

МАТЕМАТИЧЕСКИЙ АНАЛИЗ ЭКОНОМИЧЕСКИХ МОДЕЛЕЙ

Прогнозирование доходности российских акций на основе анализа сентимента инвесторов в социальных сетях

© 2025 г. Г.А. Хазиев, Т.В. Соколова

Г.А. Хазиев,

НИУ «Высшая школа экономики», Москва; e-mail: gakhaziev@hse.ru

Т.В. Соколова,

НИУ «Высшая школа экономики», Москва; e-mail: tv.sokolova@hse.ru

Поступила в редакцию 26.04.2024

Исследование выполнено за счет гранта Российского научного фонда № 22-18-002760, <https://rscf.ru/project/22-18-00276/>

Аннотация. В работе исследуется сентимент российских частных инвесторов в социальных сетях и его влияние на динамику доходности акций 78 компаний российского рынка в 2018–2022 г. Для учета сентимента при прогнозировании цен используется авторский индекс RSMI (Russian social media index), который строится на уникальной выборке сообщений из наиболее популярных у российских инвесторов социальных сетей — «Телеграм» и «Тинькофф Пульс». Индекс RSMI включает количественные (число публикаций в отношении каждой компании) и качественные (реакции инвесторов) характеристики, позволяющие определить реальное влияние той или иной публикации на инвесторов. С использованием индекса RSMI построены модели прогнозирования цен акций российских компаний методами регрессии «лассо» (lasso), «случайного леса», градиентного бустинга, экстремального градиентного бустинга, ансамблевого обучения и рекуррентной нейронной сети (LSTM). Показано, что для акций широкой выборки индикаторы технического анализа и рыночные мультипликаторы играют большую роль в построении прогнозов изменения доходности акций на часовых данных. Хотя добавление индекса сентимента и позволяет улучшить результаты прогнозирования доходности для акций широкой выборки, это не дает значительного улучшения предсказательной способности моделей и вызывает разнонаправленные результаты. Наилучшие результаты добавление индекса сентимента в прогнозные модели показывает для топ-15 наиболее обсуждаемых российских компаний. Для отдельных моделей удалось добиться среднего снижения ошибок на 4,9%, а для отдельных компаний более чем на 10% уменьшить показатель ошибки MAE и на 20% MSE. Доказано, что на динамику доходности акций второго и третьего эшелона российского фондового рынка сентимент частных инвесторов на часовых данных не оказывает существенного влияния, а добавление индекса сентимента не позволяет улучшить результаты прогнозных моделей.

Ключевые слова: акции, фондовый рынок, машинное обучение, сентимент инвесторов.

Классификация JEL: E44, G12, G14.

УДК: 336.763.2.

Для цитирования: **Хазиев Г.А., Соколова Т.В.** (2025). Прогнозирование доходности российских акций на основе анализа сентимента инвесторов в социальных сетях // *Экономика и математические методы*. Т. 61. № 1. С. 95–108. DOI: 10.31857/S0424738825010095

ВВЕДЕНИЕ

За последние несколько лет произошли значительные изменения в структуре участников российского фондового рынка. Ранее основной объем торгов приходился на профессиональных участников, но с 2020 г. начался активный рост доли частных инвестиций, который со временем привел к их доминирующему влиянию на рынке. По итогам февраля 2024 г. число физических лиц, имеющих брокерские счета на Московской бирже (МОЕХ), превысило 30,7 млн, ими открыто 54,3 млн счетов. За период 2021–2023 гг. частные инвесторы вложили в рынок акций более 950 млрд руб., при этом доля частных инвесторов в общем объеме торгов акциями постоянно растет. По итогам 2021 г. она составляла 40%, в 2022 г. — 59%, а по итогам 2023 г. выросла до 79%¹.

¹ Официальный сайт Московской биржи. Новости и пресс-релизы. «Московская биржа подвела итоги торгов в феврале 2024 года» (<https://www.moex.com/n39552>, <https://www.moex.com/n53950>).

Одновременно с изменением структуры рынка появились и новые факторы, способные влиять на изменение доходности торгуемых активов. Главную роль по-прежнему играют экономические и геополитические факторы, но наряду с ними значительное влияние на динамику торгов оказывает *настроение инвесторов* (сентимент) (Cai, Tang, Chen, 2024; Gupta, Nel, Pierdzioch, 2023).

Изучение сентимента частных инвесторов в современных работах по теме финансовой экономики ведется преимущественно по двум направлениям: объяснение и прогнозирование изменения доходности акций (Cai, Tang, Chen, 2024; Gu, Kurov, 2020; Gupta, Nel, Pierdzioch, 2023) и выявление манипуляций на фондовом рынке (Uslu, Akal, 2021; Tallboys, Zhu, Rajasegarar, 2022).

В рамках текущего исследования мы развиваем направление прогнозирования доходности акций с учетом сентимента частных инвесторов. В работе предлагается оригинальный подход к построению индекса (метрики) сентимента российских частных инвесторов, включая использование моделей классификации сентимента, адаптированных под русский язык, а также учет эмодзи, тональности реакций на сообщения, числа просмотров и пересылок сообщений другим участникам.

По сравнению с другими исследованиями (например, (Chong, Han, Park, 2017; Gupta, Nel, Pierdzioch, 2023)), во многих из которых в прогнозную модель наряду с сентиментом включается только изменение биржевых показателей акций, нами для повышения точности модели дополнительно используются индикаторы технического анализа и рыночные мультипликаторы.

Поскольку выводы предыдущих работ относительно возможности сентимента прогнозировать доходность акций противоречивы (Liu, Lee, Huang, Wu, 2023; Li et al., 2023), а в отдельных работах исследователи приходят к выводу о неэффективности применения данного индикатора, актуальным представляется дальнейшее изучение данного вопроса с использованием современных данных на российском рынке акций. Нами рассматривается период более длительный, чем в большинстве предыдущих работ (Audrino, Sigrist, Ballinari, 2020; Lin, Ma, Fildes, 2024), а также более короткий часовой таймфрейм и большой набор данных (Xu, Xue, Zhang, 2019; Liang et al., 2020), в совокупности эти факторы позволяют получить более достоверные результаты.

ОБЗОР ИССЛЕДОВАНИЙ ПО ТЕМЕ

За последние годы было проведено множество исследований влияния сентимента инвесторов в социальных сетях на динамику акций компаний, но все они преимущественно посвящены двум крупнейшим рынкам — США (Mili et al., 2024; Sarkar et al., 2022) и Китая (Niu, Pan, Xu, 2023; Liu, Leu, Holst, 2023). Работы, исследующие российский фондовый рынок, носят единичный характер (Теплова и др., 2022; Teplova, Tomtosov, Sokolova, 2022), что позволяет нам заполнить существующий пробел в исследованиях.

