

ФИНАНСОВАЯ ЭКОНОМИКА

А. И. Шайдуллин¹

Национальный исследовательский университет
«Высшая школа экономики» (Москва, Россия)

Ен Ми Ан²

Национальный исследовательский университет
«Высшая школа экономики» (Москва, Россия)

В. Е. Миткинов³

Национальный исследовательский университет
«Высшая школа экономики» (Москва, Россия)

УДК: 336.61, 336.63

doi: 10.55959/MSU0130-0105-6-61-1-5

ПРЕДСКАЗАНИЕ ВОЛАТИЛЬНОСТИ АКЦИЙ РОССИЙСКИХ КОМПАНИЙ С ИСПОЛЬЗОВАНИЕМ ДАННЫХ СОЦИАЛЬНЫХ СЕТЕЙ

В научной литературе существует противоречие между растущей ролью социальных медиа как источника информации для принятия инвестиционных решений и отсутствием формализованных методов количественной оценки их влияния на волатильность российских акций. Данное исследование направлено на преодоление этого разрыва путем создания модели, которая позволит преобразовать качественные данные из соцсетей в прогнозный фактор рыночной нестабильности. Цель работы — на основе данных социальных сетей методами машинного обучения построить модель с весами настроения и внимания инвесторов для оценки и прогнозирования волатильности акций некоторых российских компаний. В статье рассматриваются компании из разных отраслей экономики: Лукойл, Роснефть, Газпром, Новатэк, Сбербанк, Яндекс и МТС. Для оценки качества построены модели трех видов: 1) базовая модель: модель реализованной волатильности; 2) модель оценки состояния финансового рынка: модель с финансовыми весами; 3) модель настроения и внимания: модель с обработкой тональности новостей из Twitter (X). Исследуемые компании по-разному реагируют

¹ Шайдуллин Ансэль Ильгизович — аспирант, Национальный исследовательский университет «Высшая школа экономики»; e-mail: ansel.shajdullin@yandex.ru, ORCID: 0000-0002-2653-1745.

² Ан Ен Ми — магистр, Национальный исследовательский университет «Высшая школа экономики».

³ Миткинов Валентин Евгеньевич — магистр, Национальный исследовательский университет «Высшая школа экономики».

© Шайдуллин Ансэль Ильгизович, 2026 

© Ан Ен Ми, 2026 

© Миткинов Валентин Евгеньевич, 2026 

на настроения в социальных сетях. Для компании Яндекс информация о настроении и внимании в социальных сетях имеет большую объяснительную силу: факторы внимания и настроения улучшают качество прогноза в среднем по всем методам на 7% относительно модели с финансовыми весами. Для Сбербанка и МТС модель настроения улучшает качество прогноза на 3% в сравнении с моделью оценки состояния финансового рынка. Компании, связанные с нефтедобычей и газодобычей, хуже реагируют на факторы социальных сетей, и при лучшем методе дают прирост качества прогноза в среднем на 2%. Результаты исследования позволяют расширить представления о возможностях машинного обучения для анализа влияния тональности сообщений в социальных сетях на волатильность акций.

Ключевые слова: тональность новостей, реализованная волатильность, фондовый рынок, метрика SHAP, социальные сети.

Цитировать статью: Шайдудлин, А. И., Ан Ен Ми, & Миткинов, В. Е. (2026). Предсказание волатильности акций российских компаний с использованием данных социальных сетей. *Вестник Московского университета. Серия 6. Экономика*, 61(1), 87–124. <https://doi.org/10.55959/MSU0130-0105-6-61-1-5>.

A. I. Shaidullin

HSE University (Moscow, Russia)

An En Mi

HSE University (Moscow, Russia)

V. E. Mitkinov

HSE University (Moscow, Russia)

JEL: C22, C45, C51, C52, C53, C58

FORECASTING VOLATILITY OF RUSSIAN COMPANIES' STOCKS WITH SOCIAL MEDIA DATA

In academic literature, there is a contradiction between the growing role of social media as a source of information for investment decision-making and the lack of formalized methods for quantitatively assessing their impact on the volatility of Russian equities. This study aims to bridge this gap by developing a model that transforms qualitative data from social networks into a predictive factor for market instability. The purpose of the paper is to build a model based on social media data using machine learning methods with weights of investors' sentiment and attention to assess and predict the volatility of shares of some Russian companies. The article examines companies from different sectors of the economy: Lukoil, Rosneft, Gazprom, Novatek, Sberbank, Yandex and MTS. Three types of models have been built to assess quality: 1) Baseline Model: realized volatility model; 2) Financial Market State Evaluation Model: financially weighted model; 3) Sentiment and Attention Model: model with social media tone processing based on Twitter (X) data. Companies exhibit varying responsiveness to social media sentiments. For Yandex, information about sentiment and attention derived from social networks has substantial explanatory power, improving prediction quality by an average of 7% compared to the financially weighted model across all methods. In contrast, for Sberbank and

MTS, the sentiment-based model enhances forecast accuracy by approximately 3%, relative to the financial market state evaluation model. Companies engaged in oil and gas extraction demonstrate weaker reactions to social network factors, yielding only a modest improvement in forecast quality averaging at 2%. This study contributes to expanding our understanding of the way machine learning can be utilized to analyze the impact of social media message tone on stock price volatility.

Keywords: news tone, realized volatility, stock market, SHAP metric, social networks.

To cite this document: Shaidullin, A. I., An En Mi, & Mitkinov, V. E. (2026). Forecasting volatility of Russian companies' stocks with social media data. *Lomonosov Economics Journal*, 61(1), 87–124. <https://doi.org/10.55959/MSU0130-0105-6-61-1-5>

Введение

Концепции поведенческой экономики свидетельствуют о том, что «эмоциональные факторы» могут значимо влиять на принятие решений инвесторами (Obaidi et al., 2022). Например, в условиях экономической и политической неопределенности, которая может быть вызвана такими факторами, как изменения в налоговой политике, торговые войны между странами, санкции или нестабильная геополитическая ситуация, финансовые рынки могут столкнуться с повышенной турбулентностью. Такая турбулентность часто возникает вследствие панических настроений среди инвесторов, которые начинают массово продавать активы, опасаясь возможных убытков. Эти действия, в свою очередь, приводят к резким колебаниям цен на акции, облигации и другие финансовые инструменты, что усиливает волатильность рынков и увеличивает риски для всех участников.

В связи с этим актуализируются следующие исследовательские вопросы:

- 1) Коррелирует ли общественное настроение с финансовыми показателями или предсказывает их?
- 2) Улучшается ли качество моделей с финансовыми весами при добавлении весов настроения и внимания инвесторов?

Всплеск активности в области извлечения мнений и анализа настроений произошел как прямой ответ на всплеск интереса к новым системам оценивания и прогнозирования (Wankhade et al., 2022). В истории фондового рынка есть множество драматических событий, когда цены на акции менялись не только под воздействием финансовых факторов, но и под влиянием поведенческих: Великая депрессия (1929–1933 гг.), пузырь Dot-com (конец 1990-х — начало 2000-х гг.), ипотечный кризис (2007 г.) и др. Например, одной из причин кризиса Dot-com являются высокие ожидания инвесторов от быстрорастущего интернета. Они полагали, что любая компания, связанная с онлайн-бизнесом, автоматически будет успешной. Это приводило к чрезмерным инвестициям в стартапы с сомнительными бизнес-моделями. В результате многие проекты оказались

финансово несостоятельными, что спровоцировало обрушение котировок и массовые банкротства.

Цель работы — на основе данных социальных сетей методами машинного обучения построить модель с весами настроения и внимания инвесторов для оценки и прогнозирования волатильности акций некоторых российских компаний. Материалы работы представляют практическую ценность для банковских организаций и инвесторов, которые могут применять предложенную методологию для прогнозирования будущей волатильности акций.

В работе рассматриваются данные по акциям семи российских компаний: «Яндекс», «Сбербанк», МТС, «Лукойл», «Роснефть», «Газпром» и «Новатэк» (период исследования: 1 января 2020 г. — 28 февраля 2021 г.).

В рамках исследования проверяется следующая гипотеза: «Настроения в социальных сетях способны влиять на волатильность акций компаний, а, следовательно, модель с добавленными весами настроения и внимания способна улучшить качество прогнозов волатильности акций российских компаний в краткосрочной перспективе».

В работе представлена следующая структура. Вначале проводится обзор литературы. Затем описываются данные для проведения исследования, рассматриваются три набора данных: 1) данные, оценивающие реализованную волатильность акций; 2) финансовые данные; 3) данные настроения и внимания из социальной сети Twitter. После строятся модели трех типов: модель реализованной волатильности; модель оценки состояния финансового рынка; модель настроения и внимания. В итоговой части статьи описываются эмпирические результаты по каждой из моделей. В заключении подводятся выводы и перспективы дальнейших исследований.

Обзор литературы

Согласно гипотезе эффективного рынка (Fama, 1970), прогнозирование доходности акций не должно быть возможным, поскольку рыночные цены будут отражать всю доступную информацию. Более современная гипотеза смешанного распределения (The Mixture of Distribution Hypothesis, MDH) утверждает, что волатильность цен и объем торгов подчинены одной и той же скорости поступления информации.

Первые упоминания о практической возможности использования поведенческих факторов в исследованиях, посвященных анализу настроений инвесторов, встречаются в статьях начала 2000-х гг. Различные исследователи, среди которых (Daniel et al., 2002; Tseng, 2006; Revathy et al., 2022) сообщают о росте количества эмпирических данных, демонстрирующих, что фондовый рынок управляется психологией инвесторов. Следующим этапом в анализе психоэмоциональных факторов в экономике является более глубокая оценка тональности текста, а также критика отдельных

концепций поведенческой экономики, посвященных исследованию роли поведения инвесторов в волатильности цен на акции. Ряд авторов, среди которых (Antweiler, Frank, 2004; Oliveira et al. (2013), скептически отнеслись к возможностям социальных сетей предсказывать волатильность акций на основе данных поведения инвесторов.

