизированный вектор. К тому же средняя длина токенизированного текста с помощью модели Qwen превосходит среднюю длину токенизированного текста RuBert, что говорит о том, что модель Qwen обладает более широким словарем и более сильной способностью декомпзиции текста.

Дополнительно мы собрали данные о 176 компаниях, сформировав набор данных из кортежей вида:

(тикер, наименование компании, описание деятельности компании).

Такие данные необходимы в нашем случае для:

(*а*) извлечения ключевых слов из описания,

(*б*) улучшения способности языковой модели связывать события, описываемые в новостном сообщении, с конкретной компанией и оценивать влияние новости на динамику цены.

Набор данных с новостными статьями содержит следующие параметры: дата публикации, источник, заголовок и тело статьи, теги (ключевые слова). Для источников РДВ, SmartLab заголовок отсутствует, соответствующие поля заполнены специальным словом: *no title*.

В нашем случае теги могут содержать полное или краткое название компании и соответствующий тикер, наименование сектора рынка и т. п. Теги в новостных сообщениях были установлены авторами статей. В случае источника «РДВ» теги отмечались авторами в виде хештегов (например, #цифры, #аналитика), «БКС Экспресс» и «БКС Теханализ» теги обозначались в специальных полях в начале или конце новостной статьи (например, ФосАгро, Российский рынок), и извлекались из *HTML* кода страницы по соответствующим *HTML* тегам. При отсутствии тегов (РБК, SmartLab) параметр в наборе данных остается пустым.

В таблице 4 приведены примеры новостной статьи (фрагмент заголовка) и приписанные к ней теги.

Таблица 4. Примеры новостных статей (фрагмент заголовка) и приписанные теги

Источник	Фрагмент статьи (заголовок)	Теги
РДВ	Сегежа (SGZH): таргет 16.2 руб., апсайд +102...	SGZH
РДВ	Артген биотех (ABIO) завершил доклинические...	аналитика, ABIO
Финам	Индекс МосБиржи восстанавливает позиции и приб...	ФосАгро, ВСМПО-АВСМ, CNYRUB
Финам	«Ашинский метзавод» назвал АО "Урал-ВК" своим ...	АшинскийМЗ
БКС Экспресс	«Восходящее окно»: в каких бумагах замечен это...	Селигдар SELG, ЕвроТранс EUTR
БКС Экспресс	«Сила Сибири» выйдет на максимальную мощность...	Газпром GAZP
БКС Теханализ	Мечел. Что ждать от бумаг на следующей неделе	Мечел
БКС Теханализ	На предыдущей торговой сессии акции Норникеля ...	ГМК Норникель

2. Методология

Для проверки нашей гипотезы о преимуществе мультимодального подхода мы запланировали серию экспериментов.

Первая серия экспериментов была направлена на прогнозирование цен с использованием только числовых рядов свечных характеристик актива (цены close, open, high, low). Метрики качества этого эксперимента будут являться базовыми значениями, относительно которых будет оцениваться прирост качества предсказания цен с применением предложенного мультимодального подхода.

Вторая серия экспериментов направлена на получение предсказаний и вычисления метрик точности (Accuracy, MAPE) с применением мультимодального подхода, рассмотрение разных методов агрегации (Sum, Mean) векторизованного новостного потока.

2.1. Использование одной модальности

Сначала мы провели серию экспериментов по прогнозированию цены активов только на временных рядах. Для этого к ежедневным значениям цен (close, open, high, low) мы применили модели классического машинного обучения: линейная регрессия (LinReg), k-ближайших соседей (KNN), решающее дерево (DT), случайный лес (RF) и бустинговый алгоритм XGBoost (XGB), среди моделей глубокого обучения мы использовали – рекурентную нейронную сеть долгой и короткой памяти (LSTM).

Концептуально эксперимент состоит из двух задач:

($а$) предсказание направления движения цены (рост или падение) – задача бинарной классификации

($б$) предсказание цены – задача регрессии.

На данном этапе эксперимента 176 компаний были сгруппированы по 23 секторам деятельности. Мы случайным образом выбрали 9 секторов экономики, а затем внутри секторов выбрали случайным образом по 2 компании. В таблице 5 перечислены секторы и компании (тикер), которые участвовали в вычислительном эксперименте. В таблице 6 показано распределение новостей по компаниям после фильтрации. В таблице 7 приведены статистические данные о временных рядах цен закрытия выбранных активов. Тепловая карта корреляций временного ряда цен закрытия активов приведена на рисунке 1. Интересная особенность рассматриваемого интервала – рынок претерпел две смены фазы – с общего снижение цен к росту и обратно, как показано на рисунке 2 вертикальными линиями.