Традиционно в рамках исследований XX в. динамика фондового рынка объяснялась гипотезой эффективного рынка (effective market hypothesis, ЕМН) и теорией случайных блужданий. Указанные концепции не рассматривали настроения инвесторов как важный аспект, однако эти концепции не смогли объяснить неоднородного и иррационального поведения инвесторов на рынке капитала, а также отклонения наблюдаемых цен активов от справедливых, вызванных влиянием сентимента инвесторов (Ph, Rishad, 2020). В XXI в. развитие получили поведенческие финансовые теории, согласно которым иррациональное поведение шумовых трейдеров и арбитражеров приводит к несоответствию цен на активы их внутренним (фундаментальным) значениям. Согласно (Baker, Wurgler, 2007) поведение иррациональных инвесторов, принимающих решения не на основе фундаментальных факторов, а на основе рыночного сентимента, отражается в ценах акций.

Важным вопросом в рамках прогнозирования доходности акций является роль сентимента в ценообразовании акций. Ответу на этот вопрос посвящено много исследований (Li et al., 2020; Gu, Kurov, 2020; и др.), но особый интерес представляет работа (Cai, Tang, Chen, 2024). Авторы анализируют влияние настроений инвесторов в режиме реального времени на высокочастотную доходность акций фондового рынка Китая и приходят к выводу, что публикации частных инвесторов в социальных сетях могут оказывать существенное влияние на трехминутную доходность акций. Авторы строят прогнозную модель доходности акций с учетом сентимента. Наибольшее влияние сентимента проявляется в периоды низкой волатильности акций, когда отсутствуют другие рыночные сигналы на покупку или продажу, с ростом волатильности влияние сентимента ослабляется.

С точки зрения методологии в работах обычно используется одна модель (Liu et al., 2023), на основании которой принимается решение об эффективности прогнозирования цен активов с учетом

тональности. В нашей работе задействовано сразу шесть моделей искусственного интеллекта, что позволяет делать более достоверные выводы, а также сопоставить модели по метрикам качества и определить наиболее эффективную модель для прогнозирования доходности акций российских компаний на основе сентимента.

В исследованиях при выборе переменных для прогнозирования доходности наряду с тональностью сообщений обычно используются только цены и объемы торгов акциями (Ahmed, Neema, Visqanadha, 2022), в нашей работе мы учитываем дополнительно технические индикаторы и рыночные мультипликаторы.

Также важным отличием от большинства предыдущих исследований (табл. 1) является короткий часовой таймфрейм, который позволяет проверить наличие/отсутствие влияния тональности сообщений в режиме реального времени, а не на дневных или недельных данных, как это делается в работах других авторов (Xu et al., 2019; Liang et al., 2020; Teplova, Tomtosov, Sokolova, 2022).

Таблица 1. Обзор и ключевые результаты работ, посвященных исследованию влияния сентимента на динамику рыночных активов

Работа	Подход к построению индекса (переменных) сентимента	Выборка исследования	Модель	Ключевые результаты
Gao, Xie, 2020	Общерыночный индекс сентимента, строится методом главных компонент на основе пяти прокси-переменных: а) объем открытых позиций фьючерсов на индекс акций CSI 300; б) дисбаланс купли-продажи по сделкам, инициированным на фондовом рынке; в) индекс Psychological Line на индекс акций CSI 300; г) объем спотовых торгов для индекса акций CSI 300; д) скорректированная скорость оборачиваемости индекса акций CSI 300	Рынок фьючерсов на фондовый индекс CSI 300 (Китай), 2010–2014, дневные данные	Модель векторной авторегрессии	Построенный интегральный индекс настроений позволяет прогнозировать избыточную доходность и аномальные объемы торгов на рынке фьючерсов на фондовый индекс
Wang, Yu, Shen 2020	Индекс сентимента по отдельным компаниям — эмитентам акций, основанный на анализе тональности текстовых сообщений инвесторов на платформе East Money: $S_{k,t} = \ln \left[\frac{(1 + R_{k,t}^{pos})}{(1 + R_{k,t}^{neg})} \right],$ где $R_{k,t}^{pos}$ и $R_{k,t}^{neg}$ — число прочтений позитивных и негативных сообщений на платформе East Money по каждой компании-эмитенту акций i в день t . Для классификации сообщений на позитивные и негативные используются LSTM-модели	125 компаний, котирующихся на Шанхайской фондовой бирже, 2014–2018 гг., дневные данные	МНК регрессии на панельных данных	Индекс сентимента значимо положительно влияет на доходность акций и объем торгов на следующий день. Влияние сентимента на торговый дисбаланс (фиксируемый по соотношению заявок на покупку и продажу акций) также значимо и положительно
Теплова и др., 2022	Авторский индикатор сентимента по отдельным компаниям — эмитентам акций, основанный на анализе текстовых сообщений инвесторов по форуму MFD и каналам «Телеграм» инвестиционной тематики: $HYPE_t = \frac{MSG_t^{POSITIVE}}{MSG_t^{ALL}} \left(\frac{MSG_t^{ALL}}{MSG_{t-1}^{ALL}} - 1 \right),$ где $MSG_t^{POSITIVE}$ — число позитивных сообщений по каждой компании; MSG_t^{ALL} — суммарное число позитивных и негативных сообщений о каждой компании (значения индикатора HYPE ограничены шкалой от 0 до 100). Также тестировалось влияние количества сообщений позитивной, нейтральной и негативной тональности. Для классификации сообщений на позитивные и негативные применялась модель нейронной сети	60 компаний российского рынка, с 2014 по 2020 г.	Модели на панельных данных со случайными эффектами, с учетом нелинейного влияния переменных сентимента	Авторский индикатор HYPE оказывает положительное и значимое влияние на доходность акций. Выявлено нелинейное влияние количества негативных сообщений на доходность акций (парабола ветвями вниз). Влияние квадрата количества позитивных сообщений значимо негативно

Таблица 1. Окончание

Работа	Подход к построению индекса (переменных) сентимента	Выборка исследования	Модель	Ключевые результаты
Niu, Pan, Xu, 2023	Индексы сентимента, которые строятся на основе классификации сообщений инвесторов на форуме Eastern Fortune по 4 типам эмоций — счастье, гнев, страх, печаль — с применением различных моделей глубокого обучения. Лучшими метриками качества для решения задачи классификации обладает модель Bi-LSTM-ATT. Каждому дню по каждой компании — эмитенту акций приписывается преобладающая эмоция (сентимент)	5 компаний — эмитентов акций на рынке Китая из разных секторов экономики, 2011–2021 гг., дневные данные	Модели глубокого обучения для прогнозирования цен на акции. Лучшими метриками качества характеризуется модель CNN-Bi-LSTM-ATT	Учет сентимента инвесторов в моделях прогнозирования цен акций позволяет повысить точность прогнозов по сравнению с моделями, в которых учитываются только данные о торгах