Некоторые исследования приходят к выводу, что настроения общественности, полученные из социальных сетей, таких как Twitter⁴ или LiveJournal, могут быть использованы для прогнозирования движений фондового рынка. Например, (Da et al., 2011) показывают, что повышенное внимание инвесторов предсказывает рост цен в краткосрочной перспективе. В статье (Andersen et al., 2001) авторы предполагают, что внутридневная доходность может быть использована для построения оценок дневной волатильности доходности, которые являются более точными, чем те, которые строятся с использованием дневной доходности. Аналогичным образом (Aouadi et al., 2013; Dimpfl, Jank, 2015; Hamid, Heiden, 2015) демонстрируют, что объем поисковых запросов можно использовать для прогнозирования волатильности фондового рынка.

Исследованием волатильности акций российских компаний с использованием параметров настроения занимались, например, Теплова и др. (2022). Авторы используют данные социальной сети Telegram и обнаруживают нелинейные эффекты влияния настроения инвесторов на биржевые характеристики акций российского рынка. В текущей же работе применяется социальная сеть Twitter и несколько инструментов построения прогнозов: от классических до более современных. Выбор социальной сети Twitter преимущественно обусловлен простотой обработки твиттов (сообщений) для построения прогнозных моделей.

Ряд авторов отдельное внимание уделяют инструментам моделирования, а также некоторым важным деталям функционирования фондового рынка. В исследовании Аганин (2017) проводится сравнение GARCH, ARFIMA и HAR-RV моделей по качеству одношагового прогноза реализованной волатильности на один день вперед на российском фондовом рынке. Было выявлено, что HAR-RV обладает наиболее высокой прогнозной силой.

Важно отметить следующие характерные черты финансового рынка (Audrino et al., 2020; Corsi et al., 2012):

- 1) «долгая память»: реализованная волатильность показывает значимую автокорреляцию даже очень больших лагов. Чтобы учитывать эту особенность применяются разные временные шкалы (для краткосрочного влияния используется дневной лаг, для среднесрочного — недельный, для долгосрочного — месячный);

⁴ По состоянию на март 2023 г.

- 2) «эффект рычага»: доходность отрицательно коррелирована с реализованной волатильностью;
- 3) «скачки»: цены подвержены резким скачкам, которые вносят большой вклад в будущую волатильность. Оценка реализованной волатильности используется потому, что она показала хорошую устойчивость к скачкам и «нулевой» доходности.

Предыдущие исследования рассматривают ограниченный набор переменных-предикторов и/или используют только ограниченный объем данных. Например, исследования (Caropin, Poli, 2017; Mao et al., 2011) не учитывают экономические и финансовые переменные. Таким образом, нельзя достоверно утверждать, что настроения в социальных сетях и/или в новостях обладают дополнительной прогностической силой. В статье Audrino et al. (2020) используется методология, позволяющая оценить влияние различных новостей на реализованную волатильность ряда американских компаний. В настоящей работе данная методология была адаптирована для исследования акций российских компаний. Определенным изменениям подверглась и методика проведения исследования: были добавлены новые факторы, некоторые показатели были упрощены. Для построения более качественных моделей, с относительно высокой прогностической силой, были использованы методы с нейронными сетями.

Данные для эмпирического исследования

В работе рассмотрены акции семи российских компаний: «Яндекс», «Сбербанк», МТС, «Лукойл», «Роснефть», «Газпром» и «Новатэк». Выборка включает компании, представляющие три ключевых сектора российской экономики: нефтегазовый, банковский и информационно-коммуникационный/ИТ-сектор. Такой подход полностью согласуется с задачами исследования, поскольку полученные выводы носят обобщенный характер и могут быть распространены на широкий круг аналогичных компаний независимо от отраслевой принадлежности. Выбранные компании занимают ведущие позиции на российском рынке и оказывают значительное влияние на динамику фондовых индексов. Доминирование представителей нефтегазового сектора объясняется ведущей ролью добывающей промышленности в структуре российской экономики.

Выбранные компании характеризуются высокими показателями «медийности», что подтверждается их значительным присутствием в средствах массовой информации и социальных сетях. Это выражается в частоте упоминаний в новостных публикациях, аналитических обзорах и социальных медиа, а также в объеме и разнообразии контента, связанного с их деятельностью. В приложении А представлен медиарейтинг исследуемых компаний за 2020–2024 гг. Как можно заметить, все исследуемые компании на протяжении периода с 2020 по 2024 г. ежегодно входят в число

50 наиболее упоминаемых в средствах массовой информации, что свидетельствует об их высокой медийной активности и значимости в общественном дискурсе.

Все компании котируются на Московской фондовой бирже. Выбор этих компаний также мотивирован тем, что переменные настроения и внимания могут по-разному влиять на будущую реализованную волатильность в зависимости от типа рассматриваемой акции. Перспективным направлением для дальнейших изысканий является углубленный анализ компаний внутри конкретных секторов экономики (например, банковского), способный выявить новые аспекты проблемы, значимые как для научного сообщества, так и для практиков. Кроме того, будущие исследования могут быть сфокусированы на организациях из иных отраслей с целью детального изучения отраслевой специфики. В рамках настоящего исследования объектом анализа выступили крупнейшие российские компании, поскольку эффект влияния социальных сетей на их деятельность может существенно отличаться от воздействия на средний и малый бизнес.

В качестве исследуемого периода был выбран период с 1 января 2020 г. по 28 февраля 2021 г. Подобный выбор обусловлен тем, что степень влияния факторов различна по всему временному горизонту, и обучаться на многолетнем периоде — плохая идея, так как влияние более давних периодов сильно ухудшает прогнозы в поздних периодах. Таким образом, модель получала избыточные данные и переобучалась. Отметим, что из-за экономических и политических событий результаты исследования могут несколько различаться в зависимости от того, на каких промежутках времени проводится исследование. В последующих исследованиях необходимо проверить результаты на других временных промежутках.

Модели обучаются на периоде пандемии коронавируса в 2020 г., а вне выборки прогноз производится на январь-март 2021 г. Выбор данного периода обусловлен рядом причин. Если бы использовались данные за 2021 г. и была бы осуществлена попытка спрогнозировать начало 2022 г., то можно было бы столкнуться с большой ошибкой прогноза (в феврале 2022 г. был шок, который не поддается прогнозу). По этой причине было принято решение спрогнозировать относительно спокойный 2021 г. на данных 2020 г. Последний подход применялся с целью вычерпнуть эмоционально ярко окрашенные данные для построения более чувствительных прогнозов: инвесторы реагируют более эмоционально на плохие новости из-за боязни значительных потерь. Анализ динамики индексов настроений инвесторов позволяет выделять важные периоды пессимизма на фондовом рынке.

В рамках дальнейших исследований планируется задействование более свежих массивов данных с целью разработки альтернативных подходов к построению прогностических моделей. Предварительный анализ указанных данных свидетельствует о необходимости внедрения специфиче-

ских методов обработки, включая нормализацию, экстраполяцию и интерполяцию, ввиду наличия особенностей структуры и динамики, отличающихся от ранее использованных наборов данных.

Реализованная волатильность акций компаний. Для анализа была выбрана оценка реализованной волатильности: использовались переменные, учитывающие кратко-, средне- и долгосрочное поведение динамики акций. Все признаки реализованной волатильности и целевая переменная были прологарифмированы, так как имели скошенное распределение. Это было сделано с целью использовать в качестве функции потерь среднеквадратичное отклонения (MSE) для всех методов машинного обучения.

В приложении Б представлено распределение реализованной волатильности для каждой компании, в приложении В — гистограммы распределения логарифма дневной реализованной волатильности. Последние похожи на нормальное распределение, правые хвосты немного «тяжеловаты» для всех компаний, наиболее часто встречаемые значения находятся в пределах от -9 до -8 . Для базовой модели предобработка данных не осуществлялась, за исключением того, что все признаки (в том числе целевая переменная) были прологарифмированы.

В табл. 1 продемонстрированы обозначения для признаков реализованной волатильности. Идея гетерогенной авторегрессии реализованной волатильности состоит в вычислении оценок волатильности с использованием высокочастотных данных. Дневная реализованная волатильность оценивается на основе внутридневной 5-минутной доходности с использованием оценки медианной реализованной волатильности (MedRV).

Таблица 1

Обозначения для признаков реализованной волатильности

Наименование фактора	Обозначение
1. Реализованная волатильность в день t	RV_d
2. Лагированная волатильность за предыдущую неделю	RV_w
3. Лагированная волатильность за предыдущий месяц	RV_m
4. Логарифм реализованной волатильности	log_rv
5. Логарифм реализованной волатильности за предыдущий день	log_rv_d
6. Логарифм реализованной волатильности за предыдущую неделю	log_rv_w
7. Логарифм реализованной волатильности за предыдущую месяц	log_rv_m

Источник: составлено авторами.

Финансовые факторы. Далее производится анализ финансовых показателей. В итоговый набор факторов были включены 9 переменных (табл. 2). Полный перечень протестированных переменных представлен в приложении Г. Ряд переменных были исключены из последующего анализа по раз-

ным причинам: 1) наличие пропущенных данных; 2) проблема мультиколлинеарности; 3) незначимое влияние на зависимую переменную. В последующих работах можно подробнее изучить перечисленные факторы.

Все данные представляют собой временные ряды акций российских компаний, они были скачаны из сайта Investing.com и обработаны. В приложении Д представлена схема корреляций финансовых показателей для компании «Яндекс». Графики позволяют проверить переменные на мультиколлинеарность. Аналогичная работа была проделана и с другими компаниями. 4 из 9 показателя имеют корреляцию больше 0,3 по абсолютному значению для всех 7 компаний, другие 5 из 9 факторов обладают меньшей корреляцией. Предобработка данных заключалась в стандартизации масштабов данных (приведения признаков к одному масштабу). Стандартизация приводит все признаки к распределению с нулевым математическим ожиданием и единичной дисперсией:

$$\bar{\mu}_{\cdot j} = \frac{1}{n} \sum_{i=1}^n x_{ij},$$

$$\bar{\sigma}_{\cdot j} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{\mu}_{\cdot j})^2},$$

$$X^{new} = \frac{X - \bar{\mu}}{\bar{\sigma}}.$$

Для модели оценки состояния финансового рынка использовалась регрессия Lasso, так как такая модель:

- а) автоматически отбирает значимые признаки;
- б) позволяет налагать штрафы на очень маленькие веса.

