EN RU PROGRAM SYSTEMS: THEORY AND APPLICATIONS ISSN 2079-3316
Research Article ARTIFICIAL INTELLIGENCE, INTELLIGENCE SYSTEMS, NEURAL NETWORKS

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

Multimodal Stock Price Prediction: A Case Study of the Russian Securities Market

Kasymkhan Usufovich **Khubiyev**[1], Mikhail Evgenyevich **Semenov**[2]
[1, 2] Sirius University of Science and Technology, Sirius, Russia
[1]✉*kasymkhankhubievnis@gmail.com*

Abstract. Classical asset price forecasting methods primarily rely on numerical data, such as price time series, trading volumes, limit order book data, and technical analysis indicators. However, the news flow plays a significant role in price formation, making the development of multimodal approaches that combine textual and numerical data for improved prediction accuracy highly relevant.

This paper addresses the problem of forecasting financial asset prices using the multimodal approach that combines candlestick time series and textual news flow data. A unique dataset was collected for the study, which includes time series for 176 Russian stocks traded on the Moscow Exchange and 79, 555 financial news articles in Russian.

For processing textual data, pre-trained models RuBERT and Vikhr-Qwen2.5-0.5b-Instruct (a large language model) were used, while time series and vectorized text data were processed using an LSTM recurrent neural network. The experiments compared models based on a single modality (time series only) and two modalities, as well as various methods for aggregating text vector representations.

Prediction quality was estimated using two key metrics: Accuracy (direction of price movement prediction: up or down) and Mean Absolute Percentage Error (MAPE), which measures the deviation of the predicted price from the true price. The experiments showed that incorporating textual modality reduced the MAPE value by 55%.

The resulting multimodal dataset holds value for the further adaptation of language models in the financial sector. Future research directions include optimizing textual modality parameters, such as the time window, sentiment, and chronological order of news messages. (*Linked article texts in English and in Russian*).

Key words and phrases: Multimodal Forecasting, Quantitative Fiance, Machine Learning

2020 *Mathematics Subject Classification*: 68T30; 68T50, 91884

Acknowledgments: This work was supported by the grant of the state program "Scientific and technological development of the 'Sirius' Federal Territory" (Agreement No. 18-03 date 10.09.2024)

For citation: Kasymkhan U. Khubiyev, Mikhail E. Semenov. *Multimodal Stock Price Prediction: A Case Study of the Russian Securities Market*. Program Systems: Theory and Applications, 2025, **16**:1(64), pp. 83–130. (*In English, in Russian*). https://psta.psiras.ru/read/psta2025_1_83-130.pdf

Introduction

Building a price forecast for an asset is a crucial task for financial market participants, as it enables strategic planning, optimal investment portfolio management, and risk assessment. Numerous attempts have been made to apply machine learning methods to construct such forecasts [1–3].

With the growing popularity of deep learning models, researchers have shifted their focus toward the application of neural networks. At the same time, the problem of accurately accounting for the news flow as a key factor influencing market behavior is being reconsidered with the rapid development of generative artificial intelligence models and large language models (LLMs) such as ChatGPT, FinGPT, GigaChat, LLama, and others. In financial economics, LLMs are still rarely used, and their full potential remains untapped.

Researchers are exploring the use of natural language processing models to enhance the accuracy of asset price forecasts and investment portfolio management strategies.

The study [4] describes the use of sentiment analysis of news as an additional parameter. The authors employed the FinBert model, trained on financial data, to assess the sentiment of news articles as positive, negative, or neutral. The study utilized time series data from candlestick charts of the U. S. stock market index, Standard & Poor's 500 (S&P 500). A machine learning model — random forest —was used for price prediction. The study concluded that incorporating sentiment analysis of news flow improves prediction accuracy.

In the study [5], the authors aimed to develop a multimodal artificial intelligence model capable of providing well-founded and accurate forecasts for time series data. They implemented a model that generates predictions of an asset's monthly or weekly returns, accompanied by a textual explanation from a language model based on the user's input query.

The study [6] proposed an approach for fine-tuning instructions to interpret numerical values and contextualize financial data.

Kulikova et al. [7] examined the effect of classifying news into thematic groups. The authors demonstrated that, in most cases, it is advisable to use a single thematic group of news for the deep learning models considered (Temporal Convolutional Network, D-Linear, Transformer, and Temporal Fusion Transformer). They also determined the probabilities of forecast improvement for the 20 thematic groups analyzed.

In all the aforementioned studies, the models were implemented using a multimodal approach for the U. S. stock market, with English as the modality language. Notably, the news flow was not integrated directly into the predictor's input vector but rather through a preprocessing block in the form of an additional parameter, such as sentiment analysis, news frequency related to the asset, or news classification, etc.