2. ГИПОТЕЗЫ НАШЕГО ИССЛЕДОВАНИЯ

Согласно гипотезе эффективного рынка, выдвинутой (Fama, 1965), доступная инвесторам информация о компании находит отражение в ценах акций. Настроение инвесторов также оказывает влияние на динамику цен, поскольку психологический фактор влияет на инвестиционные решения финансовых агентов (Shiller, 2003). В предыдущих эмпирических исследованиях выявлено значимое влияние сентимента частных инвесторов в социальных сетях на доходность акций как для развитых, так и для развивающихся рынков, включая российский (Wang, Yu, Shen, 2020; Теплова и др., 2022; Niu, Pan, Xu, 2023).

Авторы (Chen, Ge, 2019) интегрировали механизмы внимания в нейросеть LSTM, что позволило улучшить качество моделей прогнозирования динамики доходности акций 72 компаний на рынке Гонконга. В работе (Niu, Pan, Xu, 2023) на дневных данных рынка Китая показано, что для прогнозной модели CNN-Bi-LSTM-ATT, учитывающей настроения инвесторов, характерны лучшие метрики качества (точность прогнозов) по сравнению с учетом только данных о торговых транзакциях. В то же время улучшение метрик качества моделей за счет учета сентимента оказалось незначительным.

Мы предполагаем, что эффект влияния сентимента на доходность акций российского рынка на часовых данных выражен в меньшей степени по сравнению с дневными. Среднестатистический частный инвестор — клиент брокера — на российском рынке заключает сделки лишь два дня в месяц, и 83% частных инвесторов совершают менее пяти сделок в месяц².

В связи с вышесказанным мы выдвигаем следующие гипотезы.

Гипотеза 1. Индикаторы технического анализа и рыночные мультипликаторы играют большую роль в прогнозировании изменений доходности акций широкой выборки компаний российского рынка на часовых данных по сравнению с сентиментом частных инвесторов в социальных сетях.

Гипотеза 2. Учет сентимента частных инвесторов в социальных сетях по наиболее обсуждаемым российским компаниям—эмитентам акций позволяет прогнозировать изменение доходности акций с большей точностью по сравнению с учетом только индикаторов технического анализа и рыночных мультипликаторов.

В работе (Baker, Wurgler 2007) авторы выявили, что сентимент инвесторов на цены акций зависит от характеристик компаний-эмитентов. В частности, более чувствительны к изменениям сентимента цены акций компаний с небольшой рыночной капитализацией, молодых, убыточных компаний, акции с повышенной волатильностью и не приносящие дивидендов. Данный эффект можно объяснить тем, что этим классам компаний труднее привлечь финансирование в случае пессимизма инвесторов относительно их финансового положения и перспектив роста. В работе (Xu, Xue, Zhang, 2024) для китайского рынка показано, что сентимент более сильно влияет на ликвидность акций быстрорастущих компаний, поскольку их прогнозные денежные потоки более чувствительны к ожиданиям инвесторов.

² Источник: Банк России, 2022 г. «Портрет клиента брокера. В исследовании приняли участие 18 крупнейших брокеров (25,2 млн клиентов)» (https://www.cbr.ru/Content/Document/File/143859/Portrait_client_brok.PDF).

Поэтому мы выдвигаем следующую гипотезу.

Гипотеза 3. Доходность акций второго и третьего эшелона российского фондового рынка в большей степени изменяется под воздействием сентимента сообщений инвесторов в социальных сетях (по сравнению с акциями компаний первого эшелона).

3. МЕТОДОЛОГИЯ

Предложенный нами алгоритм исследования состоит из семи этапов.

1. Сбор почасовых торговых данных по акциям: цены открытия и закрытия, минимальная и максимальная цена, объем торгов, а также изменения каждого из указанных показателей. Выборка данных производилась по всем компаниям, включенным хотя бы в один из 14 индексов Московской биржи (подробнее см. разд. 4).

2. Сбор (парсинг) сообщений инвесторов из наиболее популярных у российских инвесторов социальных сетей — «Тинькофф Пульс» и «Телеграм»³.

3. Разметка данных, предварительная обработка и классификация сообщений по тональности (т.е. определение их сентимента). Такая обработка включала токенизацию (разбиение текста на токены — слова), очистку от стоп-слов (частицы, союзы, наречия и др.), лемматизацию (приведение слова к начальной форме), удаление знаков препинания и других символов (цифр, ссылок), а также дополнительную разметку тональности 26 наиболее часто встречающихся эмодзи (смайлов).

Классификация сообщений происходила по трем классам тональности (positive — позитивный, neutral — нейтральный, negative — негативный) с использованием нескольких популярных моделей машинного обучения (ML): Catboost, rugpt3large_based_on_gpt2, TextBlob, Dostoevsky, Bert-модели для русского языка (yandex/RuLeanALBERT, sberbank-ai/ruT5-large и др.). Лучшие результаты показала модель градиентного бустинга «CatBoost» от Яндекса: Accuracy = 0,75; Precision = 0,54 — positive, 0,77 — neutral, 0,49 — negative. Обучение моделей проводилось на основе готовой базы данных из 32 тысяч сообщений в социальных сетях с размеченными вручную настроениями⁴.

4. Разработан и применен авторский индикатор измерения сентимента настроений инвесторов «RSMI» (Russian social media index), представляющий собой сумму двух индексов «Telegram Index» (Tg Index) и «Pulse Index» (Pulse Index) и показывающий популярность и тональность обсуждения конкретной компании за определенный период.

Индексы рассчитываются по формулам:

$$TgIndex = \sum \left(Sentiment \left(\frac{NViews}{1000} \right) \left(1 + \frac{NLikes}{NViews \times EngageLevel} + \frac{NReposts}{Nviews \times RepostFreq} \right) \right), \quad (1)$$

где Σ — суммирование значений сентимента для каждого сообщения по отдельной компании — эмитенту акций за определенный период (час), *Sentiment* — величина, характеризующая тональность сообщения и равная 1, если сообщение имеет положительную окраску, и -1 , если отрицательную, *NViews* — число просмотров сообщения, *EngageLevel* — средний процент реакций по отношению к просмотрам в данной социальной сети, *NLikes* — разность числа положительных и отрицательных реакций, *NReposts* — число репостов, *RepostFreq* — средний процент переадресованных сообщений по отношению к просмотрам в данной социальной сети;

$$PulseIndex = \Sigma (Sentiment \times NLikes \times NComments), \quad (2)$$

где *NComments* — число комментариев к сообщению.

