УДК 336.761.4

5.2.2. Математические, статистические и инструментальные методы в экономике

АНАЛИЗ ТОНАЛЬНОСТИ В МОДЕЛЯХ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ЦЕН АКЦИЙ НА КИТАЙСКОМ ФОНДОВОМ РЫНКЕ

Чжан Синьхао
Аспирант
РИНЦ-SCIENCE INDEX SPIN-код: 7585-0746
st106398@student.spbu.ru
Санкт-Петербургский Государственный Университет, г. Санкт-Петербург, Россия

Ласкин Михаил Борисович
д.э.н., главный научный сотрудник, профессор,
РИНЦ-SCIENCE INDEX SPIN-код: 3457-6998
laskin.m@iias.spb.su
Санкт-Петербургский ФИЦ РАН, Санкт-Петербургский Государственный Университет, г. Санкт-Петербург, Россия

В статье рассматривается применение анализа тональности инвесторов в моделях прогнозирования цен акций на китайском рынке акций класса А. Цель состоит в повышении точности прогнозирования за счёт интеграции многоисточниковых индикаторов тональности. На основании собранных комментариев новостных каналов и на инвесторских форумах за период с 2020 по 2024 годы построены индикаторы тональности, чтобы максимально сохранить информацию и снизить субъективную предвзятость. При экспериментировании с моделями LSTM и GRU прогнозирование цен акций проводились с различными комбинациями индикаторов тональности. Результаты показали наилучшую прогностическую способность у модели GRU. Включение в модель первой главной компоненты, как обобщённого индикатора тональности снизило ошибку прогнозирования и улучшило качество модели. Исследование показывает, что на китайском рынке учёт тональности инвесторов может улучшить прогностические возможности моделей

Ключевые слова: АНАЛИЗ ТОНАЛЬНОСТИ, ПРОГНОЗ ЦЕН АКЦИЙ, КИТАЙСКИЙ ФОНДОВЫЙ РЫНОК, ГЛУБОКОЕ ОБУЧЕНИЕ, НАСТРОЕНИЯ ИНВЕСТОРОВ

UDC 336.761.4

5.2.2. Mathematical, statistical and instrumental methods in economics

SENTIMENT ANALYSIS IN DEEP LEARNING MODELS FOR PREDICTING STOCK PRICES IN THE CHINESE STOCK MARKET

Zhang Xinhao
Post-graduate student,
RSCI-SCIENCE INDEX SPIN-code: 7585-0746
st106398@student.spbu.ru
St. Petersburg University, St. Petersburg, Russia

Laskin Mikhail Borisovich
Dr.Sci.Econ., Chief Scientific Officer, professor,
RSCI-SCIENCE INDEX SPIN-code: 3457-6998
laskin.m@iias.spb.su
Federal Research Center of the Russian Academy of Sciences, St. Petersburg University, St. Petersburg, Russia

The article discusses the application of investor sentiment analysis in stock price forecasting models in the Chinese Class A stock market. The goal is to improve forecasting accuracy by integrating multi-source sentiment indicators. We have collected news and comments on investor forums for the period from 2020 to 2024 and built indicators of sentiment in order to preserve information as much as possible and reduce subjective bias. When experimenting with the LSTM and GRU models, stock price forecasting was carried out with various combinations of sentiment indicators. The results showed the best predictive ability in the GRU model. The inclusion of the first main component in the model, as a generalized sentimental indicator, reduced the prediction error and improved the quality of the model. The study shows that in the Chinese market, taking into account investor sentiment can improve the predictive capabilities of models

Keywords: SENTIMENT ANALYSIS, STOCK PRICE FORECASTING, CHINESE STOCK MARKET, DEEP LEARNING, INVESTOR SENTIMENTS

http://dx.doi.org/10.21515/1990-4665-215-026

Введение.

Прогнозирование цен на акции давно является актуальной темой в финансовом секторе, привлекающей внимание многочисленных ученых и практиков, которые постоянно исследуют различные модели и методы с целью выявить рыночные закономерности и оптимизировать инвестиционные решения. Гипотеза эффективного рынка предполагает, что цены на акции полностью, точно и оперативно отражают всю доступную информацию, но будущие движения следуют модели случайного блуждания, что создает определенные трудности при прогнозировании [5]. Тем не менее, многие исследователи утверждают, что на колебания цен на акции влияют многочисленные факторы, в том числе макроэкономические условия, политические решения, фундаментальные показатели компаний, баланс спроса и предложения, его изменения во времени [1],[9]. Одновременно экономические и политические факторы могут генерировать циклические закономерности, приводящие к неслучайным тенденциям в ценах [2],[23]. Следовательно, прогнозирование движений на фондовом рынке остается сложной и трудной задачей. Высокая размерность данных, учет влияния внутренних и внешних факторов вносят значительный шум, который существенно влияет на цены акций [6].

В связи с динамичными изменениями информационной среды инвесторы все чаще склонны высказывать свое мнение в социальных сетях и просматривать неструктурированную текстовую информацию из таких источников, как новостные сайты, финансовые форумы и объявления компаний. В результате этого процесса накапливается значительный массив данных о тональности, настроениях потенциальных инвесторов, которые оказывают сложное трудноуловимое влияние на цены акций [4].

Традиционные авторегрессионные модели не всегда способны учитывать многомерные, ярко выраженные нелинейные зависимости финансовых временных рядов, полученных из разных источников. Современные методы машинного и глубокого обучения, например, такие как метод опорных векторов (SVM), искусственные нейронные сети, сети LSTM, их гибридные модели, способны эффективно обрабатывать нелинейные связи и многомерные данные и часто применяются для прогнозирования рыночной стоимости акций. Shen и соавт. (2025), [4], разработали гибридную модель прогнозирования динамики цен на акции, основанную на полном ансамблевом эмпирическом модальном разложении с адаптивным шумом (CEEMDAN), сетях LSTM и трансформере. Согласно результатам теста Диболда–Мариано (DM) авторы получили статистически значимое повышение точности прогноза по сравнению с базовыми моделями LSTM, CEEMDAN-LSTM и Transformer [17]. Liu и соавт. (2023),[12], извлекли комментарии инвесторов с платформы Stocktwits, использовали модель FinBERT для формирования эмоциональных оценок и на основе модели SVM предсказали направление движения акций. По сравнению с другими методами предложенный подход повысил точность прогноза на 4–5 %.