Таблица 2

Финансовые показатели для исследования

Название переменной	Обозначение
Индекс IMOEX	IMOEX
Индекс MSCI Russia	MSCI
Доходность облигации России 10-летние (годовая ставка), %	Bond_LR
Доходность облигации Россия годовые (годовая ставка), %	Bond_SR
Курс доллара к рублю, руб.	USD_RUB
Доходность акции, %	yield
Объем торгов, руб.	volume_brand
Объем торгов индекса МосБиржи	volume_MOEX
Индекс потребительских цен	CPI

Источник: составлено авторами.

Отдельного внимания заслуживают показатели доходности 10-летних и годовых облигаций. Несмотря на несомненную научную ценность анализа облигаций с иными сроками погашения, цель данной модели заключалась в построении модели оценки финансового рынка, выявляющей взаимосвязь между информационными потоками и рынком облигаций в широком временном диапазоне. Учет прочих факторов, включенных в модель, обусловил выбор двух крайних временных горизонтов — одного года и десяти лет. Такой подход на основе контрастных сроков погашения обеспечил сбалансированную оценку взаимодействия долгового и информационного рынков.

Факторы внимания и настроения инвесторов. Для анализа реагирования новостей и сообщений в социальных сетях на волатильность фондового рынка была выбрана социальная сеть Twitter (X). Выбор Twitter в качестве основного источника текстовых данных определяется совокупностью факторов, соответствующих поставленным задачам исследования. Во-первых, высокая технологическая доступность Twitter обусловлена наличием открытых программных интерфейсов (API), что позволяет легко интегрировать данную социальную платформу в исследовательские процессы и обеспечить автоматизированный сбор значительных объемов необходимой информации. Во-вторых, особенности формата публикаций в Twitter (краткая структура записей, короткий временной интервал их появления и наличие встроенных метаданных) делают данный ресурс оптимальным для автоматизации процесса анализа эмоциональной составляющей и тональности высказываний, минимизируя трудности выделения значимой информации о поведении пользователей.

Принимая во внимание тот факт, что аудитория русскоязычных пользователей Twitter сравнительно мала (относительно, например, Telegram и VK), важно отметить, что сама природа сообщества данной социальной сети предполагает присутствие активного сегмента профессиональных экономистов, трейдеров и инвесторов, чье мнение может оказывать существенное воздействие на поведение участников российского фондового рынка. Именно такая целевая направленность площадки формирует особую значимость ее использования в рамках проводимого исследования.

Парсинг данных осуществлялся с помощью библиотеки Snsrcape. Источники данных делятся на две группы:

- 1) новостные страницы — выгружались все посты;
- 2) пользовательские посты — отбирались посты по ключевым словам (табл. 3).

Среди каналов получения информации есть как многомиллионные (Forbes и Ria), так и каналы с немногочисленной аудиторией. Данный шаг позволяет учитывать большую вариативность настроений аудитории. Переменные делятся на две группы: переменные внимания и перемен-

ные настроения, а также по принципу интегрируемости факторов: общие и специфичные. Под настроением инвесторов понимается отношение пользователей социальной сети к фондовому рынку в целом и к определенным компаниям; под вниманием инвесторов — общая заинтересованность инвестора к фондовому рынку и к тому или иному проекту (схожая логика используется в статье (Aouadi et al., 2013)). Общие факторы — факторы, одинаковые для всех компаний; специфичные факторы — факторы, характерные для определенной компании. Поиск специфичных для компаний постов осуществлялся по следующим словам (в соответствии с компанией, для которой строится модель): «новатэк», «мтс», «газпром», «лукойл», «роснефть», «яндекс», «сбербанк», «сбер». Все данные ежедневные. Переменные 3–6, 8–9 из табл. 4 сглаживались следующим образом (формула (1)):

$$\tilde{x}_t = 0,7x_t + 0,2x_{t-1} + 0,1x_{t-2}. \quad (1)$$

Исследование сфокусировано на анализе русскоязычных текстов, что обусловлено ключевой задачей работы — выявлением специфики эмоциональной окраски сообщений на русском языке.

Англоязычные элементы, в частности тикеры⁵, не подвергались обработке в рамках данного этапа, поскольку основной целью являлся анализ эмоциональной составляющей, а не извлечение конкретных финансовых данных и их характеристик. При этом необходимо признать, что тикеры и иные элементы текстов могут содержать релевантную информацию, учет которой предусматривается в рамках перспективных исследований при обработке текстовых данных.

Таким образом, учитывалось настроение не только за текущий день, но и за один и два дня до текущего дня. Для оценки переменных настроения строился классификатор тональности текста, который имеет два класса настроения: положительное и отрицательное. Для всех переменных настроения использовались метки классов, а не вероятностные значения: любому посту сопоставляется число из дискретного набора {0,1}.

⁵ Тикер (тикерный символ) — это стандартизированный буквенный код, присваиваемый эмитенту или финансовому инструменту для их однозначной идентификации в торговых и информационных системах фондового рынка (например, «SBER» для обозначения «Сбербанка»). В лингвистике и компьютерной лингвистике тикеры могут рассматриваться как особый класс именованных сущностей, релевантных для задач информационного поиска, классификации финансовых текстов и анализа тональности.

Источники данных по постам Twitter

Источники	Тип источника	Тематика	Данные	Аудитория, чел.
Новостные страницы	Общий	Финансы и инвестиции	InvestingRu ⁶	7,7 тыс.
			Finam ⁷	12,5 тыс.
			Finam signals ⁸	6,8 тыс.
			RBC invest ⁹	31 тыс.
	Общий	Новости и события	Meduza ¹⁰	1,3 млн
			RBC ¹¹	363 тыс.
			Forbes ¹²	16,6 млн
			Ria ¹³	2,7 млн
Пользовательские посты	Общий	Посты, содержащие ключевые слова «московская биржа», «мосбиржа»		
	Специфичный	Посты, содержащие ключевые слова по компаниям		

Примечание: численность аудитории представлена по состоянию на март 2022 г.

Источник: составлено авторами.

⁶ X (Twitter). Investing.com [@investingru]. Retrieved December 11, 2021, from <https://twitter.com/investingru>

⁷ X (Twitter). Finam Blog [@finam_blog]. Retrieved December 11, 2021, from https://twitter.com/finam_blog

⁸ X (Twitter). Finam Alert [@finamalert]. Retrieved December 11, 2021, from <https://twitter.com/finamalert>

⁹ X (Twitter). РБК Quote [@rbc_quote]. Retrieved December 11, 2021, from https://twitter.com/rbc_quote

¹⁰ X (Twitter). Meduza Project [@meduzaproject]. Retrieved December 11, 2021, from <https://twitter.com/meduzaproject>

¹¹ X (Twitter). РБК [@ru_rbc]. Retrieved December 11, 2021, from https://twitter.com/ru_rbc

¹² X (Twitter). Forbes [@forbes]. Retrieved December 11, 2021, from <https://twitter.com/forbes>

¹³ X (Twitter). РИА Новости [@rianru]. Retrieved December 11, 2021, from <https://twitter.com/rianru>

Переменные настроения и внимания

Общий / специфичный	Внимание / настроение	№ п/п	Наименование фактора	Обозначение
Общий	Внимание	1	Прирост логарифма количества постов	att_invest
		2	Логарифм количества постов	att_moex
	Настроение	3	Среднее настроение, сглаженное	sent_invest
		4	Стандартное отклонение настроения, сглаженное	std_invest
		5	Среднее настроение, сглаженное	sent_general
		6	Стандартное отклонение настроения, сглаженное	std_general
Специфичный	Внимание	7	Прирост логарифма количества постов	att_brand
	Настроение	8	Среднее настроение, сглаженное	sent_brand
		9	Стандартное отклонение настроения, сглаженное	std_brand

Источник: составлено авторами.

В приложении Е представлены схемы корреляции показателей внимания и настроения для компании «Яндекс». В среднем для всех компаний 2 из 9 признаков внимания и настроения имеют значимую корреляцию с целевой переменной (более 0,3). Стоит отметить, что переменная att_moex (логарифм количества постов, содержащих упоминание «Мосбиржи») имеет корреляцию с целевой переменной больше 0,3 для 6 из 7 компаний. Все показатели внимания и настроения численные.

Модель для тестирования

Обучение моделей и прогнозирование. В анализе используются: базовая модель (бенчмарк), модель оценки состояния финансового рынка и модель настроения и внимания. Базовая модель включает в себя только первые лаги кратко-, средне- и долгосрочной реализованной волатильности. Модель оценки состояния финансового рынка является обобщением базовой и включает также финансовые признаки. Итоговая модель настроения включает вышеупомянутые признаки, а также признаки, построенные на данных социальной сети Twitter.

Для базовой модели была построена линейная регрессия без использования регуляризации. Для модели оценки состояния финансового рынка и модели настроения были построены Lasso регрессия с регуляризацией L1, случайный лес, экстремальный градиентный бустинг (XGBoost) и расширенный градиентный бустинг (Light GBM).

Базовая модель. Базовая модель строится на сумме трех AR(1) процессов: реализованная волатильность в день t , средняя недельная реализованная волатильность со дня $t-4$ по день t и средняя месячная реализованная волатильность со дня $t-21$ по день t . Все три признака важны, поэтому для базовой модели построена обычная линейная регрессия (без регуляризации), чтобы избежать обнуления. Математически данная логика выглядит следующим образом (формула (2)):

$$\begin{aligned}\log RV_{t+1}^{(d)} &= c + \beta^{(d)} \log RV_t^{(d)} + \beta^{(w)} \log RV_t^{(w)} + \beta^{(m)} \log RV_t^{(m)} + \varepsilon_{t+1} = \\ &= c + \left(\log RV_t \right)' \beta_{RV} + \varepsilon_{t+1},\end{aligned}\quad (2)$$

где $\log RV_t^{(d)}$ — оценка реализованной волатильности в день t методом MedRV;

ε_{t+1} — ошибка предсказания;

$\left(\log RV_t \right)'$ — матрица признаков волатильности размера $n \times 3$;

n — количество наблюдений;

β_{RV} — вектор размерности 3×1 .

Модель оценки состояния финансового рынка. В этой модели происходит обобщение базовой модели, к ней добавляются финансовые показатели (формула (3)):

$$\log RV_{t+1}^{(d)} = c + \left(\log RV_t \right)' \beta_{RV} + E_t' \gamma_{eco} + \varepsilon_{t+1}, \quad (3)$$

где E_t' — матрица размера $n \times 9$, в которой каждый вектор-столбец — финансовый фактор;

γ_{eco} — вектор размерности 9×1 весов для финансовых признаков.