5. Рассчитываются технические индикаторы и рыночные мультипликаторы, изменение которых дает сигналы на покупку и продажу акций: цена/прибыль (P/E), цена/балансовая стоимость (P/B), экспоненциальная скользящая средняя EMA PE, EMA PB, EMA цены акции за 12 и 26 часов, индикатор схождения/расхождения скользящих средних (MACD), сигнальная линия MACD, индикатор моментум (Momentum), индикаторы Stochastic (стохастик) и индекс относительной силы (RSI).

³ Список каналов опубликован на сайте авторов: <https://fmlab.hse.ru/appendices>

⁴ База данных сообщений и метрик сентимента инвесторов на российском рынке акций. Номер РИД: 6.0020—2020. Сведения о регистрации: 2020622801. Авторы: А.Ф. Томтосов, Д.В. Бучко, Т.В. Теплова, Т.В. Соколова.

6. На основании технических индикаторов и рыночных мультипликаторов определяются сигналы на покупку или продажу акций.

Сигнал 1 (RSI): сигналом на продажу акций служит выход RSI из зоны перекупленности, т.е. пересечение уровня 70 сверху вниз; на покупку — выход RSI выходит из зоны перепроданности, т.е. пересечение уровня 30 снизу вверх.

Сигнал 2 (MACD): если MACD, построенный как разность 12 и 26 часовых скользящих средних, пересекает свою сигнальную линию снизу вверх, поступает сигнал на покупку акций, если сверху вниз — на продажу.

Сигнал 3 (Momentum): если линия момента пересекать отметку 100% сверху вниз — фиксируется сигнал на продажу акций, если снизу вверх — на покупку. Значение индикатора моментум берется за семь торговых часов.

Сигнал 4 (Stochastic): сигнал на покупку акций фиксируется, когда индикатор пересекает уровень 20 снизу вверх, на продажу — когда индикатор пересекает уровень 80 сверху вниз.

Сигнал 5 (EMA): сигнал на покупку акций фиксируется, когда 25-дневная ЕМА пересекает 100-дневную ЕМА снизу вверх; на продажу — когда 25-дневная ЕМА пересекает 100-дневную ЕМА сверху вниз.

Сигнал 6 (P/E): если мультипликатор P/E пересекает свою 250-дневную скользящую среднюю сверху вниз, фиксируется сигнал на покупку, в обратном случае — на продажу.

Сигнал 7 (P/B): если мультипликатор P/B пересекает свою 250-дневную скользящую среднюю сверху вниз, фиксируется сигнал на покупку, в обратном случае — на продажу.

7. Сигналы по каждой компании — эмитенту акций были объединены с индексом сентимента RSMI в единый датасет. Построены и оценены прогнозные модели искусственного интеллекта (ИИ): регрессия «лассо» (lasso), случайный лес (rfr), градиентный бустинг (gbr), экстремальный градиентный бустинг (XGBoost, xg), ансамблевое обучение (wel) и метод рекуррентной нейронной сети (LSTM).

При построении моделей использование сигналов технических индикаторов и рыночных мультипликаторов, а также индекса сентимента позволяет предоставить моделям как можно больше переменных для обучения для получения большей точности прогнозов. Особенностью применяемых моделей машинного обучения является невозможность точно определить, какой вклад те или иные факторы внесли в финальные результаты модели. Для оценки качества моделей ИИ выбраны три метрики:

1) MAE (средняя абсолютная ошибка) — метрика, которая показывает среднюю абсолютную разность между прогнозируемыми значениями и фактическими значениями в наборе данных. Чем ниже MAE, тем лучше модель соответствует набору данных;

Таблица 2. Переменные, используемые для построения прогнозных моделей

Переменная	Описание	Исследования, в которых применялась переменная
Объясняемая переменная		
Return	Доходность акции за период (в рамках данного исследования — час)	Chong, Han, Park, 2017
Авторский индекс сентимента		
RSMI	Russian social media index — индекс сентимента, учитывающий настроения публикаций инвесторов в социальных сетях «Тинькофф Пульс» и «Телеграм»	—
Контрольные переменные — рыночные мультипликаторы		
P/E	Мультипликатор Цена/Прибыль	Basu, 1983; Navratil et al., 2021
P/B	Мультипликатор Цена/Балансовая стоимость	Agrawal, Sehgal, Vasishth, 2020
Контрольные переменные — сигналы, которые дают технические индикаторы акций		
EMA	Экспоненциальная скользящая средняя	Phuoc et al., 2024
MACD	Пересечение двух скользящих средних цены акции	Phuoc et al., 2024
Momentum	Технический индикатор Momentum	Bui et al., 2023
Stochastic	Стохастический осциллятор	Neely et al., 2014
RSI	Индекс относительной силы	Phuoc et al., 2024

2) MSE (среднеквадратическая ошибка) — это среднее значение квадрата разности между прогнозными и фактическими значениями;

3) RMSE (среднеквадратическая ошибка) — метрика, которая показывает квадратный корень из средней квадратичной разности между прогнозируемыми значениями и фактическими значениями в наборе данных. Чем ниже RMSE, тем лучше модель соответствует набору данных.

Перечень прогнозируемых и объясняющих переменных приведен в табл. 2.

4. ВЫБОРКА ДАННЫХ

Выборка включает 78 компаний российского рынка, входивших хотя бы в один из индексов Московской биржи за период 2018–2022 гг. Рассматриваемые компании представляют 25 различных отраслей. Данные по всем компаниям являются полными, выборка сбалансирована. Для сбора информации использовались парсеры данных, которые были написаны на языке Python специально для проведения данного исследования, а также базы данных finam.ru с торговыми показателями каждой компании.

Исследование производится на часовом интервале (таймфрейме) и включает в себя 773 862 наблюдения.

Для построения индекса сентимента используются сообщения из социальных сетей «Пульс» и «Telegram». Данные из социальной сети «Пульс» собраны за 2018–2022 гг. и включают 113 000 наблюдений, данные из социальной сети «Telegram» за период 2020–2022 гг. — более 1 млн наблюдений.

Прогнозирование доходности акций оценивается по двум периодам: обучающему и тестовому, в качестве обучающего выступает период 2019–2021 гг., в качестве тестового — 2022 г.

В табл. 3 представлены описательные статистики по выборке исследования.