С 2008 года анализ тональности (англ. sentiment analysis) стал популярным и интенсивно развивающимся направлением исследований. В 2008–2022 годах количество публикаций, в которых фигурирует концепция анализа тональности, увеличивалось с темпом роста около 34 % в год [16]. Однако оценка эффективности применения анализа тональности для прогнозирования цен акций остаётся неоднозначной в академической среде. В исследовании Nguyen и соавт. (2015), [14], отмечается, что тональность является одним из основных факторов, влияющим на волатильность цен на акции, однако комментарии отдельных инвесторов могут не отражать эффективно движения цен на акции. Розничные инвесторы обычно обладают ограниченными источниками информации и недостаточными

возможностями технического анализа, а значит высказываемые ими мнения могут быть неточными [14]. В то же время Jing и соавт. (2021),[10], показали, что гибридная модель, интегрирующая методы глубокого обучения и анализ тональности, демонстрирует более высокую эффективность и точность прогнозирования цен акций по сравнению с моделями, не учитывающими настроения. В последние годы, по мере прогресса крупных языковых моделей (LLM) и методов глубокого обучения, классификация тональности на базе LLM превосходит традиционную модель FinBERT по качеству прогнозирования цен акций [3].

В настоящей статье построена модель прогнозирования цен акций на основе тональности инвесторов и её проверка на реальных данных Шэньчжэньской фондовой биржи. Характерная особенность рынка акций класса А в Китае заключается в доминировании розничных инвесторов, на долю которых приходится более 80 % торгового оборота, а также в сильном влиянии регуляторной политики [7], что позволяет надеяться, что анализ тональности новостей и тональности инвесторов на данном рынке позволит улучшить прогнозирование.

Обзор литературы. Модель прогнозирования цены акций

Исторически прогнозирование цен на акции основывалось на традиционных моделях временных рядов. В 1970 году Бокс и Дженкинс впервые предложили модель авторегрессионного интегрированного скользящего среднего (ARIMA), которая является одной из наиболее представительных статистических моделей в анализе временных рядов. Данная модель основывается на линейном прогнозировании и использует операции дифференцирования для преобразования нестационарных рядов в стационарные. Затем динамические зависимости временного ряда моделируются с использованием составляющих авторегрессии (AR) и

скользящего среднего (MA). Однако, несмотря на прочную статистическую основу, традиционные модели временных рядов обычно опираются на сильные гипотезы о данных, которые на практике нередко не выполняются [15]. Эти ограничения побуждают исследователей обращаться к методам машинного и глубокого обучения, способным автоматически извлекать многомерные признаки и захватывать сложные нелинейные взаимосвязи.

Прогнозирование цен на акции, на основе методов машинного обучения, заключается в автоматическом обучении закономерностям и моделям, по историческим данным, для предсказания будущих цен. К классическим моделям относятся такие, как методы, основанные на решающих деревьях, нейронные сети с обратным распространением ошибки(BP), метод SVM. Хотя модели машинного обучения могут повысить точность прогнозов и изучать нелинейные зависимости, они сильно подвержены влиянию качества данных. Если качество данных низкое или их объем недостаточен, могут возникнуть проблемы переобучения или недообучения модели [11].

Модели глубокого обучения позволяют одновременно повышать точность прогнозов, снижать проблему переобучения и улучшать обобщающую способность. По сравнению с другими моделями LSTM способна сохранять и использовать зависимости на длительные временные периоды. Nelson и соавт. (2017),[13], опираясь на несколько наборов данных и систему оценочных метрик, сопоставили LSTM, многослойный перцептрон MLP, случайный лес и другие модели. Результаты показали, что сеть LSTM по сравнению с традиционными методами значительно повышает предсказательную способность, особенно при более длительном горизонте прогноза.

Анализ тональности для прогнозирования цен акций

Настроения инвесторов являются важным фактором прогнозирования динамики цен акций, и всё больше исследований учитывают их при

моделировании ценовых тенденций. Souma и соавторы (2019),[19], обучили рекуррентные нейронные сети (RNN) и сети LSTM на новостных данных Thomson Reuters, что повысило точность прогноза цен акций. Wu и соав-торы (2021),[20], предложили метод анализа тональности на основе сверточных нейронных сетей (CNN) для вычисления индекса тональности инвесторов и использовали модель LSTM для прогнозирования цен акций. Результаты показали, что предложенный подход превосходит традиционные методы. Zhen и соавторы (2025),[22], объединили несколько методов отбора признаков и модели глубокого обучения для прогнозирования цен акций на китайском рынке. Гибридная модель LSTM-CNN-Attention продемонстрировала наилучшие показатели.

Данные и методы. Сбор данных

Из базы данных CNRDS, за пятилетний период 2020–2024 г.г., собраны новостные данные и комментарии инвесторов по всем эмитентам акций класса А, которые обращаются на китайском рынке. База данных CNRDS является первой в Китае, где для сбора и аналитической обработки применяются алгоритмы искусственного интеллекта и методы определения тональности текста из области компьютерных наук, что обеспечивает высокую долю верного распознавания положительной, нейтральной и отрицательной тональности внутри выборки до 85%, она охватывает новости подавляющего большинства ведущих средств массовой информации страны. В работе использованы сетевые финансовые новости из модуля CFND и комментарии инвесторов из модуля GUBA, собранные на крупнейшем в Китае форуме инвесторов Eastmoney. Кроме того, из базы данных CSMAR отобраны 30 показателей, включая цену закрытия, цену открытия и ежедневную рыночную капитализацию отдельных акций.