The objective of the current study is to demonstrate the advantages of a new multimodal method over predictions based solely on numerical data and to present a Russian-language financial news dataset.

To achieve this objective, we formulated the following key tasks:

(*1*) Construct a multimodal dataset consisting of time series data and news articles.
(*2*) Develop a predictive model capable of utilizing one or two modalities.
(*3*) Train the predictive model and analyze the values of accuracy functions and metrics, specifically Accuracy and MAPE.

In this study, we propose a new multimodal approach for integrating news flow into time series numerical data. The text of the news articles is converted into a vector representation and fed into the model alongside the time series vector.

Our hypothesis is that the multimodal approach will enable predictive models to extract semantic information from the text, thereby improving the accuracy of asset price forecasts.

1. Data Collection and Structuring

Multimodality implies the use of more than one data modality, which affects both the data structure and the logic of predictive model development. We utilize two types of modalities:

numerical — time series of stock prices,
textual — news streams.

To train the predictive model and analyze its performance, we collected an original dataset.

The time series, represented as candlestick data with open, close, high, and low prices, were obtained through the Algopack API of the Moscow Exchange (MOEX). For the numerical experiment, we selected stock time series data spanning from July 7, 2022, to August 30, 2024, covering 176

TABLE 1. Statistical features of the dataset after tokenization, RuBert

Source	Mean	Std	Min	Max	Q25	Q50	Q75
RDV	134	88	8	512	65	123	187
Finam	221	135	18	512	116	178	284
BCS Express	20	10	4	82	13	17	26
BCS Technical Analysis	502	37	29	512	512	512	512
RBC	43	7	16	75	39	44	48
SmartLab	21	8	5	82	15	19	25

companies. During this period, the Russian stock market experienced phases of rapid growth and decline, with the IMOEX index rising from 2,213.81 to 2,650.32 points ($+19,72\%$).

We collected $79,555$ news articles from various sources, including the online publication "RBC" (1,823 articles), "BCS Express" (11,331), and "BCS Technical Analysis" (9,670), the investment company website "Finam" (20,647), the trader community website «SmartLab.ru» (30,857), as well as the Telegram channel "RDV" (5,227).

Several factors justify the selection of these sources. First, they provide news coverage for the required time period. Second, the institutional differences between sources, along with variations in writing style and levels of expertise, contribute to a more objective representation of events related to the analyzed time series.

News messages were tokenized using two models: RuBERT [8] and Vikhr-Qwen2.5-0.5b-Instruct [9] (further as Qwen). In the context of tokenized text, a word refers to a token — an element of the vector space represented as an index in the tokenizer's vocabulary.

Descriptive statistics of the dataset (in tokens), including mean, standard deviation, minimum, maximum word count, and quartiles, are presented in Tables 1 and 2. It is important to note that tokenization can increase the word count in a text, for example, by splitting words into smaller components.

Table 3 provides examples of how a phrase changes after tokenization. For instance, the word «открывает» is split into three subcomponents: «от», «##к», and «##рывает», where the "##" prefix indicates that the token is a continuation of the previous token.

TABLE 2. Statistical features of the dataset after tokenization, Qwen

Source	Mean	Std	Min	Max	Q25	Q50	Q75
RDV	215	157	3	1324	92	187	304
Finam	453	405	35	5732	211	319	501
BCS Express	36	19	5	163	23	32	47
BCS Technical Analysis	1493	310	40	2221	1448	1545	1665
RBC	75	12	28	105	68	77	83
SmartLab	33	12	7	120	25	31	39

TABLE 3. Original and tokenized texts examples

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

News articles characteristics On the "BCS Technical Analysis" platform, news articles tend to be lengthy, which imposes limitations on tokenizers. Specifically, as shown in Table 1 and Table 2, the RuBERT model truncates the tokenized vector for longer texts. Additionally, the average length of tokenized text using the Qwen model exceeds that of RuBERT, indicating that Qwen has a broader vocabulary and a stronger text decomposition capability.

Furthermore, we collected data on 176 companies, forming a dataset consisting of tuples in the format:

(ticker, company name, company activity description).

Such data are essential in our case for:

(*a*) extracting keywords from company descriptions,

(*b*) improving the language model's ability to link events described in news articles to specific companies and assess the impact of news on price dynamics.