Добавляется регуляризация, и задача сводится к следующей оптимизационной задаче (формула (4)):

$$L_{lasso} \left(X_t, RV_{t+1}^{(d)}, W \right) = \left(\log RV_{t+1}^{(d)} - c - W' X_t \right)^2 + \lambda \|W\|_1 \rightarrow \min_{c, W}, \quad (4)$$

где X_t — матрица признаков размером $n \times 12$ (3 признака волатильности + 9 финансовых признаков);

W — вектор-столбец весов размера 12;

λ — коэффициент регуляризации (гиперпараметр);

$$\|W\|_1 = \sum_{j=1}^m |w_j| \text{ — норма весов.}$$

Модель настроения и внимания. Чтобы измерить факторы настроения, была обучена модель для прогнозирования тональностей ULMFiT. В качестве предобученной модели использовалась готовая модель Russian AWD-LSTM language model¹⁴, которая обучалась на корпусе Taiga. В корпусе Taiga содержится порядка 5 млрд слов. Для обучения модели была взята выборка текстов, источниками которых являются: Arzamas, Fontanka, Interfax, KP, Lenta и NPlus1. Обучение модели классификатора проводилось на размеченных данных из предыдущего пункта на корпусе коротких текстов Рубцовой (2012). Выборка была разделена на тестовую — 20%, обучающую — 64% (0,8:0,8) и валидационную — 16% (0,8:0,2). Подвыборки были разделены случайным образом, при этом соблюдалась балансировка классов. Для того чтобы сравнить, какой прирост в качестве прогноза дают факторы внимания и настроения, была построена регрессия Lasso. Таким образом, появляется возможность улучшить итоговую модель настроения, применив другие подходы машинного обучения: случайный лес, XGBoost и Light GBM. Модель настроения описывается следующим образом (формула (5)):

$$\log RV_{t+1}^{(d)} = c + (\log RV_t^{(d)})' \beta_{RV} + E_t' \gamma_{eco} + S_t' \theta_{sent} + \varepsilon_{t+1}, \quad (5)$$

где S_t' — матрица размера $n \times 9$, в которой каждый вектор-столбец — фактор внимания или настроения;

θ_{sent} — вектор размерности 9×1 весов для признаков внимания и настроения.

При добавлении регуляризации задача сводится к следующей оптимизационной задаче (формула (6)):

$$L_{lasso}(X_t, RV_{t+1}^{(d)}, W) = (\log RV_{t+1}^{(d)} - c - W' X_t)^2 + \lambda \|W\|_1 \rightarrow \min_{c, W}, \quad (6)$$

где X_t — матрица признаков размером $n \times 21$ (3 признака волатильности + 9 финансовых признаков + 9 признаков настроения и внимания);

W — вектор-столбец весов размера 21;

λ — коэффициент регуляризации (гиперпараметр);

$$\|W\|_1 = \sum_{j=1}^m |w_j| \text{ — норма весов.}$$

¹⁴ GitHub. Russian AWD-LSTM language model. Retrieved December 11, 2021, from <https://github.com/noise-field/Russian-ULMFiT>.

Далее для проверки построенных моделей был проведен сравнительный анализ моделей и анализ влияния факторов на реализованную волатильность:

1. Сравнение базовой модели, модели оценки состояния финансового рынка и модели настроения и внимания методом регрессии. Это помогает исследовать, дает ли добавление финансовых показателей в базовую модель прирост в качестве прогноза, а также позволяет пронаблюдать, как добавление признаков настроения и внимания улучшат качество прогноза по сравнению с базовой моделью и моделью оценки состояния финансового рынка.
2. Сравнение модели оценки состояния финансового рынка и моделей настроения и внимания всеми четырьмя методами между собой и друг с другом, что позволяет:
 - а) выявить оптимальный метод машинного обучения для модели оценки состояния финансового рынка и модели настроения и внимания для каждой компании;
 - б) проанализировать, в каких методах и для каких компаний модель настроения и внимания дает прирост в качестве прогноза в сравнении с моделью оценки состояния финансового рынка;
 - в) выбрать наилучший метод машинного обучения в пределах одной модели и сравнить лучшую модель оценки состояния финансового рынка и лучшую модель настроения и внимания для каждой отдельной компании.

В качестве функции потерь взята среднеквадратичная ошибка для всех моделей, так как целевая переменная имеет симметричное распределение.

Далее рассмотрим каждый шаг подробнее.

1. Поиск по сетке (Grid Search). Для поиска оптимального значения таких параметров используется поиск по сетке. В табл. 5 приведены все гиперпараметры¹⁵, которые были подобраны для моделей, а также сетка для их поиска: использовались методы поиска по сетке для подбора гиперпараметров в Lasso регрессии и в случайном лесе, а также метод байесовской оптимизации в экстремальном градиентном бустинге и в расширенном градиентном бустинге.

2. Байесовский оптимизатор (Bayesian Optimization). В отличие от поиска по сетке этот метод полагается на информацию, полученную моделью во время предыдущих оптимизаций с целью найти наиболее оптимизированный список параметров. Также, метод требует меньшего количества выборок для получения наилучших значений.

3. Кросс-валидация (Cross Validation). Для поиска гиперпараметров обучающая выборка должна разделяться еще на две: обучающую и вали-

¹⁵ Гиперпараметр модели — это параметр, который устанавливается не во время обучения, а задается извне.

дационную. Далее осуществляется одинаковая процедура для всех методов, где подбор гиперпараметра происходит через поиск по сетке, производится разделение обучающей выборки методом кросс-валидации. Учитывая особенности исследования (временные ряды), нет возможности использовать обычную перекрестную кросс-валидацию.

Таблица 5

Сетка гиперпараметров

Гиперпараметр	Название	Интервал
<i>Регрессия Lasso</i>		
Коэффициент регуляризации	alpha	[0,01, 10], шаг 0.5
<i>Случайный лес (Random Forest Regressor)</i>		
Минимальное количество объектов, необходимое для разделения внутреннего узла	min_samples_split	[2, 5], шаг 1
Минимальное количество объектов, необходимое в листе	min_samples_leaf	[1, 3], шаг 1
Количество признаков, которые следует учитывать при поиске лучшего разделения	max_features	$\left[d, \log_2 d, \frac{d}{3} \right]$, d – количество факторов
<i>XGBoost (XGB Regressor)</i>		
Количество деревьев с градиентным бустингом	n_estimators	[200, 500]
Максимальная глубина деревьев	max_depth	[3, 15]
Скорость обучения	learning_rate	[0,01, 1], 'log-uniform'
Отношение размера подвыборки к размеру выборки	subsample	[0,2, 1], 'uniform'
Минимальное снижение потерь, необходимое для дальнейшего разбиения на листовом узле дерева	gamma	[1e-9, 0,5], 'log-uniform'
Отношение столбцов подвыборки при построении каждого дерева	colsample_bytree	[0,5, 1], 'uniform'
<i>Light GBM</i>		
Количество деревьев с градиентным бустингом	n_estimators	[200, 500]
Максимальная глубина деревьев	max_depth	[3, 15]
Скорость обучения	learning_rate	[0,01, 1], 'log-uniform'
Отношение размера подвыборки к размеру выборки	subsample	[0,2, 1], 'uniform'

Гиперпараметр	Название	Интервал
Отношение столбцов подвыборки при построении каждого дерева.	colsample_bytree	[0,5, 1], 'uniform'
Максимальное количество ячеек, в которые помещены значения признаков	max_bin	[100, 1000]
Максимальное количество листьев на одном дереве	num_leaves	[2, 100]

Источник: составлено авторами.

Для временных рядов такой способ деления не подходит, так как нельзя использовать будущую информацию для прогноза в прошлом. Поэтому необходимо брать в обучающую выборку только те данные, которые известны до той даты, на которую делается прогноз. Выборка с 1 января 2021 г. по 28 февраля 2021 г. отложена для теста: при обучении модель не видит ее. Выборка с 01.01.2020 по 31.12.2020 (250 рабочих дней) будет разделена на 12 фолдов (частей данных) для кросс-валидации, чтобы подобрать оптимальные гиперпараметры для модели. Кросс-валидация происходит следующим образом: модель обучается на 9 первых фолдах, валидация проводится на фолде 10. Необходимо двигаться по скользящему окну, каждый раз имея одинаковую по размеру обучающую выборку в 9 фолдов (рис. 1). Подвыборки были разделены случайно, при этом соблюдалась балансировка классов (таким же образом, как в предыдущем пункте). Размеры выборок приведены в табл. 6.

1	2	3	4	5	6	7	8	9	10	11	12
Обучающая									Валид.		
	Обучающая									Валид.	
		Обучающая									Валид.
			Тестовая выборка								

Рис. 1. Деление на обучающую и тестовую (отложенную) выборку

Источник: составлено авторами.

Таблица 6

Размеры обучающей, валидационной и тестовой выборки

Обучающая выборка		Валидационная выборка		Тестовая выборка	
0	90926	0	22731	0	18186
1	92490	1	23123	1	18498

Примечание: 0 — метка «твит негативный», 1 — метка «твит позитивный».

Источник: составлено авторами.

Чтобы оценить качество классификатора использовалась отложенная тестовая выборка. Матрица ошибок (Confusion matrix) приведена на рис. 2: по строкам расположены реальные данные, по столбцам – прогнозные. Точность классификатора равна 75%.

В табл. 7 приведены другие метрики качества: точность, полнота, F1 score. Точность — это доля объектов, принадлежащих классу, относительно всех объектов, отнесенных классификатором к данному классу. Полнота — это доля объектов, отнесенная к данному классу классификатором, относительно всех объектов класса. В классификаторе точность и полнота примерно одинаковы, что является признаком того, что классификатор работает хорошо. С одной стороны, классификатор умеет правильно классифицировать в 79% случаев для обоих классов. С другой стороны, классификатор может выявить примерно 79% объектов каждого класса. Все три метрики приблизительно равны. F1 score учитывает сразу точность и полноту и рассчитывается следующим образом (формула (7)):

$$F1 = 2 \cdot \frac{\text{Точность} \cdot \text{Полнота}}{\text{Точность} + \text{Полнота}} \tag{7}$$

Матрица ошибок			
		Предсказанные	
		0	1
Текущие	0	14081	4105
	1	3688	14810

Рис. 2. Матрица ошибок классификатора для данных из тестовой выборки
 Источник: составлено авторами.