Построение индикаторов тональности

В данном исследовании в качестве источников данных о тональности выбраны интернет-новости и социальные медиа, на основе которых

формируется индекс тональности. Интернет-новости отражают текущие изменения в политике и фундаментальных показателях компаний, включая оценки профессиональных инвестиционных институтов. Этот индекс тональности далее обозначается как Sentiment_news. Sentiment_news формируется по данным, агрегированным по торговым дням. Данные социальных медиа представлены комментариями с форумов фондового рынка, которые по сравнению с традиционными социальными медиа обладают большей профессиональной направленностью и позволяют фиксировать динамику тональности частных инвесторов в режиме реального времени. Этот индекс тональности будем обозначать как Sentiment_guba. Sentiment_guba рассчитывается на основе данных, собранных по календарным дням.

Каждый источник данных о тональности включает положительную, отрицательную и нейтральную тональности. Индекс тональности S_t рассчитывается по следующей формуле [18]:

$$S_t = \frac{pos_t - neg_t}{total_t} \quad (1)$$

где pos_t обозначает количество положительных новостей/комментариев, neg_t обозначает количество отрицательных новостей/комментариев, $total_t$ обозначает общее количество новостей/комментариев, включая нейтральные.

Далее мы строим два сводных индекса тональности и проверяем можно ли улучшить прогноз, включая какой либо из сводных индексов тональности в модель. Первый сводный индекс построим как среднее арифметическое индексов тональности социальных медиа и тональности новостей:

$$S_t^{avg} = (S_t^s + S_t^n)/2, \quad (2)$$

где S_t^{avg} арифметическое среднее индексов тональности (также будем обозначать его как Sentiment), S_t^s – тональность социальных медиа,

S_t^n- тональность новостей. Формула (2) означает равные веса, что носит достаточно субъективный характер. Новостные и социальные медиа различаются по информативности, поэтому агрегирование с равными весами не выглядит достаточно обоснованным. Второй сводный индекс тональности мы строим как первую главную компоненту для стандартизированных рядов индекса тональности новостей и индекса тональности социальных медиа. Формула расчёта PC1:

$$S_t^{PC1} = a\, Z_t^s + b\, Z_t^n \quad (3)$$

где S_t^{PC1}- сводный индекс тональности (первая главная компонента), Z_t^s обозначает стандартизированный ряд индекса тональности социальных медиа, Z_t^n обозначает стандартизированный ряд индекса тональности новостей, a и b обозначают нагрузки первой главной компоненты метода главных компонент. Геометрический смысл первой главной компоненты – направление, вдоль которого достигается максимальная дисперсия из всех возможных, что указывает на наибольшую информативность главной компоненты, которая является линейной комбинацией нормированных исходных индексов.

Методы оценки

В качестве метрик оценки результатов точности моделей выбирались *MSE*, *MAE* и *RMSE*. Формулы расчёта:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (4)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (5)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (6)$$

где y_i обозначает фактическую цену закрытия акции в день i, $\hat{y}_i$ обозначает прогнозную цену закрытия в тот же день, N обозначает объём тестовой выборки. Чем меньше значения этих метрик, тем лучше качество модели и тем выше точность прогноза.

Результаты экспериментов и обсуждение.

Были выбраны акции, по которым в базе данных CNRDS в модулях новостей и инвесторских комментариев имеются полные данные за непрерывный пятилетний период. Таких компаний оказалось две. Первая компания — это девелоперская группа Vanke (код 000002), являющаяся одной из ведущих компаний в сфере недвижимости Китая. Крупнейшим акционером компании является Shenzhen Metro Group. По состоянию на 12 декабря 2025 года совокупная рыночная капитализация составила 59,89 млрд юаней, что соответствует четвертому месту в отрасли. Вторая компания — это государственное предприятие в сфере производства продуктов питания и напитков Luzhou Laojiao (код 000568). Она является одной из репрезентативных публичных компаний китайской алкогольной отрасли и была признана одной из четырех крупнейших известных марок китайского байцзю[1] на первой национальной дегустационной конференции в 1952 году. По состоянию на 12 декабря 2025 г. ее совокупная рыночная капитализация составила 181,02 млрд юаней, что соответствует четвертому месту в отрасли. Обе компании обладают высокой репрезентативностью в своих отраслях, характеризуются высокой ликвидностью акций и активным обсуждением на инвестиционных форумах, что способствует эффективному сопоставлению данных о тональности с рыночным торговым поведением. Следует отметить, что в силу ограниченного объема выборки полученные эмпирические результаты в основном отражают характеристики типичных компаний соответствующих отраслей, а выводы для отрасли в целом носят предварительный характер. Нелинейный характер динамики цен акций не исключает наличия потенциальной линейной связи между ценой и входными признаками [8]. По итогам предварительного анализа из 30 рыночных и корпоративных показателей базы CSMAR были отобраны 5, которые вместе с индикаторами

[1] Байцзю – китайский крепкий дистиллированный напиток, крепостью 40-60%.

тональности сформировали итоговый набор данных.

Таблица 1 — Описание используемых показателей

Название поля	Значение поля
Close_price	Цена закрытия
AdjReturn	Скорректированная доходность
PE	Коэффициент P/E
Turnover	Оборачиваемость (доля оборота)
Liquidity	Ликвидность
Cnvaltrdtl	Совокупный оборот рынка

Все переменные были нормированы по формуле min–max:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (7)$$

где x' обозначает нормализованное значение переменной по методу *min–max*, x обозначает исходное значение переменной, $min(x)$ *и* $max(x)$ обозначают соответственно минимальное и максимальное значения переменной x в выборке.