Таблица 5. Секторы экономики и компании (тикер), включенные в набор данных

Сектор	Компания (тикер)
Металлы и добыча	Мечел (MLTR), Трубная металлургическая компания (TRMK)
Нефть и газ	Сургутнефтегаз (SNGS), Газпромнефть (SIBN)
Потребительский сектор	Магнит (MGNT), Лента (LENT)
Строительство	ПИК (PIKK), Самолет (SMLT)
Телекоммуникации	МТС (MTSS), Ростелеком (RTKMP)
Транспорт	Аэрофлот (AFLT), Совкомфлот (FLOT)
Финансы	Банк Санкт-Петербург (BSPB), ЭсЭфАй (SFIN)
Химия	ФосАгро (PHOR), Казаньоргсинтез (KZOSP)
Электроэнергетика	РусГидро (HYDR), МРСК Центра (MRKC)

Таблица 6. Распределение новостей по компаниям после фильтрации

Компания (тикер)	Количество новостей
Мечел (MLTR)	4258
Трубная металлургическая компания (TRMK)	11739
Сургутнефтегаз (SNGS)	12674
Газпромнефть (SIBN)	11421
Магнит (MGNT)	1236
Лента (LENT)	311
ПИК (PIKK)	897
Самолет (SMLT)	3392
МТС (MTSS)	1101
Ростелеком (RTKMP)	628
Аэрофлот (AFLT)	1429
Совкомфлот (FLOT)	14476
Санкт-Петербургская биржа (BSPB)	14278
ЭсЭфАй (SFIN)	1647
ФосАгро (PHOR)	2773
Казаньоргсинтез (KZOSP)	168
РусГидро (HYDR)	1921
МРСК Центра (MRKC)	1576

Таблица 7. Описательные характеристики для акций компаний

Тикер	Mean	Std	Min	Max	Q25	Q50	Q75
MTLR	191.8245	72.5652	81.2800	332.8800	123.8500	187.6700	251.6400
TRMK	153.1245	64.9362	55.8200	271.0000	87.1400	166.4200	218.7800
SNGS	27.0104	4.0119	17.3500	36.9600	23.7750	27.3300	30.0250
SIBN	601.5097	163.9205	335.5500	934.2500	452.0500	582.6500	748.9000
MGNT	5691.6429	1161.7684	4040.0000	8444.0000	4665.0000	5495.0000	6375.0000
LENT	814.3870	154.9502	650.0000	1263.0000	716.5000	749.0000	843.5000
PIKK	732.6617	94.8650	518.0000	955.5000	656.7000	732.9000	811.5000
SMLT	3120.8996	594.1018	1926.5000	4145.5000	2572.0000	3045.0000	3713.0000
MTSS	264.5382	32.0791	183.0000	346.9500	239.0000	266.2500	289.7500
RTKMP	68.1797	9.2753	52.2500	92.1000	60.4500	68.0000	74.7000
AFLT	38.1316	10.3131	22.4400	64.4000	27.9700	38.8800	44.1200
FLOT	88.0111	39.5834	29.9200	149.3000	42.1000	97.2000	124.1800
BSPB	211.1501	101.2533	67.5700	387.6800	100.8400	210.9900	295.3400
SFIN	762.9939	428.5679	425.8000	1975.0000	497.4000	518.0000	992.0000
PHOR	6774.6040	618.1977	4997.0000	8153.0000	6416.0000	6763.0000	7278.0000
KZOSP	25.8603	5.2029	15.3500	40.5700	21.9400	27.0700	29.8500
HYDR	0.7697	0.0810	0.5178	1.0278	0.7318	0.7721	0.8210
MRKS	0.5247	0.2382	0.2025	1.0745	0.2735	0.5550	0.7475

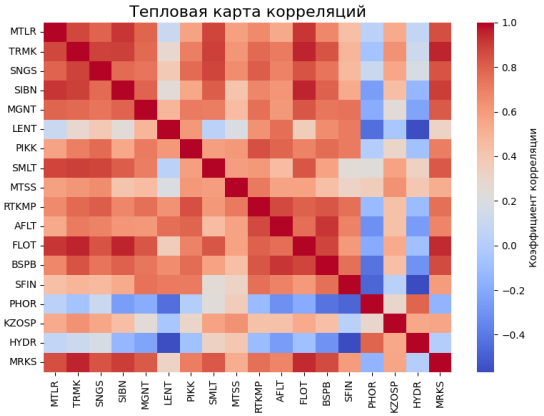

Рисунок 1. Тепловая карта корреляций цен закрытия активов

Рисунок 2. Нормированные исторические цены закрытия активов с указанием точек смены фазы рынка (вертикальные пунктирные линии)

Для оценки качества предсказания в задаче классификации использовалась метрика *Accuracy* (точность), для регрессии – *MAPE* (средняя абсолютная ошибка в процентах). Выбор этих метрик аргументируется постановкой задач. В случае классификации модель должна наиболее точно предсказать направление движение цены – рост (знак плюс «+».) или падение (знак минус «–»). Метрика *MAPE* наилучшим образом подходит для оценки качества задачи регрессии с точки зрения доменной области – финансов: *MAPE* демонстрирует среднее отклонение от цены актива в процентах, что легко переводится в денежный эквивалент.

На рисунке 3 приведен процесс разработки модели для использования одной и двух модальностей.