Таблица 7

Метрики качества: точность, полнота, F1 score для каждого класса по отдельности и их среднее значение

	Точность	Полнота	F1 – score
Негативный	0,79245	0,77428	0,78326
Позитивный	0,78298	0,80063	0,79170
Среднее	0,78771	0,78745	0,78748

Примечание: F1 также показывает неплохое качество классификатора.
 Источник: составлено авторами.

Поскольку все данные были стандартизированы, имеет смысл сравнивать важность признаков по значениям их весов. Для оценки значимости признаков в методах, основанных на решающих деревьях, использовалась метрика SHAP (Shapley Additive Explanations). Этот метод считается лучше

традиционных методов scikit-learn (обычно тем важнее признак, чем большее количество раз он входит в углы деревьев).

Результаты тестирования модели

Базовая модель. Для всех семи компаний краткосрочная и среднесрочная реализованные волатильности значимы на 1%-м уровне значимости. Средняя месячная реализованная волатильность — незначимый фактор ни для одной компании (даже на 10-м% уровне значимости). Коэффициенты и значения MSE для базовой модели приведены в табл. 8. Средняя реализованная волатильность за месяц вносит значительно меньший вклад по сравнению с другими признаками. При этом наблюдается положительная корреляция между реализованной волатильностью за предыдущий день и предыдущую неделю с целевой переменной.

Краткосрочная и среднесрочная волатильности являются важными факторами, влияющими на целевую переменную, что может быть связано с повышенной чувствительностью инвесторов и трейдеров к краткосрочным изменениям на рынке. Отсутствие значимости средней месячной волатильности может быть обусловлено ее меньшей чувствительностью к краткосрочным колебаниям, что делает ее менее полезной для прогнозирования. Положительная корреляция между реализованной волатильностью за предыдущий день и предыдущую неделю с целевой переменной указывает на инерцию рынка, когда текущие тенденции продолжают влиять на будущие движения. Данные результаты будут считаться бенчмарком (табл. 8, базовая модель).

Модель оценки состояния финансового рынка. В регрессии Lasso в среднем для всех компаний обнулился только один коэффициент. Некоторые признаки имеют вес, близкий к нулю, и, как правило, это переменные, имеющие слабую корреляцию с целевой переменной. Регуляризация обнулила веса следующих факторов для каждой компании: 1) «Яндекс»: MSCI, CPI; 2) «Сбербанк»: yield; 3) МТС: CPI; 4) «Газпром»: yield. Обнуление коэффициентов для переменных MSCI, CPI и yield указывает на то, что эти факторы не оказывают значительного влияния на целевую переменную, что может быть связано с тем, что они не коррелируют с динамикой финансового рынка или их влияние слишком слабое.

Зачастую наиболее важные признаки для модели оценки состояния финансового рынка (на основе результатов метода решающих деревьев) — это признаки реализованной волатильности, индекс МосБиржи, объем торгов, объем торгов индекса МосБиржи, курс доллара к рублю. В некоторых случаях в топ появляются доходности кратко- или долгосрочных облигаций, которые могут отражать ожидания рынка относительно будущей инфляции и экономической стабильности. Включение этих показателей в модель позволяет учитывать макроэкономические факторы, что улучшает точность прогнозов.

В табл. 8 (модель оценки состояния финансового рынка) приведены результаты прогноза модели на тестовой выборке. Как можно заметить, ошибка в среднем стала меньше с добавлением финансовых показателей.

Таблица 8

Значения среднеквадратичной ошибки MSE для базовой и модели оценки состояния финансового рынка на тестовой выборке для каждой отдельной компании

Компания	Базовая модель					Модель оценки состояния финансового рынка			
	MSE	intercept	log_rv_d	log_rv_w	log_rv_m	Lasso	RF	XGB	LGBM
«Яндекс»	0,326	-1,469	0,472	0,300	0,043	0,160	0,123	0,144	0,132
«Сбербанк»	0,220	-1,585	0,376	0,510	-0,070	0,213	0,207	0,202	0,189
МТС	0,187	-1,058	0,354	0,563	-0,037	0,255	0,233	0,261	0,283
«Лукойл»	0,208	-1,135	0,496	0,408	-0,041	0,193	0,213	0,192	0,200
«Роснефть»	0,196	-1,121	0,395	0,469	0,002	0,234	0,186	0,183	0,214
«Газпром»	0,254	-1,352	0,446	0,380	0,016	0,233	0,231	0,219	0,223
«Новатэк»	0,247	-0,993	0,436	0,435	0,004	0,263	0,223	0,238	0,250

Примечание: серым цветом выделен наилучший метод для каждой компании в модели оценки состояния финансового рынка.

Источник: составлено авторами.

Модель настроения и внимания. Стоит отметить, что в сравнении с аналогичной моделью оценки состояния финансового рынка, большее количество признаков было обнулено регуляризацией (в среднем по всем компаниям регрессия Lasso отобрало четыре финансовых показателя). Это может быть связано с тем, что факторы настроения и внимания менее стабильны и могут иметь более низкую корреляцию с целевой переменной.

Из факторов внимания и настроения регуляризация отобрала в среднем два признака, зачастую это att_moex — логарифм количества постов, содержащих упоминание Мосбиржи и att_invest — прирост логарифма количества постов по инвестиционным страницам. Это указывает на то, что количество постов может быть важным индикатором настроения и активности на рынке.

Регуляризация отобрала следующие признаки настроения и внимания для каждой компании: 1) «Яндекс»: att_invest; 2) «Сбербанк»: att_invest, att_moex; 3) МТС: att_invest, att_moex; 4) «Лукойл»: att_brand, att_moex; 5) «Роснефть»: att_invest, att_brand, att_moex; 6) «Газпром»: att_invest, att_moex; 7) «Новатэк»: att_brand, att_moex.

Ни один из признаков, связанных с тональностью твитов, не был выбран регрессией Lasso. Все отобранные признаки связаны с изменением внимания, т.е. изменением в количестве постов. Однако для моделей, ос-

нованных на деревьях, важность признаков по тональности существеннее, чем для Lasso регрессии. Зачастую std_invest (стандартное отклонение в настроении по инвестиционным страницам) стоит в топе важности признаков для разных методов и компаний. В табл. 9 приведены результаты прогноза модели настроения на тестовой выборке. Лучшие результаты дали модели случайного леса и XGBoost. Добавление факторов настроения и внимания не дали явный и ощутимый прирост в улучшении прогноза, однако, большинство моделей уменьшило (хоть и в незначительной степени) ошибку на тесте.

Таблица 9

Значения среднеквадратичной ошибки MSE для модели настроения на тестовой выборке для каждой отдельной компании

Компания	Lasso	RF	XGB	LGBM	MSE по базовой модели	MSE по экон. модели
«Яндекс»	0,158	0,121	0,118	0,123	0,326	0,123 (RF)
«Сбербанк»	0,206	0,200	0,183	0,208	0,220	0,189 (LGBM)
МТС	0,222	0,239	0,267	0,284	0,187	0,233 (RF)
«Лукойл»	0,212	0,184	0,186	0,195	0,208	0,192 (XGB)
«Роснефть»	0,218	0,182	0,192	0,185	0,196	0,183 (XGB)
«Газпром»	0,241	0,227	0,215	0,224	0,254	0,219 (XGB)
«Новатэк»	0,260	0,218	0,257	0,224	0,247	0,223 (RF)

Примечание: по столбцам слева направо — регрессия Lasso, случайный лес, экстремальный градиентный бустинг, расширенный градиентный бустинг. В последних двух столбцах для сравнения также представлены MSE по базовой модели и MSE по модели оценки состояния финансового рынка.

Источник: составлено авторами.

Сравнение базовой модели, модели оценки состояния финансового рынка и модели настроения и внимания. В табл. 10 продемонстрировано сравнение базовой и моделей оценки состояния финансового рынка, модели настроения и базовой модели. В плохом смысле выделяется только компания МТС, для нее финансовые факторы не дали прироста в точности. Отметим, что отрицательные значения свидетельствуют о снижении точности прогноза. «Яндекс», «Сбербанк» и «Газпром» хорошо отреагировали на добавление финансовых факторов. «Лукойл», «Роснефть» и «Новатэк» дали неоднозначные результаты, хотя для 3 из 4 моделей добавление финансовых факторов оказало положительное влияние на качество прогноза. В среднем, если не брать в расчет МТС, модель оценки состояния финансового рынка дает прирост в 13% в качестве прогноза по сравнению с бенчмарком, а модель настроения — прирост в 15%.

Для более детального анализа следует сравнивать модель оценки состояния финансового рынка и модель настроений. Из табл. 11 отметим компа-

нию «Яндекс», где каждый алгоритм с добавлением признаков по данным социальных сетей улучшил точность прогноза. Для «Сбербанка», МТС, «Лукойла», «Роснефти», «Газпрома» и «Новатэка» добавление признаков внимания и настроения улучшили результат для 3 из 4 моделей.

Таблица 10

**Уменьшение среднеквадратичной ошибки MSE на тестовой выборке
в модели оценки состояния финансового рынка и модели настроения
относительно базовой модели, %**

Компания	В модели оценки состояния финансового рынка				В модели настроения			
	Lasso	RF	XGB	LGBM	Lasso	RF	XGB	LGBM
«Яндекс»	51	62	56	60	52	63	64	62
«Сбербанк»	3	6	8	14	6	9	17	5
МТС	-36	-25	-40	-51	-19	-28	-43	-52
«Лукойл»	8	-2	8	4	-2	12	11	6
«Роснефть»	-19	5	7	-9	-11	7	2	6
«Газпром»	8	9	14	12	5	11	15	12
«Новатэк»	-6	10	4	-1	-5	12	-4	9

Примечание: по столбцам слева направо — регрессия Lasso, случайный лес, экстремальный градиентный бустинг, расширенный градиентный бустинг.

Источник: составлено авторами.