Для интеграции индикаторов тональности из новостей и социальных сетей и выбора агрегирующего индикатора тональности мы применили метод главных компонент (PCA) и рассмотрели первую главную компоненту (PC1) как второй сводный индекс тональности, т.к. PC1 представляет собой направление с максимальной дисперсией в исходных данных. Сначала мы преобразовали исходные индикаторы тональности S_t^s и S_t^n в числовой формат и заполнили пропущенные значения средним значением по соответствующему столбцу. Затем мы предварительно стандартизировали эти индикаторы, получив Z_t^s и Z_t^n (дисперсии равны 1). Первая главная компонента - PC1, вторая - PC2. С помощью функции PCA библиотеки *scikit-learn* (при параметрах *n_components=2* и *random_state=0*) могут быть получены главные компоненты $S_t^{PC1} = 0.7071 \cdot Z_t^s + 0.7071 \cdot Z_t^n$ и

$S_t^{PC2} = 0.7071 \cdot Z_t^s - 0.7071 \cdot Z_t^n$. Но в нашем случае этот результат и так очевиден, так как у нас всего два индекса тональности. Поскольку, мы предполагаем, что Z_t^s и Z_t^n имеют коэффициент корреляции $\rho > 0$, то собственные числа ковариационной матрицы равны $1 \pm \rho$, собственные вектора $\begin{pmatrix}1\\1\end{pmatrix}$ и $\begin{pmatrix}1\\-1\end{pmatrix}$ и из условия нормировки получаем сводный нормированный индекс тональности $S_t^{PC1} = \frac{1}{\sqrt{2}} Z_t^s + \frac{1}{\sqrt{2}} Z_t^n$, где Z_t^s и Z_t^n определены выше в формуле (3).

Доли объяснённой дисперсии для PC1 и PC2 в выборках двух компаний различаются. Для компании «Vanke» доля объясненной дисперсии 50,13% для PC1 и 49,87% для PC2. Это указывает на то, что в данной выборке два исходных индекса тональности практически не коррелируют друг с другом. Для компании «Luzhou Laojiao» доля объяснённой дисперсии для PC1 составляет 57,3%, для PC2 - 42,7%. Поэтому мы сохраняем PC1 в качестве второго интегрального показателя тональности. С экономической точки зрения PC1, агрегируя изменения тональности, содержащиеся в новостях и комментариях инвесторов, позволяет несколько полнее отразить общее рыночное настроение. Для выбранных показателей и сводных индексов тональности построена корреляционная матрица (на рисунке1, для примера показана матрица корреляций для компании Vanke).

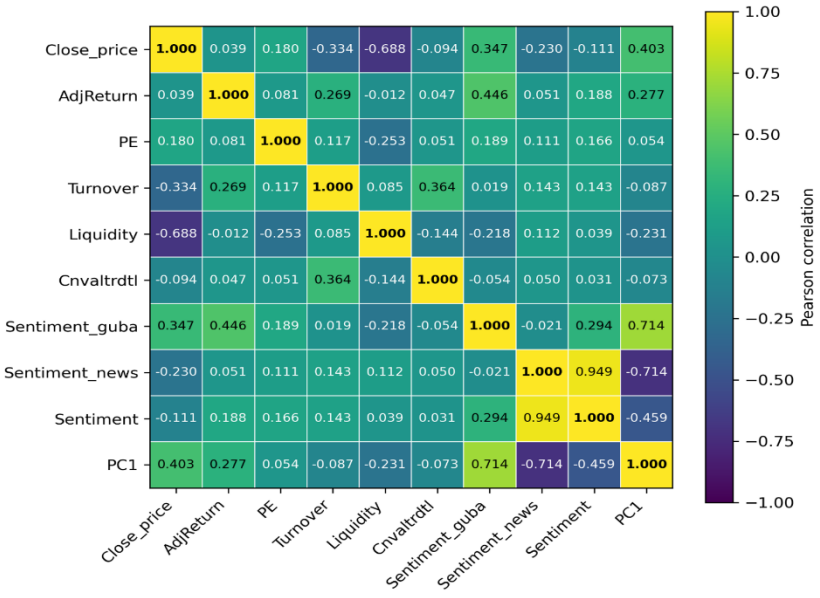

Рис. 1 Матрица корреляции для компании Vanke

Для выбранных показателей и индикаторов тональности можно отметить:

- почти нулевое значение коэффициента корреляции между индексами тональности Sentiment_guba и Sentiment_news, т. е. выбранные индексы тональности практически не коррелированы;

- сильная отрицательная корреляция между ликвидностью и целевой переменной;

- коэффициент корреляции между индикатором тональности в социальных сетях (Sentiment_guba) и ценой закрытия (Close_price) составляет приблизительно +0,347, что указывает на возможное незначительное повышение цены акций при росте данного индикатора;

- коэффициент корреляции между новостным индикатором тональности (Sentiment_news) и ценой закрытия (Close_price) составляет около –0,230, что свидетельствует о тенденции к незначительному снижению цены акций при росте новостного индикатора;

- сводный показатель тональности Sentiment имеет выраженную сильную корреляцию с новостным индексом тональности Sentiment_news, при этом оба индекса слабо влияют на понижение целевой переменной

Close_price;

- наибольшее влияние из индексов тональности на целевую переменную Close_price у второго сводного индекса тональности, выбранного как главная компонента.

Оценка модели

Были использованы две модели глубокого обучения для прогнозирования цен акций: **LSTM** и **GRU**. Сеть LSTM хорошо улавливает долгосрочные зависимости, а GRU, хотя и является упрощённой версией LSTM, лучше подходит для задач с небольшим объёмом данных. Поскольку за пятилетний период объём дневных признаков по акциям невелик, мы выбрали именно эти две модели.