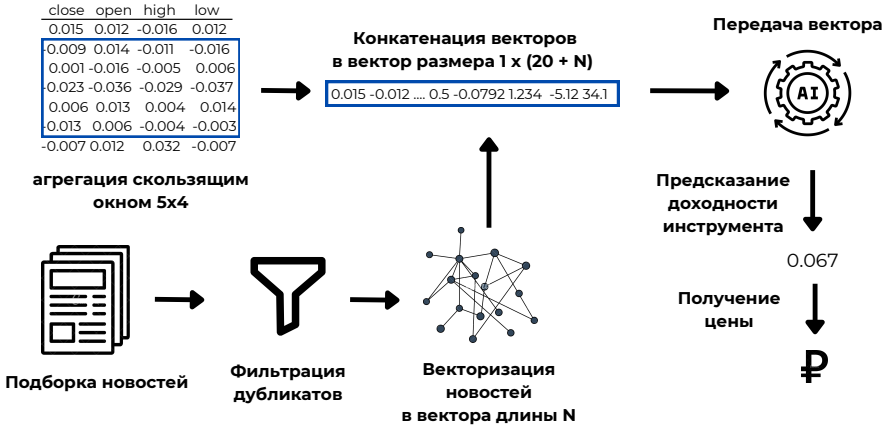

Рисунок 3. Процесс разработки модели для использования одной и двух модальностей

В качестве входного параметра в модель передавался вектор доходности инструмента, рассчитанный для цены закрытия (close) за предыдущие пять торговых сессий:

$$Return(d+1) = \frac{close(d+1)}{close(d)} - 1. \tag{1}$$

На выходе модели – предсказание на следующую торговую сессию.

Для оценки качества предсказания направления движения цены в качестве предсказанного класса использовался знак ($\pm$) предсказанного значения доходности, так как физический смысл доходности инструмента

это величина относительного прироста. Таким образом, положительное значение доходности говорит о росте цены, отрицательное – падении. Для оценки качества прогноза цены актива предсказанный вектор доходности инструмента преобразовывался в цену (в рублях):

$$price(d+1) = (Return(d+1)+1) \cdot price(d). \tag{2}$$

Полученный в результате преобразования поточечно спрогнозированный вектор цен сравнивался с историческим вектором цен активов по метрике *MAPE*.

Выбор величины доходности инструмента (а не цены инструмента) в качестве целевого значения для предсказательной модели обоснован тем, что при выходе цены при росте (падения) рынка за пределы исторического максимума (минимума) возможность использования методов ограничена.

Исходя из этого рассуждения свечные характеристики (цены close open, high, low) учитывались в виде значений *относительных приростов цен*, рассчитанных по формуле аналогично (1).

Далее из значений относительных приростов цен скользящим окном в пять торговых дней формируется вектор-строка и он подается в предсказательную модель. Таким образом, на вход модель получает вектор из двадцати параметров и на выходе предсказывает одно значение – величину доходности инструмента на конец следующей торговой сессии.

2.2. Использование двух модальностей

Для проведения эксперимента с применением новостного потока мы отобрали по ключевым словам новости, которые соответствуют анализируемым активам (таблица 5). Ключевые слова были выбраны как топ-30 слов извлеченных методом TF-IDF. Данный метод вычисляет важность слов в тексте относительно частоты появления слова и его уникальность во всем тексте. Пример извлеченных ключевых слов методом TF-IDF приведен в таблице 8.

Получив список ключевых слов с помощью метода TF-IDF, мы дополнительно добавили ключевые слова с помощью модели ChatGPT-4o, увеличив тем самым вариабельность ключевых слов с помощью перестановок, замен букв, изменения окончаний (таблица 9). Отобранные новости для каждой компании (тикера) были отображены в векторы и отфильтрованы на предмет дубликатов.

Таблица 8. Ключевые слова, полученные из описаний компаний

Тикер	Ключевые слова
MTLR	мечел, горнодобывающей, руда, сырье, энергия, ферросплавы, уголь
SNGS	газ, геологоразведка, нефть, сургутнефтегаз, нефтепродукты, электроэнергия, бурение
SMLT	аренда, девелопмент, девелопер, недвижимость, строительство, московский регион, жилые кварталы
MTSS	абонент, автоматизация, интернет, мобильной связи, провайдер, коммуникационных
BSPB	банк, вклад, дивиденды, финансовые услуги, калининград, спбанк, санкт-петербург

Таблица 9. Ключевые слова, полученные дополнительно

Тикер	Ключевые слова
MTLR	мечел, метчел, мечал, mechel, Mchel, ферросплавы, фурросплав
SNGS	сургутнефтегаз, surgutneftegaz, surgut, сурнефтегаз, сургаз, сургут, сур-нфтгз
SMLT	самолет, smlt, samolet, samalet, Самлет
RTKMP	ростелеком, телеком, rostelecom, telecom, rtkm, ртк, r-telecom, растелком
HYDR	русгидро, rushydro, rshydro, r-gidro, гидрорус, гидра, русгидра

На рисунке 4 приведена диаграмма распределения новостных статей для компаний после фильтрации.