Таблица 3. Описательные статистики за период 2019–2022 гг. (число наблюдений = 773864)

Переменная	Среднее	Минимальное значение	Максимальное значение	Стандартное отклонение
Return	0,001	−0,49	0,80	0,01
P/E	3,34	−5,19	165,30	11,99
P/B	1,32	−10,54	18,87	2,76
EMA 12	3491,07	0,01	179480,71	17950,23
EMA 26	3491,35	0,01	178677,12	17954,05
MACD	−0,28	−12423,58	4566,04	111,12
Сигнальная линия MACD	−0,28	−10523,78	3644,23	103,05
Momentum	1,00	0,39	1,82	0,20
RS	2,25	0,00	1223,80	8,78
Stochastic	0,50	0,00	1,00	0,39
RSI	49,28	0,00	99,91	22,70

Примечание. Отрицательные значения коэффициентов P/E и P/B не являются нормальной ситуацией для компаний, но возможны в случае убыточности компании для коэффициента P/E, а также ситуации, в которой обязательства компании превышают ее активы — для коэффициента P/B.

5. РЕЗУЛЬТАТЫ МОДЕЛИРОВАНИЯ

Без использования индекса RSMI самыми эффективными по трем показателям ошибок на обучающей выборке оказались методы случайного леса, градиентного бустинга gbr и экстремально-градиентного бустинга XgBoost (табл. 4). На тестовой выборке минимальные средние ошибки достигаются для методов случайного леса и ансамблевого обучения.

После добавления в модель индекса сентимента RSMI на тех же данных были получены результаты, представленные в табл. 5. На обучающей выборке по метрикам качества лидируют методы «случайного леса», градиентного бустинга gbr и экстремально-градиентного бустинга XgBoost, а на тестовой — случайного леса и ансамблевого обучения. Если комплексно анализировать

Таблица 4. Средние ошибки каждой модели без индекса RSMI на обучающей и тестовой выборках

Модель	Обучающая выборка			Тестовая выборка		
	mae	mse	rmse	mae	mse	rmse
gbr	0,003588	0,000034	0,005603	0,004099	0,000098	0,009498
lasso	0,004547	0,000101	0,009671	0,004519	0,000105	0,009896
lstm	0,004499	0,000062	0,007343	0,006911	0,000266	0,015779
rfr	0,001533	0,000012	0,003304	0,004099	0,000092	0,009223
wel	0,003270	0,000057	0,007260	0,003815	0,000084	0,008813
xg	0,003337	0,000027	0,005044	0,004089	0,000098	0,009528

Таблица 5. Средние ошибки каждой модели с индексом RSMI

Модель	Обучающая выборка			Тестовая выборка		
	mae	mse	rmse	mae	mse	rmse
gbr	0,003585	0,000034	0,005599	0,004098	0,000098	0,009520
lasso	0,004547	0,000101	0,009671	0,004519	0,000105	0,009896
lstm	0,004445	0,000062	0,007317	0,006788	0,000264	0,015693
rfr	0,001534	0,000012	0,003307	0,004099	0,000092	0,009244
wel	0,003282	0,000057	0,007287	0,003813	0,000084	0,008812
xg	0,003333	0,000027	0,005039	0,004089	0,000099	0,009568

Таблица 6. Процентное изменение ошибок каждой модели после добавления индекса RSMI в обучающую и тестовую выборки

Модель	Обучающая выборка			Тестовая выборка		
	mae	mse	rmse	mae	mse	rmse
gbr	−0,09%	−0,10%	−0,07%	−0,02%	0,66%	0,22%
lasso	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
lstm	−1,21%	−0,75%	−0,34%	−1,78%	−0,85%	−0,55%
rfr	0,06%	0,20%	0,09%	0,01%	0,67%	0,23%
Wel	0,36%	0,89%	0,37%	−0,04%	−0,02%	−0,01%
Xg	−0,13%	−0,18%	−0,10%	0,01%	0,80%	0,42%

результаты обучающей и тестовой выборки, то наиболее стабильные и эффективные результаты показывает метод случайного леса.

В табл. 6 приводятся результаты изменения метрик качества (показателей ошибок) моделей после добавления в модели индекса сентимента RSMI. Добавление индекса RSMI в прогнозные модели не позволило улучшить результаты в большинстве моделей, и по некоторым моделям даже привело к ухудшению результатов. Мы объясняем это, во-первых, тем, что наблюдаемый нами период очень большой, особенно с учетом часовых таймфреймов. Во-вторых, активность российских инвесторов в социальных сетях по компаниям в среднем низкоинтенсивная, это зашумляет наблюдения, потому что преобладают периоды с нулевым значением индекса RSMI. Наконец, как было указано в разделе «Гипотезы», для большинства частных инвесторов характерна низкая торговая активность (около двух сделок в месяц в среднем), и появление и обсуждение той или иной новости о компании-эмитенте не приводит к оперативному совершению сделок с акциями.

Таким образом, для компаний широкого рынка (всех эшелонов) подтвердилась первая гипотеза, предполагающая большую роль индикаторов технического анализа и рыночных мультипликаторов в прогнозировании доходности акций широкой выборки компаний российского рынка по сравнению с сентиментом частных инвесторов в социальных сетях.

Таблица 7. Средние ошибки каждой модели без индекса RSMI для топ-15 компаний по обсуждаемости

Модель	Обучающая выборка			Тестовая выборка		
	mae	mse	rmse	mae	mse	rmse
gbr	0,003292	0,000028	0,005204	0,003727	0,000099	0,009586
lasso	0,004222	0,000086	0,009063	0,004213	0,000096	0,009528
lstm	0,004092	0,000050	0,006625	0,006545	0,000262	0,015706
rfr	0,001407	0,000011	0,003178	0,003740	0,000095	0,009398
wel	0,002961	0,000049	0,006775	0,003507	0,000079	0,008622
xg	0,003160	0,000025	0,004878	0,003751	0,000104	0,009866

Таблица 8. Средние ошибки каждой модели с индексом RSMI для топ-15 компаний по обсуждаемости

Модель	Обучающая выборка			Тестовая выборка		
	mae	mse	rmse	mae	mse	rmse
gbr	0,003288	0,000028	0,005192	0,003731	0,000094	0,009399
lasso	0,004222	0,000086	0,009063	0,004213	0,000096	0,009528
lstm	0,004126	0,000050	0,006595	0,006670	0,000265	0,015833
rfr	0,001406	0,000011	0,003188	0,003743	0,000093	0,009310
wel	0,002929	0,000046	0,006588	0,003509	0,000079	0,008608
xg	0,003149	0,000024	0,004849	0,003750	0,000103	0,009821

Таблица 9. Процентные изменения ошибок каждой модели после добавления индекса RSMI в обучающую и тестовую выборки для топ-15 компаний по обсуждаемости