Для всех групп экспериментов, представленных в таблице 2, фиксировались архитектура сети и обучающие гиперпараметры. Отличие экспериментов заключается в различных комбинациях признаков тональности ($S_t^s, S_t^n, S_t^{avg}, S_t^{PC1}$) на входе. Входные выборки формировались методом скользящего окна: по последовательности признаков за предыдущие 50 торговых дней прогнозировалась цена закрытия следующего торгового дня. Для модели LSTM данные разделялись по времени, 80% наблюдений использовались для обучения и 20% для тестирования. Архитектура сети: 2 скрытых рекуррентных слоя по 150 нейронов в каждом слое, dropout составляет 0,2, в качестве функции активации используются стандартные для PyTorch активации LSTM, то есть функции sigmoid и tanh. В качестве функции потерь использовалась *MSE*, оптимизатором выбран Adam, скорость обучения равна 0,0005, размер мини пакета составляет 32, число эпох равно 100.

Архитектура модели GRU соответствует настройкам LSTM, число скрытых слоев равно 2, в каждом слое по 150 нейронов, dropout равен 0,2, оптимизатор Adam, скорость обучения 0,0005, размер мини-пакета 32, число эпох 40. Аналогично, в механизме ворот используются функции

sigmoid и tanh в качестве функций активации. При временном разбиении 80% - обучающая выборка, 20% - тестовая, наблюдалось снижение качества прогноза на тестовой выборке и увеличение размера ошибок. При разбиении 70% на 30% модель GRU ошибки выглядели более стабильными. Поэтому для модели GRU использовано разбиение 70% на 30%. Все переменные нормировались методом min–max, причем параметры нормировки оценивались только на обучающей выборке и затем применялись к тестовой. Случайное зерно фиксировалось как 3407.

Следует отметить, что основная цель работы состоит в оценке дополнительного вклада различных комбинаций признаков тональности в улучшение качества прогнозирования, поэтому единые архитектура и обучающие гиперпараметры используются как контрольное условие, чтобы обеспечить сопоставимость результатов внутри одной модели. Ниже представлены результаты оценки моделей при интеграции различных признаков настроений по метрикам *MSE*, *MAE* и *RMSE*.

Таблица 2 — Результаты прогнозирования моделей LSTM и GRU группы **Vanke**

Модели		O	$O + S_t^s$	$O + S_t^n$	$O + S_t^{avg}$	$O + S_t^{PC1}$
LSTM	*MSE*	0,000463	0,000435	0,000481	0,000489	0,000396
	MAE	0,017387	0,016066	0,017810	0,018113	0,015202
	RMSE	0,021522	0,020855	0,021921	0,022113	0,019908
GRU	*MSE*	0,000476	0,000357	0,000466	0,000453	0,000348
	MAE	0,015375	0,014325	0,016769	0,016426	0,014227
	RMSE	0,021808	0,018889	0,021586	0,021276	0,018656

Обозначения столбцов таблицы 2: O – модель без учета тональности, $O + S_t^s$ – модель с учетом тональности социальных медиа, $O + S_t^n$ – модель с учетом тональности новостных медиа, $O + S_t^{avg}$ – модель с учетом первого сводного индекса тональности (среднее индексов новостных и социальных медиа), $O + S_t^{PC1}$ – модель с учетом второго сводного индекса тональности (первая главная компонента). Из таблицы 2 видно, что показатели производительности моделей LSTM и GRU демонстрируют, что модель, интегрирующая первый главный компонент PCA, показывает наилучшую

точность модели. При этом модель GRU в целом проявляет себя несколько лучше, с минимальным значением MSE 0,000348 ($O + S_t^{PC1}$) и максимальным 0,000476 (O), что демонстрирует более сильную способность GRU к обработке малых партий временных последовательностей, особенно при интеграции признаков тональности. В целом, модели с интегрированными индикаторами тональности показывают улучшение в прогнозировании цен акций по сравнению с моделями без индикаторов тональности.

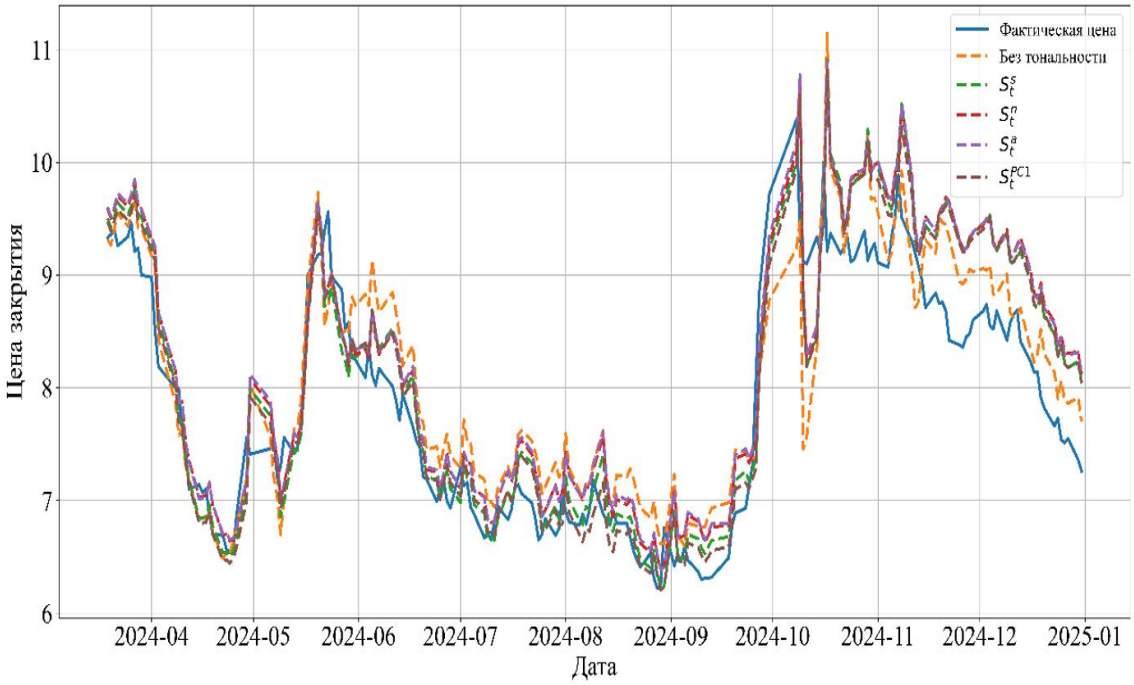

Рис. 2 Сравнение прогнозов LSTM с фактическими значениями при различных индикаторах тональности (Vanke)