В качестве векторизатора новостного потока на русском языке мы применили две модели: RuBert [8] и Qwen [9].

Работая с новостным потоком мы столкнулись с двумя проблемами. Первая проблема – это проблема рерайтинга, поэтому необходима фильтрация дублирующих новостей. Для того чтобы наша модель учитывала новость только один раз, необходимо реализовать алгоритм идентификации дубликатов.

Вторая проблема – выявление активов, на которые оказывает влияние конкретная новость. Данную проблему можно интерпретировать как

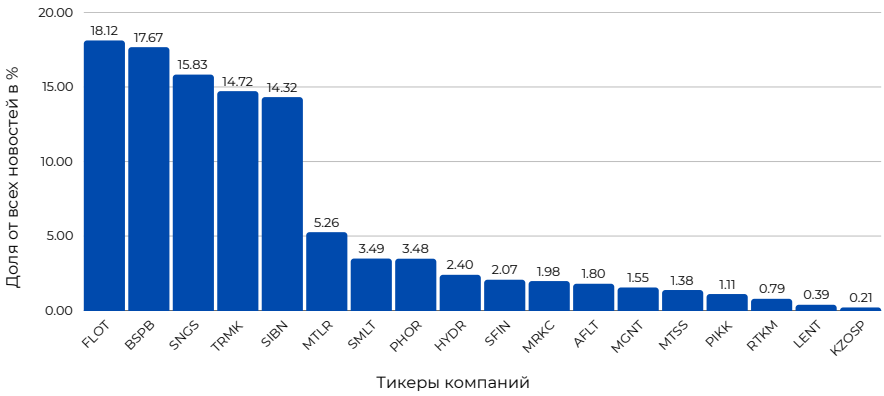

Рисунок 4. Распределение новостей по компаниям после фильтрации (цифры на диаграмме – процент новостей про компанию в наборе данных)

задачу классификации новости, для которой тикеры выступают в качестве меток классов.

Для решения проблемы рерайтинга мы спроектировали сиамскую нейронную сеть. Для этого мы составили обучающий набор данных с помощью API GigaChat следующим образом: для каждой статьи генерировались три перефразированных заголовка и тела статьи, далее случайным образом составлялись в равном соотношении пары из оригинальной и перефразированных новостей и их заголовков.

Сиамская нейронная сеть спроектирована следующим образом: на вход подается пара новостей, с помощью модели RuBert [8] извлекаются векторные представления новостей, над обоими векторами выполняется операция конкатенации, итоговый вектор передается далее в полносвязанную нейронную сеть (MLP). Для того чтобы определить оптимальную глубину модели MLP мы провели серию экспериментов, в ходе которой оценивались качество предсказания и время обработки новостного потока. В результате проведенных экспериментов мы выбрали глубину MLP в 3 слоя.

Затем отфильтрованные новостные статьи переводятся в векторы, чтобы при поступлении новых новостей классификацию дубликатов проводить в One-Shot режиме, что позволяет снизить затраты на время обработки новостного потока и вычислительные ресурсы (в нашем случае GPU V100).

Для решения второй проблемы – сопоставления выборок новостных статей по датам и их последующего использования в прогнозировании цен – необходимо формализовать процесс отбора данных и генерации прогнозов. Мы предполагаем, что предсказание цены закрытия актива производится для каждого торгового дня на момент открытия торгов. При этом в выборку включаются только те новости, которые были опубликованы до начала текущего торгового дня.

Формирование выборок осуществляется путем группировки новостей по дате публикации. Для прогноза цены на заданный день используются материалы, опубликованные в предшествующий торговый день. Например, статьи аналитического характера, такие как материалы под рубрикой «технический анализ» от источника «БКС», публикуемые ежедневно до начала торгов, включаются в выборку для прогнозирования цен активов, которые проанализированы в сообщении. Такой подход позволяет учитывать наиболее актуальную информацию и улучшить качество предсказания.

Для использования двух модальностей обучающие серии сформированы через конкатенацию векторов приростов цен за предыдущие пять дней и векторов новостного потока. Векторы относительных приростов цен конструировались аналогично эксперименту с одной модальностью, а новости выбирались за предыдущий торговый день в соответствии с выбранным активом. Далее эти новости отображались в векторы и агрегировались.

Если за предыдущий день или до открытия торгов текущего дня публикаций не было, то к вектору относительных приростов конкатенируется нулевой вектор длины 768 для модели RuBert и 896 для модели Vikhr-Qwen2.5-0.5b-Instruct (Qwen), иначе добавляется агрегированный вектор новостного потока той же длины. Такие длины конечных векторов соответствуют размерам оригинального выходного слоя предобученных моделей RuBert и Qwen.

В исследовании мы рассмотрели два варианта агрегации новостных векторов: сумма векторов (Sum) и усредненная сумма (Mean). Под суммой векторов мы подразумеваем суммирование значений соответствующих координат векторов. Под усредненной суммой мы подразумеваем, что в соответствующих координатах вектора выставляется среднее арифметическое значение координат агрегируемых векторов.