Таблица 11

**Уменьшение среднеквадратичной ошибки MSE
на тестовой выборке в модели настроения относительно модели
оценки состояния финансового рынка, %**

Компания	Lasso	RF	XGB	LGBM
«Яндекс»	1	2	18	7
«Сбербанк»	3	3	9	-10
МТС	13	-3	-2	0
«Лукойл»	-10	14	3	3
«Роснефть»	7	2	-5	14
«Газпром»	-3	2	2	0
«Новатэк»	1	2	-8	10

Примечание: по столбцам слева направо — регрессия Lasso, случайный лес, экстремальный градиентный бустинг, расширенный градиентный бустинг.

Источник: составлено авторами.

В среднем для всех компаний добавление признаков настроения и внимания к модели оценки состояния финансового рынка улучшают качество прогноза на 3%. Для дальнейшего сравнения были отобраны лучшая модель оценки состояния финансового рынка и лучшая модель настроения. Лучшие алгоритмы прогнозирования для модели оценки состояния финансового рынка и модели настроения, их ошибки на прогнозе и прирост в улучшении прогноза для модели настроения относительно модели оценки состояния финансового рынка приведены в табл. 12.

Включение весов настроения и внимания в целом положительно влияет на точность прогноза. Уменьшение в MSE на тесте незначительное, но все-таки присутствует. Для компании «Яндекс» информация о настроении и внимании в социальных сетях имеет большую объяснительную силу: факторы внимания и настроения улучшают качество прогноза в среднем по всем методам на 7% относительно модели оценки состояния финансового рынка. Далее следует «Сбербанк», для которого, при выборе оптимального метода модель настроения, улучшает качество прогноза на 3% в сравнении с моделью оценки состояния финансового рынка. Исследуемые компании, связанные с нефтедобычей и газодобычей, хуже реагируют на факторы социальных сетей, и даже при лучшем методе дают прирост качества прогноза в 2% в среднем.

Таблица 12

**Лучшие алгоритмы прогнозирования для модели оценки
состояния финансового рынка и модели настроения и внимания,
сравнение качества моделей**

	Базовая модель	Модель оценки состояния финансового рынка		Модель настроения		Модель настроения/ модель оценки состояния финансового рынка	Модель настроения/ базовая модель
Компания	MSE	Метод	MSE	Метод	MSE	Уменьшение MSE, %	Уменьшение MSE, %
«Яндекс»	0,326	RF	0,123	XGB	0,118	4	64
«Сбербанк»	0,220	LGBM	0,189	XGB	0,183	3	17
МТС	0,187	RF	0,233	Lasso	0,222	5	-19
«Лукойл»	0,208	XGB	0,191	RF	0,184	4	12
«Роснефть»	0,196	XGB	0,183	RF	0,182	1	7
«Газпром»	0,254	XGB	0,219	XGB	0,215	2	15
«Новатэк»	0,247	RF	0,223	RF	0,218	2	12

Источник: составлено авторами.

В качестве примера, далее рассмотрим компанию «Яндекс» и проанализируем важность признаков в моделях со случайным лесом и XGBoost при помощи метрики SHAP (рис. 3)¹⁶.

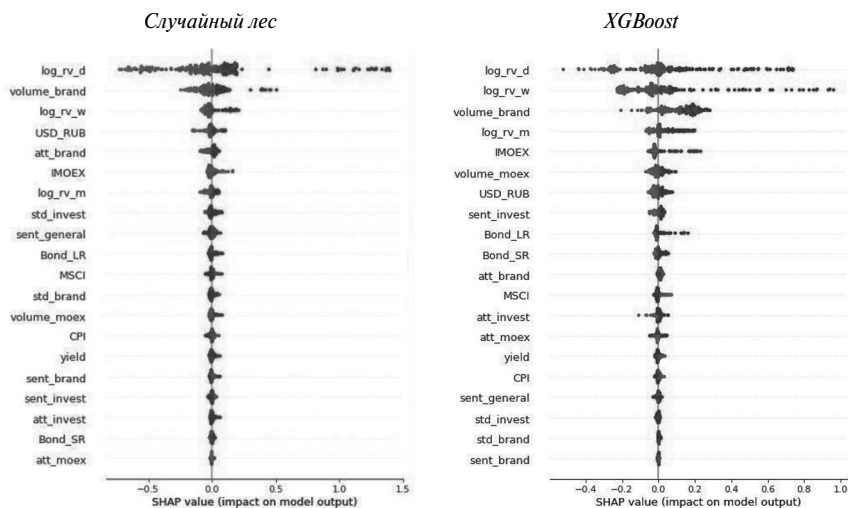


Рис. 3. Линейчатые диаграммы, отображающие важность признаков методом SHAP для модели настроения и внимания компании «Яндекс»
Источник: составлено авторами.

SHAP в моделях на основе дерева может давать двум одинаково важным характеристикам разные оценки в зависимости от того, какой уровень разделения был выполнен с использованием этих факторов. Признаки, которые первыми разделяют модель, могут иметь большее значение (Chan et al., 2023). SHAP вычисляет локальную важность признака для каждого наблюдения, обеспечивая единую основу для интерпретации предсказаний модели и атрибуции признаков (Hickey et al., 2021).

Согласно результатам, большой вклад вносят признаки `log_rv_d` (логарифм реализованной дневной волатильности), `log_rv_w` (логарифм реализованной недельной волатильности) и `volume_brand` (объем торгов). Самые важные признаки по данным социальных сетей — `sent_invest` (тональность твитов на инвестиционных страницах) и `att_brand` (изменение в количестве твитов, содержащих упоминание компании), причем они занимают достаточно высокое место (5-е и 8-е соответственно).

Высокая дневная волатильность может указывать на повышенную неопределенность и риск на рынке, поэтому инвесторы могут восприни-

¹⁶ Аналогичная работа была проделана и с другими компаниями.

мать высокую волатильность как сигнал о нестабильности, что, в свою очередь, может повлиять на их решения о покупке или продаже акций. Недельная волатильность обычно менее подвержена краткосрочным колебаниям и может лучше отражать долгосрочные тренды: инвесторы могут использовать этот показатель для оценки долгосрочной стабильности компании и ее акций. Объем торгов (volume_brand) показывает активность на рынке и может быть индикатором интереса к конкретным акциям: высокий объем торгов может указывать на повышенный интерес к компании или сектору, что может привести к росту цен на акции, в то время как низкий объем может свидетельствовать о снижении интереса и потенциальном снижении цен.

Заключение

В работе измерялись сигналы настроения и внимания, связанные с отдельными акциями и индексом фондового рынка, использовались текстовые данные из социальной сети Twitter; проверялась гипотеза о том, что модель настроения и внимания может улучшать качество прогноза реализованной волатильности российских компаний.

Для проведения анализа рассматривались различные вариации реализованной волатильности, которые характеризуют кратко-, средне- и долгосрочную память. В последующих работах также можно проверить построенные модели на основе данных других социальных сетей: Telegram, VK, «Одноклассники» и др. Работа с данными из VK или «Одноклассников» обладает определенными сложностями, например, есть проблемы с выгрузкой данных, предобработкой, предобучением и т.д. Подробнее об обработке текстовых данных из социальной сети VK можно прочитать в статье (Shulyak, 2022).

Результаты проведенного исследования представляют особую ценность для широкого круга инвесторов, заинтересованных в объективной оценке риск-факторов, обусловленных изменениями информационного поля. Однако, несмотря на продемонстрированную перспективность предложенной модели, существует необходимость дальнейшего совершенствования ее прикладного аспекта. Реализация полной функциональности требует дополнительного раскрытия деталей построения комплексной системы, основанной на итогах исследования. Такой подход позволил бы обеспечить полноценное понимание механизмов взаимодействия информационного окружения и динамических характеристик фондовых инструментов. Помимо расширения теоретико-методологического раздела, актуальным представляется дополнение работы описанием практических шагов, необходимых для внедрения разработанной методологии в реальный инвестиционный процесс. Это подразумевает создание четких инструкций по определению доминирующих настроений на рынке, формализации

критериев выбора инвестиционного инструмента и применению полученного знания для выработки оптимальных стратегических решений.

Список литературы

Аганин, А. Д. (2017). Сравнение GARCH и HAR-RV моделей для прогноза реализованной волатильности на российском рынке. *Прикладная эконометрика*, 48, 63–84.

Интерфакс. (2024, 26 декабря). Медиарейтинг компаний 2024 год. SCAN-Интерфакс. Дата обращения 26.01.2026, <https://scan-interfax.ru/ratings/mediareiting-kompanii-2024-god-2/>.

Интерфакс. (2023, 28 декабря). Медиарейтинг компаний 2023 год. SCAN-Интерфакс. Дата обращения 26.01.2026, <https://scan-interfax.ru/ratings/mediareiting-kompanii-2023-god-2/>.

Интерфакс. (2022, 27 декабря). Медиарейтинг компаний 2022 год. SCAN-Интерфакс. Дата обращения 26.01.2026, <https://scan-interfax.ru/ratings/mediareiting-kompanij-2022-god/>.

Интерфакс. (2021, 27 декабря). Рейтинг медиа-заметности компаний 2021 год. SCAN-Интерфакс. Дата обращения 26.01.2026, <https://scan-interfax.ru/ratings/rejting-media-zametnosti-kompanij-za-2021-god/>.

Интерфакс. (2020, 28 декабря). Рейтинг медиа-заметности компаний 2020 год. SCAN-Интерфакс. Дата обращения 26.01.2026, <https://scan-interfax.ru/ratings/rejting-media-zametnosti-kompanij-2020/>.

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), 53–84. <https://doi.org/10.31737/2221-2264-2022-53-1-3>.

Рубцова, Ю. (2012). Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора. *Инженерия знаний и технологии семантического веба*, 1, 109–116.

Ahmed, H. J. A., Hassan, A., & Nasir, A. M. (2005). The relationship between trading volume, volatility and stock market returns: A test of Mixed Distribution Hypothesis for a Pre and Post crisis on Kuala Lumpur Stock Exchange. *Investment Management and Financial Innovations*, 3(3), 146–158.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, 59(3), 1259–1294.

Aouadi, A., Arouri, M., & Teulon, F. (2013). Investor attention and stock market activity: Evidence from France. *Economic Modelling*, 35, 674–681.

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.

Caporin, M., & Poli F. (2017). Building news measures from textual data and an application to volatility forecasting. *Econometrics*, 5(3).

Chan, J. Y. L., Bea, K. T., & Leow S. M. H. (2023). State of the art: a review of sentiment analysis based on sequential transfer learning. *Artificial Intelligence Review*, 56, 749–780. <https://doi.org/10.1007/s10462-022-10183-8>.