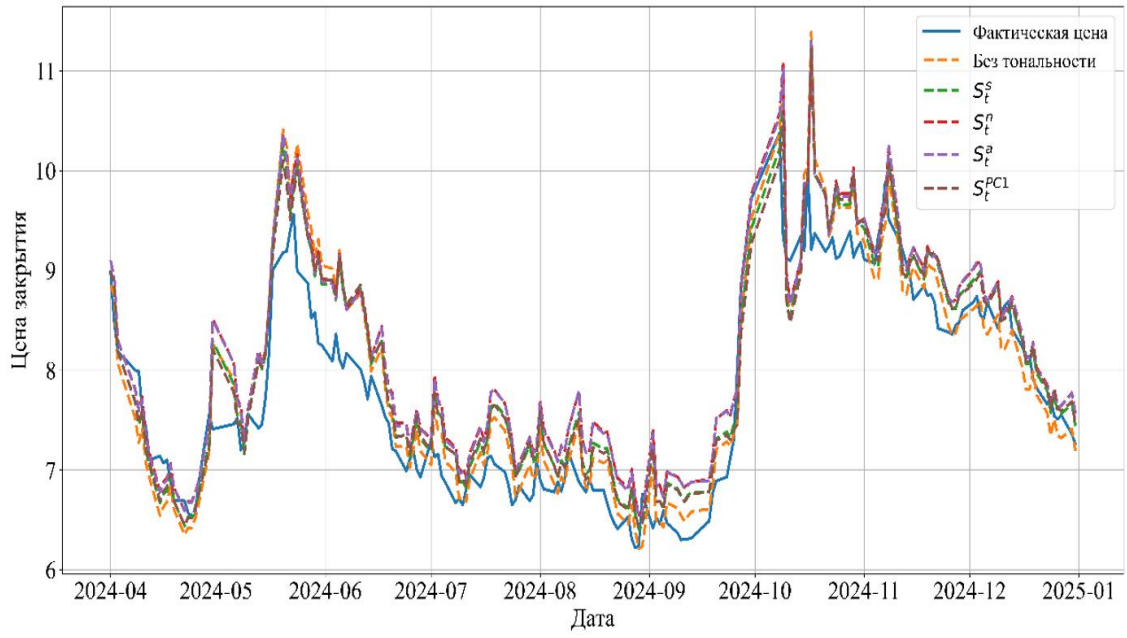

Рис. 3 Сравнение прогнозов GRU с фактическими значениями при различных индикаторах тональности (Vanke)

На рис. 2 и рис. 3 показано сравнение прогнозируемых моделей LSTM и GRU с фактическими ценами. В период с сентября 2023 года по январь 2025 года прогнозируемые линии под всеми комбинациями индикаторов тональности тесно следуют тенденции фактических цен, но в середине 2024 года наблюдаются сильные колебания. Общая тенденция показывает снижение цены примерно с 9 юаней до 7 юаней, прогнозируемые кривые после интеграции индикаторов тональности становятся более гладкими, уменьшая переобучение. В то же время прогнозы модели GRU с вторым сводным индексом тональности (главная компонента) ближе к фактическим значениям, особенно в середине 2024 года. Теперь посмотрим результаты моделирования для компании Luzhou Laojiao (LZLJ).

Таблица 3 — Результаты прогнозирования моделей LSTM и GRU компании Luzhou Laojiao (LZLJ)

Модели		O	$O + S_t^s$	$O + S_t^n$	$O + S_t^{avg}$	$O + S_t^{PC1}$
LSTM	*MSE*	0,000341	0,000340	0,000314	0,000320	0,000266
	MAE	0,013457	0,014651	0,013395	0,013543	0,011604
	RMSE	0,018466	0,018431	0,017728	0,017888	0,016296
GRU	*MSE*	0,000312	0,000274	0,000285	0,000284	0,000280
	MAE	0,014133	0,012413	0,012626	0,012604	0,012518

	RMSE	0,017673	0,016545	0,016867	0,016850	0,016728

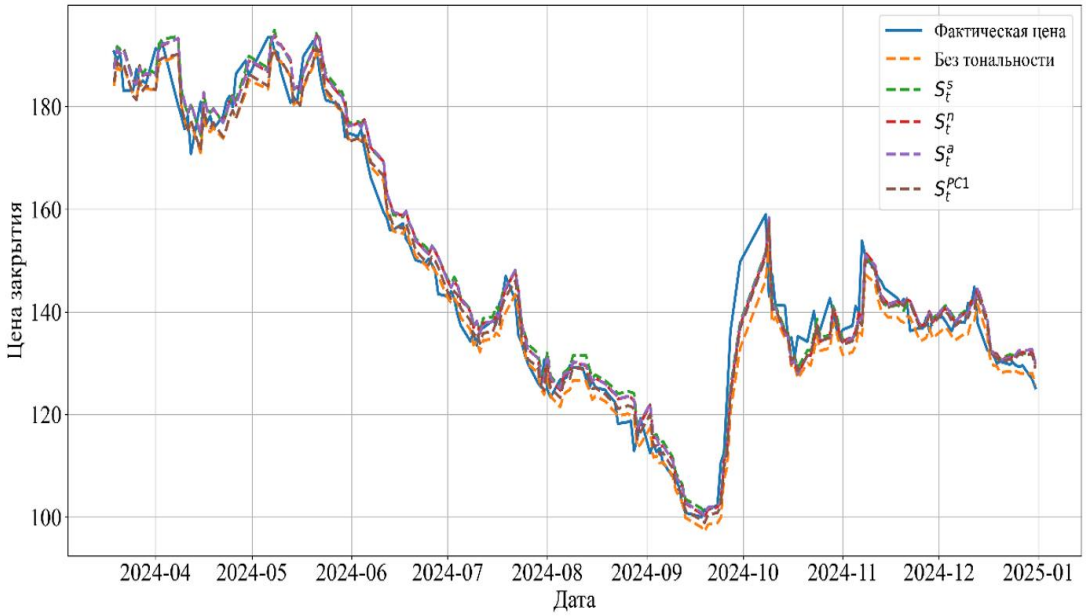

Рис. 4 Сравнение прогнозов LSTM с фактическими значениями при различных индикаторах тональности (LZLJ)