Базовая модель RuBert имеет ограниченный размер контекстного окна, равный 512 токенам. По этой причине, статьи, превышающие лимит

контекстного окна, усекались или дробились для раздельной обработки, поэтому одной новостной статье могли соответствовать более одного вектора. Модель Qwen имеет размер контекстного окна – 32768 токенов (в 64 раза больше), это достаточно для обработки статей без усечений. Далее мы сравниваем, как влияет на качество предсказание цен векторизатор новостного потока.

Поточечно предсказанные вектора доходностей переводились в цены активов по формуле (2). Качество предсказания оценивалось по двум метрикам: точность (Accuracy) и величина среднего абсолютного отклонения в процентах (MAPE). Точность оценивалась как доля верно предсказанных знаков значений элементов вектора доходностей: положительный или отрицательный. Метрика MAPE демонстрирует, насколько процентов в среднем предсказанная стоимость отличается от истинного значения. Таким образом, мы можем оценить качество предсказания в денежных единицах (рублях).

3. Вычислительный эксперимент

В данном разделе приведем результаты вычислительных экспериментов для двух предсказательных моделей (одна и две модальности). Для разработки предсказательной модели мы использовали фреймворк Transformers (платформа Hugging Face). Для проведения вычислений была использована видеокарта V100.

3.1. Результаты использования одной модальности

Результаты эксперимента предсказания векторов доходностей инструментов только на временных рядах для моделей классического и глубокого машинного обучения представлены в таблице 10.

В таблице 11 приведены усредненные оценки качества предсказания по моделям, данные в этой таблице отсортированы в порядке возрастания средней величины абсолютной ошибки отклонения предсказания от цены актива в процентах (столбец «Отклонение»).

Из результатов эксперимента видно, что рекуррентная модель LSTM демонстрирует наилучшее качество классификации, то есть предсказывает рост или падение, и регрессии – наименьшее среднее отклонение прогнозируемой цены от истинной, но при этом отстает по метрике средней абсолютной ошибки.

Таблица 10. Результаты предсказания векторов доходности с использованием только временных рядов. Точность (слева) и отклонение (справа) в процентах

	Источник	LSTM		XGB		KNN		RF		LinReg		DT	
Металлы и добыча	MTLR	**56.364**	**0.410**	40.000	2.089	42.273	2.050	50.909	2.020	50.000	2.029	42.727	2.679
	TRMK	**56.364**	**0.362**	40.909	2.105	38.182	2.167	47.273	2.154	49.091	2.114	52.727	2.308
Нефть и газ	SNGS	50.303	**0.352**	49.091	1.776	48.182	1.775	50.000	1.735	**60.909**	1.744	52.727	1.857
	SIBN	**58.182**	**0.341**	40.000	1.766	**58.182**	1.746	46.364	1.788	41.818	1.839	51.818	1.813
Потребительский сектор	MGNT	46.667	**0.331**	39.091	1.517	43.636	1.493	49.091	1.519	40.000	1.709	**60.000**	1.672
	LENT	**56.364**	**0.371**	54.546	2.202	39.091	2.178	52.723	2.145	51.818	2.220	51.818	2.589
Строительство	PIKK	49.091	**0.484**	40.909	1.565	50.909	1.563	50.000	1.558	44.545	1.637	**51.818**	1.592
	SMLT	**53.939**	**0.328**	42.727	1.577	38.182	1.552	46.364	1.539	49.091	1.536	41.818	1.683
Телекоммуникации	MTSS	**56.970**	**0.541**	42.727	1.290	40.000	1.306	45.455	1.520	53.636	1.419	50.000	1.395
	RTKMP	**55.152**	**0.246**	45.455	1.299	42.723	1.303	42.727	1.335	50.909	1.355	48.182	1.411
Транспорт	AFLT	55.152	**0.419**	46.364	2.079	**57.273**	2.017	52.727	2.062	60.909	1.976	51.818	2.194
	FLOT	47.273	**0.258**	43.637	2.116	38.182	2.124	42.727	2.104	45.454	2.074	**49.091**	2.294
Финансы	BSPB	46.061	**0.410**	49.091	1.612	50.909	1.695	50.909	1.598	**54.545**	1.602	45.455	1.829
	SFIN	**49.697**	**0.447**	40.000	1.603	30.909	1.647	39.091	1.743	48.182	1.960	41.818	1.959
Химическая промышленность	PHOR	41.818	**0.231**	42.727	1.194	**52.723**	1.149	48.182	1.168	50.000	1.227	45.455	1.218
	KZOSP	**57.576**	**0.458**	49.091	1.198	42.723	1.237	49.091	1.210	46.364	1.217	54.545	1.581
Электроэнергетика	HYDR	59.394	**0.380**	51.182	1.124	**60.000**	1.130	48.182	1.214	45.455	1.151	49.091	1.355
	MRKC	40.000	**0.768**	51.182	1.182	49.091	1.225	54.545	1.214	50.000	1.224	**55.455**	1.403

Таблица 11. Результаты предсказания с использованием одной модальности (временные ряды)

Модель	Точность, %	Отклонение, %
LSTM	**52.020**	**0.397**
XGB	45.000	1.627
KNN	46.010	1.631
RF	48.384	1.646
LinReg	50.152	1.669
DT	49.798	1.824

3.2. Результаты использования двух модальностей

Результаты второго эксперимента, состоявшего в слиянии новостного потока и числовых временных рядов и сравнении предложенного мультимодального подхода с прогнозом построенным исключительно на временных ряда свечей активов, представлены в таблице 12. В таблице 13 представлены усредненные значения метрик предсказания по рассмотренным моделям, данные в таблице отсортированы по столбцу «Отклонение» – в порядке возрастания средней величины абсолютной ошибки отклонения предсказания от цены актива в процентах.