Модель	Обучающая выборка			Тестовая выборка		
	mae	mse	rmse	mae	mse	rmse
gbr	−0,09%	−0,47%	−0,24%	0,11%	−4,91%	−1,95%
lasso	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
lstm	0,83%	−0,16%	−0,46%	1,91%	0,86%	0,81%
rfr	−0,05%	0,74%	0,30%	0,07%	−2,39%	−0,94%
wel	−1,08%	−5,77%	−2,76%	0,06%	−0,14%	−0,16%
xg	−0,34%	−1,55%	−0,60%	−0,02%	−1,40%	−0,46%

Далее мы отобрали топ-15 компаний по числу сообщений в социальных сетях на рассматриваемом периоде 2018–2022 гг. и построили для них новые модели. Эти компании характеризуются и большими значениями индекса RSMI ($RSMI > 1,3$ млн): MGNT (Магнит), POLY (Полиметалл), YNDX (Яндекс), NVTX (Новатэк), ROSN (Роснефть), CHMF (Северсталь), ALRS (Алроса), GAZP (Газпром), SBER (Сбербанк), BELU (Белуга), VTBR (ВТБ), GMKN (ГМК Норильский Никель), AFLT (Аэрофлот), DSKY (Детский мир), AFKS (АФК Система). Мы построили новые модели только для этих компаний.

Табл. 7 содержит средние ошибки каждой модели без индекса RSMI (табл. 8 — с добавлением в модель индекса RSMI) для топ-15 наиболее обсуждаемых компаний. Среди моделей, построенных на данных по самым обсуждаемым компаниям, наилучшие прогнозные свойства показали модели случайного леса (rfr), градиентного бустинга (gbr), ансамблевого обучения (wel).

В табл. 9 приведены изменения метрики качества моделей при добавлении фактора сентимента. Для модели ансамблевого обучения учет сентимента на обучающей выборке позволяет уменьшить ошибку MSE на 5,77%, для модели градиентного бустинга на тестовой снижение ошибки MSE составляет 4,91%, для модели случайного леса — 2,39%, а для модели экстремального градиентного бустинга — 1,4%.

По отдельным компаниям благодаря добавлению в модель индекса RSMI удалось снизить показатель ошибки MAE более чем на 10% (NLMK, MGNT, RTKMP, ALRS, NVTK), а показатель ошибки MSE более чем на 20% (NLMK, SBER, MRKC, VTBR).

Высокое качество построенных нами моделей можно подтвердить сравнением результатов с другими работами. Так, в работе (Aslim et al., 2024) с помощью добавления индекса сентимента удалось построить LSTM-модель с наилучшими показателями ошибок $RMSE = 0,0991$ и $MAPE = 0,1660$ по сравнению с $RMSE = 0,0086$ и $MAE = 0,003509$ для модели ансамблевого обучения в нашей работе. В исследовании (Ahmed, Neema, Visqanadha, 2022), посвященном использованию методов искусственного интеллекта и индекса сентимента для прогнозирования цен акций, применялись модели линейной регрессии, рекуррентной нейронной сети, опорных векторов и случайного леса с лучшим показателем ошибки $RMSE = 0,0510$, и за счет добавления индекса сентимента показатель ошибки снизился на 2,5%. В нашем исследовании для модели ансамблевого обучения с индексом сентимента показатель ошибки — $RMSE = 0,0086$.

Таким образом, можно утверждать, что *вторая гипотеза нашла свое подтверждение*: анализ сентимента публикаций инвесторов в социальных сетях и добавление индекса сентимента RSMI в модели прогнозирования доходности акций позволяет повысить их точность.

Для проверки третьей гипотезы исследования была построена модель, включающая только акции второго и третьего котировальных списков, для которых характерен преимущественно средний и низкий уровень ликвидности соответственно. Всего из 78 исследуемых компаний — эмитентов акций в эту группу попали 32 компании, из которых достаточное количество публикаций в социальных сетях (позволяющее получить индекс RSMI более 50 000) было только по 11 тикерам (SNGSP, FESH, SNGS, AKRN, BANEP, RASP, MRKP, UNAC, AMEZ, APTK, VSMO). Построенные модели показали худшие метрики качества, чем модели, построенные на всем объеме данных и модели по самым обсуждаемым акциям (табл. 10–11).

Таблица 10. Сравнение показателей средних ошибок по всем моделям

Выборка акций	mae	mse	rmse
Все акции	0,004026	0,000086	0,008413
Топ-15 акций по обсуждаемости	0,003728	0,000081	0,008164
Акции второго и третьего эшелонов	0,004625	0,000103	0,009298

Таблица 11. Результаты оценки эффективности моделей, построенных на данных по акциям второго и третьего эшелонов

Модель	Обучающая выборка			Тестовая выборка		
	mae	mse	rmse	mae	mse	rmse
Без индекса RSMI						
gbr	0,004158	0,000045	0,006451	0,004766	0,000113	0,010233
lasso	0,005259	0,000133	0,011148	0,005212	0,000130	0,010975
lstm	0,005168	0,000089	0,008832	0,007347	0,000279	0,016068
rfr	0,001787	0,000015	0,003772	0,004766	0,000106	0,009929
wel	0,003820	0,000075	0,008371	0,004426	0,000101	0,009680
xg	0,003795	0,000034	0,005667	0,004752	0,000115	0,010292
С индексом RSMI						
gbr	0,004163	0,000045	0,006458	0,004766	0,000114	0,010274
lasso	0,005259	0,000133	0,011148	0,005212	0,000130	0,010975
lstm	0,005260	0,000090	0,008891	0,007493	0,000282	0,016173
rfr	0,001786	0,000015	0,003767	0,004766	0,000106	0,009935
wel	0,003802	0,000074	0,008359	0,004427	0,000101	0,009678
xg	0,003804	0,000035	0,005677	0,004756	0,000114	0,010246

Таблица 11. Окончание

Модель	Обучающая выборка			Тестовая выборка		
	mae	mse	rmse	mae	mse	rmse
Изменение показателей ошибок после добавления индекса RSMI						
gbr	0,13%	0,14%	0,12%	0,01%	0,63%	0,40%
lasso	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
lstm	1,78%	1,07%	0,66%	2,00%	0,98%	0,65%
rfr	−0,03%	−0,38%	−0,13%	−0,02%	0,05%	0,07%
wel	−0,46%	−0,62%	−0,15%	0,04%	−0,08%	−0,02%
xg	0,22%	0,34%	0,18%	0,10%	−0,78%	−0,45%

Третья гипотеза исследования не нашла своего подтверждения. Добавление индекса RSMI в модели не дало существенного улучшения предсказательной способности. Количество публикаций инвесторов в социальных сетях относительно данных компаний очень мало, так, средний индекс сентимента RSMI по компаниям второго и третьего эшелонов в 50 раз меньше, чем его совокупное среднее значение по всем компаниям. К похожим результатам в своей работе приходят (Gao et al., 2022): было установлено, что влияние настроений инвесторов в Интернете на скачки волатильности акций компаний малого и среднего бизнеса и быстрорастущих компаний относительно ограничено. Как показано в работе (Xu, Xue, Zhang, 2024), на китайском фондовом рынке положительная связь между сентиментом инвесторов и ликвидностью акций более ярко проявляется для крупных компаний. Авторы объясняют это тем, что более крупные компании или компании с высокой долей институциональных инвесторов в структуре собственного капитала, как правило, привлекают больше внимания инвесторов, что делает их более чувствительными к изменениям в настроениях инвесторов.