Corsi, F., Audrino, F., & Reno, R. (2012). HAR Modeling for Realized Volatility Forecasting. In: *Handbook of Volatility Models and Their Applications*, New Jersey, USA: John Wiley & Sons, Inc., 363–382.

- Da, Z., Engelberg J., & Gao P. (2011). In search of attention. *The Journal of Finance*, 66(5), 1461–1499.
- Da, Z., Engelberg, J., & Gao P. (2015). The sum of all FEARS investor sentiment and asset prices. *Review of Financial Studies*, 28(1), 1–32.
- Daniel, K., Hirshleifer, D., & Teoh S. H. (2002). Investor psychology in capital markets: evidence and policy implications. *Journal of Monetary Economics*, 49(1), 139–209.
- Dimpfl, T., & Jank S. (2012). Can Internet Search Queries Help to Predict Stock Market Volatility? *Paris December 2012 Finance Meeting EUROFIDAI-AFFI Paper, European Financial Management, Forthcoming*.
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417.
- Hamid, A., & Heiden, M. (2015). Forecasting volatility with empirical similarity and Google trends. *Journal of Economic Behavior & Organization*, 117, 62–81.
- Hickey, J. M., Di Stefano, P. G., & Vasileiou V. (2021). Fairness by explicability and adversarial SHAP learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 174–190. https://doi.org/10.1007/978-3-030-67664-3_11.
- Mao, H., Counts, S., & Bollen J. (2011). Predicting financial markets: Comparing survey, news, Twitter and search engine data. *Computer Science*, 10.
- Obaidi, M., Nagel, L., Specht, A., & Klünder, J. (2022). Sentiment analysis tools in software engineering: A systematic mapping study. *Information and Software Technology*, 151. <https://doi.org/10.1016/j.infsof.2022.107018>.
- Oliveira, N., Cortez, P., & Areal, N. (2013). On the predictability of stock market behavior using StockTwits sentiment and posting volume. In: *L. Correia, L. P. Reis, & J. Cascalho (Eds.), Progress in artificial intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg*, 355–365.
- Revathy, G., Alghamdi, S., Alahmari, S., Yonbawi, S., Kumar, A., & Haq, M. (2022). Sentiment analysis using machine learning: Progress in the machine intelligence for data science. *Sustainable Energy Technologies and Assessments*, 53. <https://doi.org/10.1016/j.seta.2022.102557>.
- Shulyak, E. (2022). Macroeconomic Forecasting Using Data from Social Media. *Russian Journal of Money and Finance*, 81(4), 86–112.
- Tseng, K. C. (2006). Behavioral finance, bounded rationality, neurofinance, and traditional finance. *Investment Management and Financial Innovations*, 3(4), 7–18.
- Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55, 5731–5780. <https://doi.org/10.1007/s10462-022-10144-1>.

References

- Aganin, A. D. (2017). Comparison of GARCH and HARV models for forecasting realized volatility in the Russian market. *Applied Econometrics*, 48, 63–84. (In Russ.)
- Interfax. (2024, December 26). Media and Social Media Monitoring Reputation Management System. Retrieved January 1, 2026, from <https://scan-interfax.ru/ratings/mediareiting-kompanii-2024-god-2/> (In Russ.)
- Interfax. (2023, December 28). Media and Social Media Monitoring Reputation Management System. Retrieved January 28, 2026, from <https://scan-interfax.ru/ratings/mediareiting-kompanii-2023-god-2/> (In Russ.)

Interfax. (2022, December 27). Media and Social Media Monitoring Reputation Management System. Retrieved January 26, 2026, from <https://scan-interfax.ru/ratings/mediarejting-kompanij-2022-god/> (In Russ.)

Interfax. (2021, December 27). Media and Social Media Monitoring Reputation Management System. Retrieved January 26, 2026, from <https://scan-interfax.ru/ratings/rejting-media-zametnosti-kompanij-za-2021-god/> (In Russ.)

Interfax. (2020, December 28). Media and Social Media Monitoring Reputation Management System. Retrieved January 26, 2026, from <https://scan-interfax.ru/ratings/rejting-media-zametnosti-kompanij-2020> (In Russ.)

Rubtsova, Yu. (2012). Automatic construction and analysis of a corpus of short texts (microblogging posts) for the task of developing and training a tone classifier. *Knowledge engineering and semantic web technologies*, 1, 109–116. (In Russ.)

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. <https://doi.org/10.31737/2221-2264-2022-53-1-3> (In Russ.)

Приложение А

Таблица А.1

Медиарейтинг исследуемых компаний за 2020–2024 гг.

	Позиция в рейтинге	Индекс заметности компаний	Упоминания в СМИ	Охват аудитории
2024 г.				
«Сбербанк»	2	89 207 904	572 324	481 262 578
«Газпром»	6	44 105 052	320 359	394 892 533
«Яндекс»	10	25 143 416	231 813	389 935 279
МТС	13	22 596 188	164 069	305 133 366
«Роснефть»	19	16 140 821	103 785	315 294 061
«Лукойл»	20	15 259 393	121 032	302 490 352
«Новатэк»	46	6 906 923	63 388	239 656 328
2023 г.				
«Сбербанк»	2	92 106 808	563 822	480 348 507
«Газпром»	7	42 271 736	290 801	399 099 858
«Яндекс»	11	23 583 756	210 535	370 738 675
МТС	13	20 104 246	139 001	296 304 416
«Роснефть»	17	17 203 984	105 561	344 740 792
«Лукойл»	25	13 188 673	96 888	273 799 105
«Новатэк»	36	9 571 476	63 251	284 081 643
2022 г.				
«Сбербанк»	2	93 699 920	642 380	453 951 591
«Газпром»	1	109 303 952	639 764	444 475 048
«Яндекс»	10	27 180 652	300 777	417 899 891
МТС	16	15 596 154	120 650	201 454 986
«Роснефть»	14	20 008 940	122 186	341 888 142
«Лукойл»	15	17 080 980	103 644	352 064 181
«Новатэк»	29	10 438 297	60 107	259 483 250
2021 г.				
«Сбербанк»	1	120 087 776	-	477 911 179
«Газпром»	3	104 114 816	-	451 133 519

	Позиция в рейтинге	Индекс заметности компании	Упоминания в СМИ	Охват аудитории
«Яндекс»	6	38 286 756	-	395 512 967
МТС	29	14 454 590	-	271 728 315
«Роснефть»	16	25 057 750	-	340 343 496
«Лукойл»	23	16 567 838	-	289 404 905
«Новатэк»	17	11 477 023	-	336 605 354
2020 г.				
«Сбербанк»	2	127 970 864	-	-
«Газпром»	3	73 697 632	-	-
«Яндекс»	5	52 043 492	-	-
МТС	24	14 031 774	-	-
«Роснефть»	10	29 843 492	-	-
«Лукойл»	21	16 165 449	-	-
«Новатэк»	36	9 149 078	-	-

Источник: составлено авторами на основе данных сайта Интерфакс.

Приложение Б

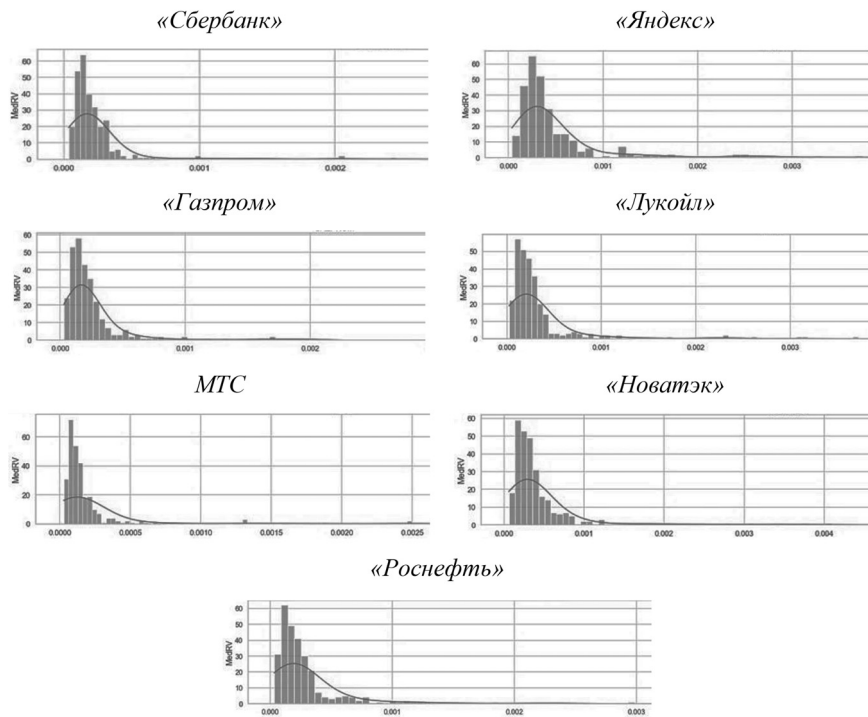


Рис. Б.2. Распределение оценки реализованной волатильности для каждой компании
Источник: составлено авторами.

Приложение В

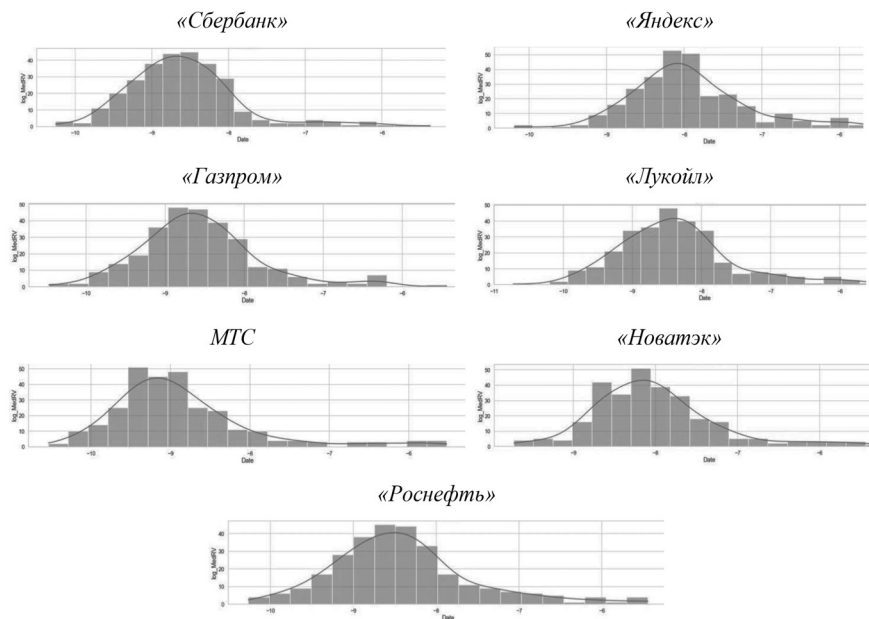


Рис. В.3. Распределение логарифмированной оценки реализованной волатильности для каждой компании

Источник: составлено авторами.