Базовой моделью во втором эксперименте была выбрана нейронная сеть LSTM. С ней мы провели сравнение различных векторизаторов (RuBert, Qwen) и методов агрегации векторов (Sum, Mean).

На рисунке 5 приведена зависимость величин функции ошибки среднеквадратичного отклонения (MSE Loss) от количества итераций обучения для различных моделей для тренировочного (с 7 июля 2022 года по 27 марта 2024 года) и тестового наборов (с 28 марта по 30 августа 2024 года). По графику видно, что после 30 эпох обучения кривые выходят на стационарное значение.

Таблица 12. Результаты предсказания векторов доходности с использованием двух модальностей. Точность (слева) и отклонение (справа) в процентах

	Источник	vanilla LSTM		LSTM_RuBert_SUM		LSTM_RuBert_MEAN		LSTM_QWEN_SUM		LSTM_QWEN_MEAN	
Металлы и добыча	MTLR	**56.364**	0.410	39.394	0.409	38.788	0.410	45.455	0.522	52.121	**0.246**
	TRMK	**56.364**	0.362	35.152	0.392	42.424	**0.192**	36.364	0.504	35.758	0.419
Нефть и газ	SNGS	50.303	0.352	53.939	0.865	**58.182**	1.824	44.848	0.307	49.697	**0.106**
	SIBN	**58.182**	0.341	**58.182**	0.265	**58.182**	0.216	39.394	0.368	47.879	**0.165**
Потребительский сектор	MGNT	46.667	0.331	**53.333**	0.417	47.879	0.299	46.061	0.307	48.485	**0.235**
	LENT	**56.364**	0.371	49.091	0.400	50.909	0.359	53.333	0.346	52.121	**0.331**
Строительство	PIKK	49.091	0.484	50.303	0.462	**57.576**	0.436	47.273	0.529	53.333	**0.322**
	SMLT	**53.939**	0.328	38.788	**0.200**	46.061	0.270	36.364	0.311	43.030	0.241
Телекоммуникации	MTSS	**56.970**	0.541	53.939	0.473	55.152	0.368	47.879	0.316	45.455	**0.193**
	RTKMP	**55.152**	0.246	49.697	0.274	45.455	0.271	44.848	**0.171**	44.242	0.178
Транспорт	AFLT	**55.152**	0.419	51.515	0.641	50.303	0.348	45.455	0.259	52.121	**0.182**
	FLOT	47.273	**0.258**	43.636	0.532	**52.121**	0.262	43.636	0.392	43.636	0.345
Финансы	BSPB	46.061	0.410	47.879	0.406	50.909	0.326	47.879	0.369	**52.121**	**0.227**
	SFIN	49.697	0.447	44.848	0.445	47.273	0.390	**56.970**	**0.195**	**56.970**	0.272
Химическая промышленность	PHOR	41.818	0.231	53.333	0.264	55.152	0.238	**60.000**	0.354	44.848	**0.219**
	KZOSP	**57.576**	0.458	42.424	0.492	41.212	0.491	48.485	0.369	49.697	**0.352**
Электроэнергетика	HYDR	59.394	0.380	58.788	0.326	55.758	0.321	47.879	0.292	**61.212**	**0.178**
	MRKC	40.000	0.768	42.424	0.742	**43.030**	0.839	42.424	0.660	41.818	**0.543**

Таблица 13. Результаты предсказания с использованием двух модальностей

Модель	Точность, %	Отклонение, %
LSTM-Qwen-Mean	48.552	**0.256**
LSTM-Qwen-Sum	46.970	0.367
LSTM	**52.020**	0.397
LSTM-RuBert-Mean	49.798	0.437
LSTM-RuBert-Sum	48.148	0.445