ЗАКЛЮЧЕНИЕ

В нашей работе на данных фондового рынка России показано, что учет сентимента частных инвесторов в социальных сетях позволяет более точно прогнозировать изменение доходности российских акций.

Согласно результатам исследования, для наиболее обсуждаемых акций российских компаний-эмитентов добавление индекса сентимента частных инвесторов в социальных сетях позволяет более точно прогнозировать изменение доходности акций. Наиболее эффективными моделями для прогнозирования цен акций с учетом сентимента являются модели случайного леса (rfr), градиентного бустинга (gbr) и ансамблевого обучения (wel).

Для акций широкой выборки компаний российского рынка добавление сентимента позволяет улучшить результаты прогнозирования доходности на часовых данных, но индикаторы технического анализа и рыночные мультипликаторы играют большую роль в построении прогнозов. Это связано с низкой торговой активностью российских инвесторов (среднее число сделок на одного участника фондового рынка — 2 сделки в месяц), а также с отсутствием постоянного интереса ко многим компаниям. Для учета сентимента в прогнозировании доходности акций широкой выборки следует рассматривать более длительные временные периоды.

Было установлено, что на доходность акций второго и третьего эшелонов российского фондового рынка сентимент частных инвесторов на часовых данных не оказывает существенного влияния, а добавление индекса сентимента не позволяет улучшить результатов прогнозных моделей изменения цен при горизонте наблюдений более 1 года. Число публикаций инвесторов в социальных сетях относительно компаний второго и третьего эшелонов очень мало, так, средний индекс сентимента RSMI по компаниям второго и третьего эшелонов в 50 раз меньше, чем его совокупное среднее значение по всем компаниям (что соотносится с числом публикаций).

Использование стратегий принятия решений, учитывающих сентимент инвесторов в социальных сетях, позволяет улучшить результаты прогнозирования доходности российских акций, но должно учитывать эффективность добавления сентимента по каждой компании, а также интенсивность обсуждений. Как показано в нашей работе, в краткосрочных стратегиях

(прогнозирование на часовых данных на час вперед) добавление сентимента может улучшить качество моделей по наиболее обсуждаемым акциям. В долгосрочных стратегиях принятия решений учет сентимента может приводить к созданию дополнительного шума, особенно в периоды, когда число публикаций по отдельной компании минимально.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

- Теплова Т.В., Соколова Т.В., Томтосов А.Ф., Бучко Д.В., Никулин Д.Д. (2022). Сентимент частных инвесторов в объяснении различий в биржевых характеристиках акций российского рынка // *Журнал Новой экономической ассоциации*. № 1 (53). С. 53–84. [Teplova T.V., Sokolova T.V., Tomtosov A.F., Buchko D.V., Nikulin D.D. (2022). The sentiment of private investors in explaining the differences in the trade characteristics of the Russian market stocks. *Journal of the New Economic Association*, 1 (53), 53–84 (in Russian).]
- Agrawal T.J., Sehgal S., Vasishth V. (2020). Firm attributes, corporate fundamentals and investment strategies: An empirical study for Indian stock market. *Management and Labour Studies*, 45 (3), 366–387. DOI: 10.1177/0258042X20927995
- Ahmed D., Neema R., Visqanadha N. (2022). Analysis and prediction of healthcare sector stock price using machine learning techniques: Healthcare stock analysis. *International Journal of Information System Modeling and Design*, 13, 9, 1–15. DOI: 10.4018/IJISMD.303131
- Aslim M.F., Firmansyah G., Tjahjono B., Akbar H., Widodo A.M. (2024). Utilization of LSTM (Long Short Term Memory) based sentiment analysis for stock price prediction. *Asian Journal of Social & Humanities*, 1, 12, 1241–1255. DOI: 10.59888/ajosh.v1i12.141
- Audrino F., Sigrist F., Ballinari D. (2020). The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting*, 36, 2, 334–357.
- Baker M., Wurgler J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21, 2, 129–152.
- Basu S. (1983). The relationship between earnings yield, market value and return for NYSE Common Stocks. *Journal of Financial Economics*, 12, 129–156. DOI: 10.1016/0304-405X(83)90031-4
- Bui D.G., Kong D.-R., Lin C.-Y., Lin T.-C. (2023). Momentum in machine learning: Evidence from the Taiwan stock market. *Pacific-Basin Finance Journal*, 82. Article 102178. DOI: 10.1016/j.pacfin.2023.102178
- Cai Y., Tang Z., Chen Y. (2024). Can real-time investor sentiment help predict the high frequency stock returns? Evidence from a mixed-frequency-rolling decomposition forecasting method. *North American Journal of Economics & Finance*, 72. Article 102147. DOI: 10.1016/j.najef.2024.102147
- Chen S., Ge L. (2019). Exploring the attention mechanism in LSTM-based Hong Kong stock price movement prediction. *Quantitative Finance*, 19, 9, 1507–1515.
- Chong E., Han C., Park F.C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, 83, 187–205.
- Fama E.F. (1965). The behavior of stock-market prices. *Journal of Business*, 38, 34, Article 105.
- Gao B., Xie J. (2020). Forecasting excess returns and abnormal trading, using investor sentiment: Evidence from Chinese stock index futures market. *Emerging Markets Finance and Trade*, 56, 3, 593–612.
- Gao Y., Zhao C., Sun B., Zhao W. (2022). Effects of investor sentiment on stock volatility: New evidences from multi-source data in China's green stock markets. *Financial Innovation*, 8, Article 77.
- Gu C., Kurov A. (2020). Informational role of social media: Evidence from Twitter sentiment. *Journal of Banking & Finance*, 121, Article 105969.
- Gupta R., Nel J., Pierdzioch C. (2023). Investor confidence and forecastability of US stock market realized volatility: Evidence from machine learning. *Journal of Behavioral Finance*, 24, 1, 111–122.
- Li T., Chen H., Liu W., Yu G., Yu Y. (2023). Understanding the role of social media sentiment in identifying irrational herding behavior in the stock market. *International Review of Economics & Finance*, 87, 163–179. DOI: 10.1016/j.iref.2023.04.016
- Li Y., Bu H., Li J., Wu J. (2020). The role of text-extracted investor sentiment in Chinese stock price prediction with the enhancement of deep learning. *International Journal of Forecasting*, 36 (4), 1541–1562.
- Liang C., Tang L., Li Y., Wei Y. (2020). Which sentiment index is more informative to forecast stock market volatility? Evidence from China. *International Review of Financial Analysis*, 71, Article 101552.
- Lin P., Ma S., Fildes R. (2024). The extra value of online investor sentiment measures on forecasting stock return volatility: A large-scale longitudinal evaluation based on Chinese stock market. *Expert Systems with Applications*, 238, Article 121927. DOI: 10.1016/j.eswa.2023.121927