Полный перечень переменных, протестированных в работе

№ п/п	Наименование фактора	Обозначение	Общий / специфичный	Источник
2.1.1	Индекс IMOEX	211_IMOEX	Общий	https://ru.investing.com/indices/mcx-historical-data
2.1.2	Индекс RV1	212_RV1_level	Общий	https://ru.investing.com/equities/retail-value-historical-data
2.1.3	Изменение в Индексе RV1, %	213_RV1_change	Общий	https://ru.investing.com/equities/retail-value-historical-data
2.1.4	Индекс MSCI Russia	214_MSCI	Общий	https://ru.investing.com/indices/msci-russia-historical-data
2.1.5	Доходность акции, %	215_Return	Специфичный	https://ru.investing.com/
2.1.6	Объем торгов, руб.	216_Volume	Специфичный	https://ru.investing.com/
2.1.7	Dow Jones Industrial Average	216_DJ	Общий	https://ru.investing.com/indices/us-30-historical-data
2.1.8	Индекс РТС	217_IRTS	Общий	https://ru.investing.com/indices/rtsi
2.2.1	Доходность облигации России 10-летние (годовая ставка), %	221_OFZ_rate_LR	Общий	https://ru.investing.com/rates-bonds/russia-10-year-bond-yield-historical-data
2.2.2	Доходность облигации Россия годовые (годовая ставка), %	222_OFZ_rate_SR	Общий	https://ru.investing.com/rates-bonds/russia-1-year-bond-yield-historical-data

№ п/п	Наименование фактора	Обозначение	Общий / специфичный	Источник
2.2.3	Дневное изменение в доходности облигации Россия 10-летние (годовая ставка), %	223_Return_OFZ_LR	Общий	https://ru.investing.com/rates-bonds/russia-10-year-bond-yield-historical-data
2.2.4	Дневное изменение в доходности облигации Россия годовые (годовая ставка), %	224_Return_OFZ_SR	Общий	https://ru.investing.com/rates-bonds/russia-1-year-bond-yield-historical-data
2.2.5	Разница между доходностью 10-летней облигации и годовой	225_Yield_diff_LR_SR	Общий	https://ru.investing.com/rates-bonds/russia-10-year-bond-yield-historical-data https://ru.investing.com/rates-bonds/russia-1-year-bond-yield-historical-data
2.3.1	Курс доллара к рублю, руб.	231_USD/RUB	Общий	https://ru.investing.com/currencies/usd-rub-historical-data
2.4.1	Объем торгов индекса МосБиржи	241_IMOEX_vol	Общий	https://ru.investing.com/indices/mcx-historical-data
2.4.2	Объем торгов Dow Jones Industrial Average	242_DJI_vol	Общий	https://ru.investing.com/indices/us-30-historical-data
2.5.1	Индекс потребительских цен (ИПЦ)	251_CPI	Общий	https://ru.investing.com/economic-calendar/russian-cpi-1180
2.5.2	Ожидаемый индекс потребительских цен (ОИПЦ)	252_CPI_Expected	Общий	https://ru.investing.com/economic-calendar/russian-cpi-1180
2.5.3	Объем промышленного производства в России	253_Indst_production	Общий	https://ru.investing.com/economic-calendar/russian-industrial-production-553
2.5.4	Денежная масса, млрд руб.	254_M1	Общий	https://www.cbr.ru/statistics/macro_itm/dkfs/

№ п/п	Наименование фактора	Обозначение	Общий / специфичный	Источник
2.5.5	Денежная масса сезонно скорректированная, млрд руб.	255_M1_SA	Специфичный	https://www.cbr.ru/statistics/macro_itm/dkfs/
2.5.6	Первая разность денежной массы, млрд руб.	256_FD_M2	Специфичный	https://www.cbr.ru/statistics/macro_itm/dkfs/
2.5.7	Первая разность денежной массы сезонно скорректированной, млрд руб.	257_FD_M2_SA	Общий	https://www.cbr.ru/statistics/macro_itm/dkfs/
2.5.8	Доходность индекса CRB	258_Return_CRB	Общий	https://ru.investing.com/indices/thomson-reuters---jefferies-crb-historical-data
2.5.9	Уровень безработицы, %	259_Unempl	Общий	https://ru.investing.com/economic-calendar/russian-unemployment-rate-556
2.5.10	Индекс производственной активности PMI России	2510_PMI	Общий	https://ru.investing.com/economic-calendar/russian-market-manufacturing-pmi-1630
2.5.11	Фьючерс на нефть Brent	_Brent_oil	Общий	https://ru.investing.com/commodities/brent-oil-historical-data

Источник: составлено авторами.

Приложение Д

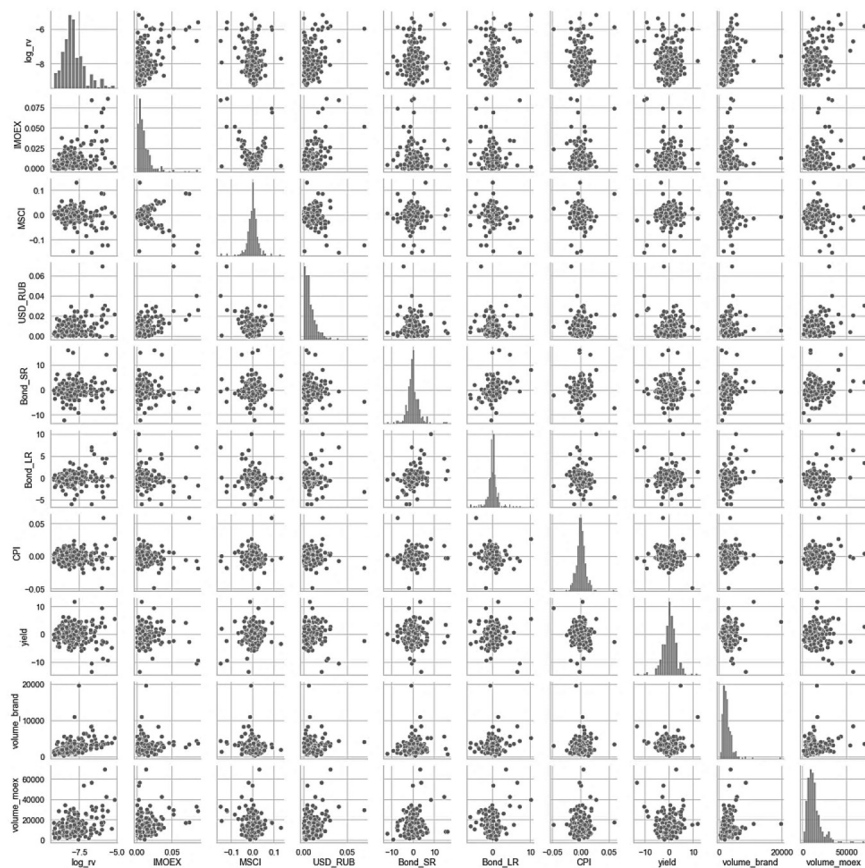


Рис. Д.5. Схема корреляций финансовых показателей для компании «Яндекс»

Источник: составлено авторами.

Приложение Е

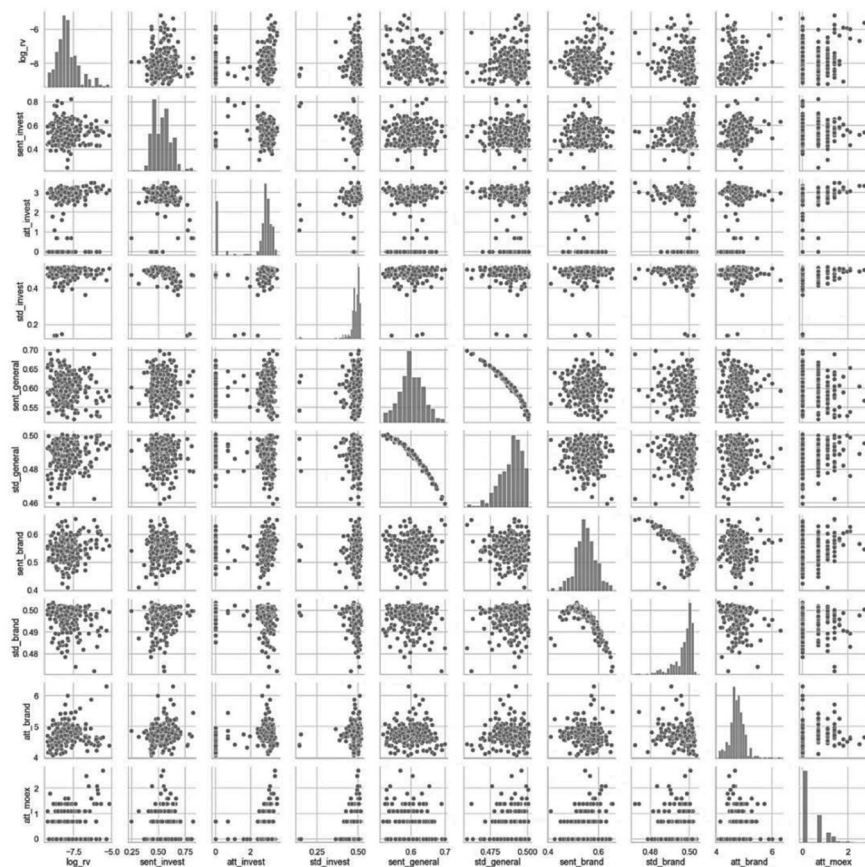


Рис. Е.6. Схема корреляций показателей модели внимания и настроения для компании «Яндекс»

Источник: составлено авторами.