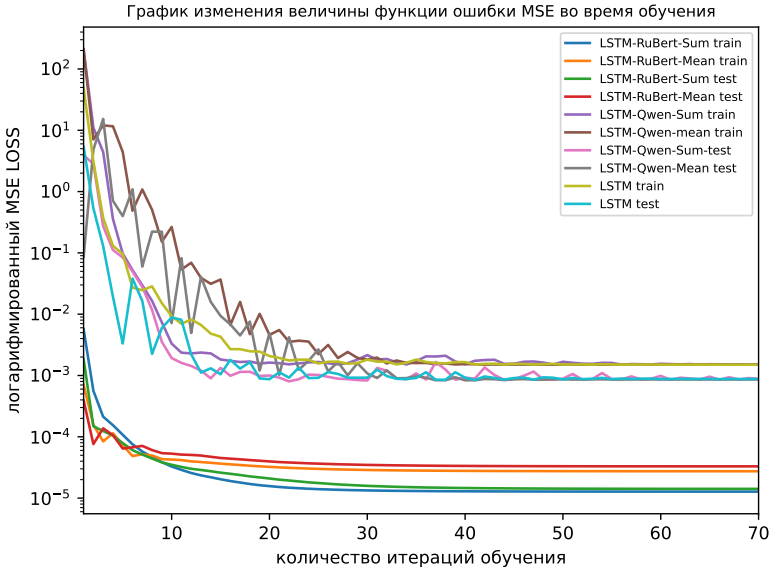

Рисунок 5. Зависимость величин функции ошибки среднеквадратичного отклонения от количества итераций обучения для различных моделей. Тренировочный и тестовой наборы

Из таблиц результатов следует, что прогноз построенный на векторизованном новостном потоке с помощью большой языковой модели превосходит прогноз, построенный исключительно на свечных данных активов, демонстрируя наименьшее значение отклонения поточечного прогноза цены от истинного вектора цен. При этом, усреднение векторов (Mean) дает наилучший результат.

Набор данных (176 акций российских компаний, торгуемых на Московской бирже, и 79555 русскоязычных финансовых новостей), собранный для проведения исследований, доступен по ссылке [11].

Заключение

В результате проведенных экспериментов мы продемонстрировали, что добавление текстовой модальности – анализ новостного потока – положительно влияет на качество предсказания цены. В среднем значение метрики MAPE (отклонение прогнозируемой цены от истинной) уменьшается на 55%: с 0.397 (модель LSTM) до 0.256 (модель LSTM-Qwen-Mean). К тому же качество предсказания на основе векторов, полученных с помощью большой языковой модели Vikhr-Qwen2.5-0.5b-Instruct, оказалось лучше, чем RuBert. От части это следует из того факта, что модель Qwen обладает бóльшим контекстным окном и обучена на бóльшем корпусе текста с поддержкой «цепочки размышлений» (Chain-of-Thoughts, CoT), что улучшает способность модели размышлять и улавливать сложные семантические зависимости внутри текста. Из результатов эксперимента следует, что метод усреднения векторов (Mean) оказался лучше, чем суммирование (Sum), и является предпочтительным методом агрегации векторов новостного потока.

При этом, важно заметить, что тестовые данные, на основе которых рассчитывались финальные значения метрик, включают интервал с 28 марта по 30 августа 2024 года. На этот временной интервал у рынка российских ценных бумаг наблюдается общая тенденция к спаду. Наличие явного тренда является важным фактором, упрощающим задачу предсказания. Однако, даже в этой постановке, предложенный мультимодальный подход оказался наилучшим среди рассмотренных.

Обучение и валидация модели для решения задачи рерайтинга происходило на новостных статьях, объем которых не превосходил контекстное окно модели RuBert, потому артефакты, связанные с величиной контекстного окна, выявились только во время этапа построения прогнозов, когда в новостную выборку попадали статьи длиной в среднем около 290 слов. Потому на будущее, для улучшения модели фильтрации и классификации новостей по компаниям необходимо воспользоваться моделями с бóльшим контекстным окном, например Qwen.

Собранный набор данных [11] демонстрирует хорошую структурированность и может быть применен для тонкой настройки русскоязычных и адаптированных под русский язык больших языковых моделей для применения в финансовой сфере.

Для количественного сравнения предложенной модели мы провели вычислительный эксперимент, в котором за основу взяли подход и метрики статьи [7]. Следуя работе [7], в качестве набора данных мы использовали временные ряды цен пяти крупных американских компаний: AAPL, AMZN, GOOGL NFLX, TSLA и набор англоязычных новостей с разметкой по компаниям за период с 12 октября 2012 года по 31 января 2020 года. (таблица 14).

Таблица 14. Сравнение метрик прогнозирования мультимодального подхода (LSTM-Qwen-Mean) с подходом, основанным на учете оценки тональности новостей (Baseline) [7]

Модель	Тикер	R2	MAPE, %	MAE
LSTM-Qwen-Mean	AAPL	**0.989**	**0.628**	**0.003**
Baseline	AAPL	0.947	2.333	0.018
LSTM-Qwen-Mean	AMZN	**0.968**	**1.601**	**0.013**
Baseline	AMZN	0.870	1.730	0.015
LSTM-Qwen-Mean	GOOGL	**0.935**	**1.394**	**0.008**
Baseline	GOOGL	0.788	2.286	0.020
LSTM-Qwen-Mean	NFLX	**0.955**	**2.361**	0.076
Baseline	NFLX	0.919	2.512	**0.019**
LSTM-Qwen-Mean	TSLA	0.915	**3.206**	**0.006**
Baseline	TSLA	**0.930**	7.423	0.034

Заметим, что использованный набор данных включает текстовую составляющую на английском языке, поэтому для векторизации новостей мы использовали оригинальную модель Qwen2.5-0.5b-Instruct [10]. Для построения прогнозов мы выбрали и обучили модель *LSTM-Qwen-Mean*, так как она показала в среднем лучшее качество в нашем исследовании. В качестве метрик мы использовали коэффициент детерминации ($R2$), среднюю абсолютную ошибку (MAE) и среднюю абсолютную ошибку в процентах ($MAPE$).