- Liu J.-X., Leu J.-S., Holst S. (2023). Stock price movement prediction based on stocktwits investor sentiment using FinBERT and ensemble SVM. *PeerJ Computer Science*, 9, Article e1403. DOI: 10.7717/peerj-cs.1403
- Liu Q., Lee W.-S., Huang M., Wu Q. (2023). Synergy between stock prices and investor sentiment in social media. *Borsa Istanbul Review*, 23, 1, 76–92. DOI: 10.1016/j.bir.2022.09.006
- Mili M., Sahut J.-M., Teulon F., Hikkerova L. (2024). A multidimensional Bayesian model to test the impact of investor sentiment on equity premium. *Annals of Operations Research*, 334, 1–3, 919–39. DOI: 10.1007/s10479-023-05165-0
- Navratil R., Taylor S., Vecer J. (2021). On equity market inefficiency during the COVID-19 pandemic. *International Review of Financial Analysis*, 77, Article 101820. DOI: 10.1016/j.irfa.2021.101820
- Neely C., Rapach D.E., Tu J., Zhou G. (2014). Forecasting the equity risk premium: The role of technical indicators. *Management Science*, 60, 7, 1772–1791. DOI: 10.1287/mnsc.2013.1838
- Niu H., Pan Q., Xu K. (2023). Hybrid deep learning models with multi-classification investor sentiment to forecast the prices of China's leading stocks. *PLoS ONE*, 18, 11, Article e0294460. DOI: 10.1371/journal.pone.0294460
- Ph H., Rishad A. (2020). An empirical examination of investor sentiment and stock market volatility: Evidence from India. *Financial Innovations*, 6, Article 34.
- Phuoc T., Anh P.T.K., Tam P.H., Nguyen C.V. (2024). Applying machine learning algorithms to predict the stock price trend in the stock market — the case of Vietnam. *Humanities and Social Sciences Communications*, 11, Article 393. DOI: 10.1057/s41599-024-02807-x
- Sarkar A., Chakraborty S., Ghosh S., Naskar S.K. (2022). Evaluating impact of social media posts by executives on stock prices. *FIRE '22: Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, 74–82. DOI: 10.1145/3574318.3574339
- Shiller R.J. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives*, 17, 83, Article 104.
- Tallboys J., Zhu Y., Rajasegarar S. (2022). Identification of stock market manipulation with deep learning. In: *International Conference on Advanced Data Mining and Applications*, 408–420. Cham: Springer.
- Teplova T., Tomtosov A., Sokolova T. (2022). A retail investor in a cobweb of social networks. *PLoS One*, 17, 12, Article e0276924.
- Uslu N.C., Akal F. (2021). A machine learning approach to detection of trade-based manipulations in Borsa Istanbul. *Computational Economics*, 60, 1, 25–45.
- Wang G., Yu G., Shen X. (2020). The effect of online investor sentiment on stock movements: An LSTM approach. *Complexity*, 1–11, Article 4754025.
- Xu L., Xue C., Zhang J. (2024). The impact of investor sentiment on stock liquidity of listed companies in China. *Investment Management and Financial Innovations*, 21, 2, 1–14.
- Xu Q., Wang L., Jiang C., Zhang X. (2019). A novel UMIDAS-SVQR model with mixed frequency investor sentiment for predicting stock market volatility. *Expert Systems with Applications*, 132, 12–27.

Forecasting Russian stock returns based on investor sentiment analysis in social networks

© 2025 G.A. Khaziev, T.V. Sokolova

G.A. Khaziev,

*National Research University Higher School of Economics (HSE University), Moscow, Russia;
e-mail: gakhaziev@hse.ru*

T.V. Sokolova,

*National Research University Higher School of Economics, (HSE University), Moscow, Russia;
e-mail: tv.sokolova@hse.ru*

Received 26.04.2024

The research was supported by the Russian Science Foundation under grant 22-18-002760, <https://rscf.ru/project/23-28-00740/>

Abstract. The study explores the sentiment of Russian private investors in social networks and its impact on the dynamics of the stock return of 78 companies on the Russian stock market (MOEX) in the period from 2018 to 2022. To take into account sentiment when forecasting returns, the authors RSMI index (Russian social media index) is used, which is based on a unique sample of messages from the most popular social networks among Russian investors — “Telegram” and “Tinkoff Pulse”. The RSMI index includes quantitative (the number of publications in relation to each company) and qualitative (investor reactions) characteristics, allowing to determine the real impact of a particular publication on investors. Using the RSMI index, several models for predicting stock prices of Russian companies were used: lasso regression, random forest, gradient boosting, extreme gradient boosting, ensemble learning and long short-term memory. It is demonstrated that for a wide sample of stocks, indicators of technical and fundamental analysis play a large role in building forecasts of changes in stock returns based on hourly data. Although the addition of the sentiment index improves the results of predicting returns for a wide sample of stocks, it does not significantly improve the predictive ability of the models and shows mixed results. The best results of adding the sentiment index to forecast models are shown for the top 15 most discussed Russian companies. For individual models, we achieved an average error reduction of 4.9%, and at the level of specific companies, the MAE error rate was reduced by more than 10% and MSE by 20%. It has been proven that the returns of low-liquidity stocks of the second and third tiers of the Russian stock market are not significantly influenced by the sentiment of private investors on hourly data, and the addition of the sentiment index does not improve the results of forecast models.

Keywords: stocks, stock market, machine learning, investor sentiment.

JEL Classification: E44, G12, G14.

UDC: 336.763.2.

For reference: **Khaziev G.A., Sokolova T.V.** (2025). Forecasting Russian stock returns based on investor sentiment analysis in social networks. *Economics and Mathematical Methods*, 61, 1, 95–108. DOI: 10.31857/S0424738825010095 (in Russian).