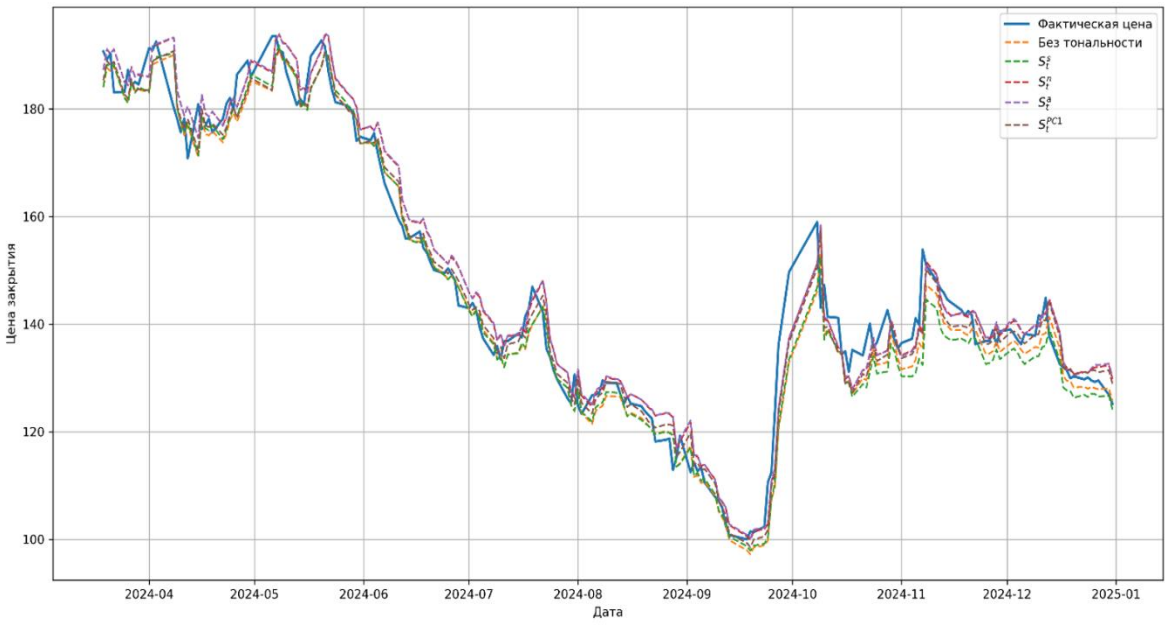

Рис. 5 Сравнение прогнозов GRU с фактическими значениями при различных индикаторах тональности (LZLJ)

Результаты экспериментов подтвердили наши выводы на датасете

Vanke. Из таблицы 3 видно, что на данных компании Luzhou Laojiao (LZLJ) лучшую точность показывает модель с добавлением первого компонента PC1, модель без добавления тональности имеет наибольшую ошибку. В целом, из рисунков 2-5, таблиц 2,3 видно, что модели с интегрированными тональностями повышают точность прогнозирования.

Заключение

В данной работе из социальных медиа и новостей в интернете извлечены две категории индикаторов тональности. Первая отражает тональность розничных инвесторов, вторая показывает тональность политики и профессиональных институтов. Из этих индикаторов построены два сводных индекса тональности: с помощью арифметического среднего и метода главных компонентов PCA. Эффективность индикаторов проверена в моделях LSTM и GRU. В исследовании использованы две акции с полной историей данных. Эксперименты показали, что модель с индикатором тональности по первой главной компоненте PCA дает наилучшие результаты. Полученные результаты указывают, что на китайском фондовом рынке коллективная тональность розничных инвесторов, формирующих наибольший объем оборота, а также политическая тональность могут выполнять роль ориентира, а добавление индикаторов тональности может повышать точность прогнозов.

Настоящее исследование предоставляет практическую основу для анализа влияния тональности инвесторов на цены акций в условиях китайского рынка, однако имеет ограничения. Объем данных невелик, поэтому в дальнейшем целесообразно расширить объем данных путем сбора статей и новостных материалов о других компаниях. Перспективным направлением исследования может быть увеличение индикаторов тональности из разных новостных и медийных источников и рассмотрение большого количества индикаторов тональности. В том случае можно ожидать более интересных комбинаций индикаторов тональностей,

участвующих в формировании главных компонент.

Текущие данные в основном текстовые, в будущем возможно включение изображений и видео с переходом к мультимодальному подходу.

В работе рассматривались только две компании, представленные на Шэньчженьской бирже. Следующим этапом исследований может быть расширение числа компаний для анализа тональностей, чтобы можно было распространить технику учета тональности в прогнозах на активы, торгующиеся на китайских площадках (Шанхайская (SSE), Шэньчжэньская (SZSE), Пекинская (BSE), Гонконгская (HKEX) биржи). В условиях глобализации рынки разных стран зависят от внешней информации, поэтому еще одним перспективным направлением является распространение исследования на несколько рынков.

Исследование выполнено при поддержке Лаборатории азиатских экономических исследований Санкт-Петербургского государственного университета и Китайского государственного стипендиального совета.