Таким образом, мы работали с одними и теми же временными рядами и метриками. По всем метрикам, за исключением MAE для NFLX и $R2$ для TSLA, предложенный мультимодальный подход с усреднением векторов превосходит метрики наилучшего запуска подхода [7]. На основании проведенных вычислений можно сделать вывод, что предложенный мультимодальный подход продемонстрировал лучшее качество прогноза и универсальность применительно к российскому и зарубежному рынку.

В дальнейшем, необходимо исследовать каким образом учитывать входной новостной поток в предсказательной модели – за какой период времени необходимо использовать новости и как учитывать новостные сообщения (например, варьировать вес новости в зависимости от ее хронологического порядка в выборке).

Список использованных источников

[1] Mishev K., Gjorgjevikj A., Vodenska I., Chitkushev L., Trajanov D. *Evaluation of sentiment analysis in finance: from lexicons to transformers* // IEEE Access.– 2020.– Vol. **8**.– Pp. 131662–131682. ↑107

[2] Ho T.-T., Huang Y. *Stock price movement prediction using sentiment analysis and CandleStick chart representation* // Sensors.– 2021.– Vol. **21**.– No. 23.– id. 7957.– 18 pp. ↑107

[3] Jaggi M., Mandal P., Narang S., Naseem U., Khushi M. *Text mining of stocktwits data for predicting stock prices* // Applied System Innovation.– 2021.– Vol. **4**.– No. 1.– id. 13.– 22 pp. ↑107

[4] Fazlija B., Harder P. *Using financial news sentiment for stock price direction prediction* // Mathematics.– 2022.– Vol. **10**.– No. 13.– id. 2156.– 20 pp. ↑107

[5] Xinli Y., Zheng Ch., Yuan L., Shujing D., Zongyi L., Yanbin L. *Temporal data meets LLM — Explainable financial time series forecasting.*– 2023.– 13 pp. arXiv 2306.11025 [cs.LG] ↑107

[6] Boyu Zh., Hongyang Y., Liu X.-Y. *Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models.*– 2023.– 7 pp. arXiv 2306.12659 [cs.CL] ↑107

[7] Куликова Т. Д., Ковтун Е. Ю., Буденный С. А. *Получаем ли мы пользу от категоризации потока новостей в задаче прогнозирования цен акций?* // Доклады Российской академии наук. Математика, информатика, процессы управления.– 2023.– Т. **514**.– № 2.– С. 385–394. ↑107, 127

[8] Kuratov Y., Arkhipov M. *Adaptation of deep bidirectional multilingual transformers for Russian language.*– 2019.– 8 pp. arXiv 1905.07213 [cs.CL] ↑109, 118, 119

[9] Nikolich A., Korolev K., Shelmanov A., Kiselev I. *Vikhr: The family of open-source instruction-tuned large language models for Russian.*– 2024.– 8 pp. arXiv 2405.13929 [cs.CL] ↑109, 118

[10] Yang A., Yang B., Hui B., Zheng B., Yu B., Zhou Ch., Li Ch., Li Ch., Liu D., Huang F., Dong G., Wei H., Lin H., J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, Sh. Wang, Sh. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou,

X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Zh. Zhang, Zh. Guo, Zh. Fan *Qwen2 Technical Report.*– 2024.– 26 pp. arXiv 2407.10671 [cs.CL] ↑127

[11] Khubiev K. *Russian financial news dataset.*– Kaggle Platform.– 2025. ↑126

Поступила в редакцию 24.12.2024;
одобрена после рецензирования 30.01.2025;
принята к публикации 27.02.2025;
опубликована онлайн 11.03.2025.

Рекомендовал к публикации *к.т.н. Е. П. Куршев*

Информация об авторах:

Касымхан Юсуфович Хубиев

исследователь в центре социально-экономического прогнозирования, магистрант направления «Финансовая математика и финансовые технологии», Университет «Сириус». Научные интересы: искусственные интеллект и его приложения в науке, финансах, промышленности и бизнесе.

0009-0007-1719-1455
e-mail: *kasymkhankhubievnis@gmail.com*

Михаил Евгеньевич Семенов

к.ф.-м.н., научный руководитель направления «Финансовая математика и финансовые технологии», Университет «Сириус», Научные интересы: информационные технологии, интеллектуальные технологии обработки и анализа данных.

0000-0002-0716-5065
e-mail: *semenov.me@talantiuspeh.ru*

Авторы внесли равный вклад в подготовку публикации.

Декларация об отсутствии личной заинтересованности: *благополучие авторов не зависит от результатов исследования.*