ЛИТЕРАТУРА

1. Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The quarterly journal of economics*, *131*(4), 1593-1636.
2. Belo, F., Gala, V. D., & Li, J. (2013). Government spending, political cycles, and the cross section of stock returns. *Journal of financial economics*, *107*(2), 305-324.
3. Chen, Q., & Kawashima, H. (2024, December). Stock price prediction using llm-based sentiment analysis. In *2024 IEEE International Conference on Big Data (BigData)* (pp. 4846-4853). IEEE.
4. Chen, X., Xie, H., Li, Z., Zhang, H., Tao, X., & Wang, F. L. (2025). Sentiment analysis for stock market research: A bibliometric study. *Natural Language Processing Journal*, *10*, 100125.
5. Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, *25*(2), 383-417.
6. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223-2273.
7. He, F., Yan, Y., Hao, J., & Wu, J. G. (2022). Retail investor attention and corporate green innovation: Evidence from China. *Energy Economics*, *115*, 106308.
8. Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, *124*, 226-251.
9. Hirshleifer, D. (2015). Behavioral finance. *Annual Review of Financial*

Economics, *7*(1), 133-159.

10. Jing, N., Wu, Z., & Wang, H. (2021). A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Systems with Applications*, *178*, 115019.
11. Kim, H. Y., & Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Systems with Applications*, *103*, 25-37.
12. 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*, e1403.
13. Nelson, D. M., Pereira, A. C., & De Oliveira, R. A. (2017, May). Stock market's price movement prediction with LSTM neural networks. In *2017 International joint conference on neural networks (IJCNN)* (pp. 1419-1426). Ieee.
14. Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), 9603-9611.
15. Petrică, A. C., Stancu, S., & Tindeche, A. (2016). Limitation of ARIMA models in financial and monetary economics. *Theoretical & Applied Economics*, *23*(4).
16. Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P. M. (2023). A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, *223*, 119862.
17. Shen, Y., Dai, J., Wang, M., & Arce, G. R. (2025). Stock price trend forecasting based on multi-channel complementary network with CEEMDAN decomposition and Transformer residual prediction. *Expert Systems with Applications*, 130028.
18. Soo, Cindy, Quantifying Animal Spirits: News Media and Sentiment in the Housing Market (October 2015). Ross School of Business Paper No. 1200.
19. Souma, W., Vodenska, I. & Aoyama, H. Enhanced news sentiment analysis using deep learning methods. J Comput Soc Sc 2, 33–46 (2019). https://doi.org/10.1007/s42001-019-00035-x
20. Wu, S., Liu, Y., Zou, Z., & Weng, T. H. (2022). S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis. *Connection Science*, *34*(1), 44-62.
21. Yun, M. K. K. (2025). Effect of exogenous market sentiment indicators in stock price direction prediction. *Expert Systems with Applications*, *281*, 127696.
22. Zhen, K., Xie, D., & Hu, X. (2025). A multi-feature selection fused with investor sentiment for stock price prediction. *Expert Systems with Applications*, *278*, 127381.
23. Zhu, Y., & Zhu, X. (2014). European business cycles and stock return predictability. *Finance research letters*, *11*(4), 446-453.

REFERENCES

1. Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The quarterly journal of economics*, *131*(4), 1593-1636.
2. Belo, F., Gala, V. D., & Li, J. (2013). Government spending, political cycles, and the cross section of stock returns. *Journal of financial economics*, *107*(2), 305-324.
3. Chen, Q., & Kawashima, H. (2024, December). Stock price prediction using llm-based sentiment analysis. In *2024 IEEE International Conference on Big Data (BigData)* (pp. 4846-4853). IEEE.
4. Chen, X., Xie, H., Li, Z., Zhang, H., Tao, X., & Wang, F. L. (2025). Sentiment analysis for stock market research: A bibliometric study. *Natural Language Processing Journal*, *10*, 100125.
5. Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, *25*(2), 383-417.

6. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223-2273.
7. He, F., Yan, Y., Hao, J., & Wu, J. G. (2022). Retail investor attention and corporate green innovation: Evidence from China. *Energy Economics*, *115*, 106308.
8. Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, *124*, 226-251.
9. Hirshleifer, D. (2015). Behavioral finance. *Annual Review of Financial Economics*, *7*(1), 133-159.
10. Jing, N., Wu, Z., & Wang, H. (2021). A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Systems with Applications*, *178*, 115019.
11. Kim, H. Y., & Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Systems with Applications*, *103*, 25-37.
12. 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*, e1403.
13. Nelson, D. M., Pereira, A. C., & De Oliveira, R. A. (2017, May). Stock market's price movement prediction with LSTM neural networks. In *2017 International joint conference on neural networks (IJCNN)* (pp. 1419-1426). Ieee.
14. Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), 9603-9611.
15. Petrică, A. C., Stancu, S., & Tindeche, A. (2016). Limitation of ARIMA models in financial and monetary economics. *Theoretical & Applied Economics*, *23*(4).
16. Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P. M. (2023). A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, *223*, 119862.
17. Shen, Y., Dai, J., Wang, M., & Arce, G. R. (2025). Stock price trend forecasting based on multi-channel complementary network with CEEMDAN decomposition and Transformer residual prediction. *Expert Systems with Applications*, 130028.
18. Soo, Cindy, Quantifying Animal Spirits: News Media and Sentiment in the Housing Market (October 2015). Ross School of Business Paper No. 1200.
19. Souma, W., Vodenska, I. & Aoyama, H. Enhanced news sentiment analysis using deep learning methods. J Comput Soc Sc 2, 33–46 (2019). https://doi.org/10.1007/s42001-019-00035-x
20. Wu, S., Liu, Y., Zou, Z., & Weng, T. H. (2022). S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis. *Connection Science*, *34*(1), 44-62.
21. Yun, M. K. K. (2025). Effect of exogenous market sentiment indicators in stock price direction prediction. *Expert Systems with Applications*, *281*, 127696.
22. Zhen, K., Xie, D., & Hu, X. (2025). A multi-feature selection fused with investor sentiment for stock price prediction. *Expert Systems with Applications*, *278*, 127381.
23. Zhu, Y., & Zhu, X. (2014). European business cycles and stock return predictability. *Finance research letters*, *11*(4), 446-453.