TABLE 4. Examples of news articles (header snippet) and assigned tags

Source	Article fragment (heading)	Tags
RDV	Сегежа (SGZH): таргет 16.2 руб., апсайд +102...	SGZH
RDV	Артген биотех (ABIO) завершил доклинические...	аналитика, ABIO
Finam	Индекс МосБиржи восстанавливает позиции и приб...	ФосАгро, ВСМПО-АВСМ, CNYRUB
Finam	«Ашинский метзавод» назвал АО "Урал-ВК" своим ...	АшинскийМЗ
BCS Express	«Восходящее окно»: в каких бумагах замечен это...	Селигдар SELG, ЕвроТранс EUTR
BCS Express	«Сила Сибири» выйдет на максимальную мощность...	Газпром GAZP
BCS Technical Analysis	Мечел. Что ждать от бумаг на следующей неделе	Мечел
BCS Technical Analysis	На предыдущей торговой сессии акции Норникеля ...	ГМК Норникель

The dataset of news articles includes the following parameters: publication date, source, title, article body, and tags (keywords). For sources such as “RDV” and “SmartLab”, article titles are absent, and the corresponding fields are filled with a label: *no title*.

In our case, tags may include the full or abbreviated company name along with the corresponding ticker, the name of the market sector, and similar information. Tags in news articles were assigned by the article authors.

For the “RDV” source, tags were marked by authors in the form of hashtags (e. g. #цифры, #аналитика). In “BCS Express” and “BCS Technical Analysis”, tags were specified in dedicated fields at the beginning or end of the news article (e. g. PhoseAgro, Russian market) and were extracted from the *HTML* code of the page using the corresponding *HTML* tags. When tags were absent (“RBC”, “SmartLab”), the parameter in the dataset remained empty.

Table 4 provides examples of news articles (headline fragments) along with their assigned tags.

2. Methods

To validate our hypothesis regarding the advantages of the multimodal approach, we have planned a series of experiments.

The first series of experiments focused on predicting prices using only numerical time series of candlestick characteristics (close, open, high, and low prices). The quality metrics obtained from this experiment serve as baseline values against which improvements in price prediction accuracy using the proposed multimodal approach will be evaluated.

The second series of experiments aims to generate predictions and compute accuracy metrics (Accuracy, MAPE) using the multimodal approach while exploring different aggregation methods (Sum, Mean) for the vectorized news stream.

2.1. The Single-Modality Approach

We first conducted a series of experiments on asset price prediction using only time series data. For this, we applied classical machine learning models to the daily price values (close, open, high, low), including linear regression (LinReg), k-nearest neighbors (KNN), decision tree (DT), random forest (RF), and the boosting algorithm XGBoost (XGB). Among deep learning models, we utilized a long short-term memory recurrent neural network (LSTM).

Conceptually, the experiment consists of two tasks:

(a) predicting the price movement direction (increase or decrease), which is a binary classification task;

(b) predicting the actual price, which is a regression task.

At this stage of the experiment, 176 companies were grouped into 23 industry sectors. We randomly selected 9 economic sectors and, within each sector, randomly chose two companies. Table 5 lists the selected sectors and companies (tickers) that participated in the computational experiment.

Table 7 provides statistical data on the closing price time series of the selected assets. Table 6 shows the distribution of news by companies after filtering. The correlation heat map of the closing price time series is shown in Figure 1. An interesting feature of the examined period is that the market underwent two phase shifts — from a general price decline to growth and back again — as indicated by the vertical lines in Figure 2.

TABLE 5. Economic sectors and companies (tickers) included into the dataset

Sector	Company (ticker)
Metal and Mining	Mechel (MLTR), TMK-Group (TRMK)
Oil and Gas	Surgutneftegas (SNGS), Gaspromneft (SIBN)
Consumer sector	Magnit (MGNT), Lenta (LENT)
Construction	PIK (PIKK), Samolet (SMLT)
Telecommunications	MTS (MTSS), Rostelecom (RTKMP)
Transport	AEROFLOT (AFLT), Sovcomflot (FLOT)
Finance	Bank Saint-Petersburg (BSPB), SFI (SFIN)
Chemical Industry	Phosagro (PHOR), Kazanorgsintez (KZOSP)
Power Engineering	Rushydro (HYDR), Rosseti Center (MRKC)

TABLE 6. Descriptive characteristics for company shares

Company (ticker)	Number of news items
Mechel (MLTR)	4258
Trubnaya Metallurgical Company (TRMK)	11739
Surgutneftegaz (SNGS)	12674
Gazpromneft (SIBN)	11421
Magnit (MGNT)	1236
Lenta (LENT)	311
PIK (PIKK)	897
Samolet (SMLT)	3392
MTS (MTSS)	1101
Rostelecom (RTKMP)	628
Aeroflot (AFLT)	1429
Sovcomflot (FLOT)	14476
Saint Petersburg Exchange (BSPB)	14278
SFAI (SFIN)	1647
PhosAgro (PHOR)	2773
Kazanorgsintez (KZOSP)	168
RusHydro (HYDR)	1921
MRSK Center (MRKC)	1576

TABLE 7. Descriptive characteristics for company shares

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

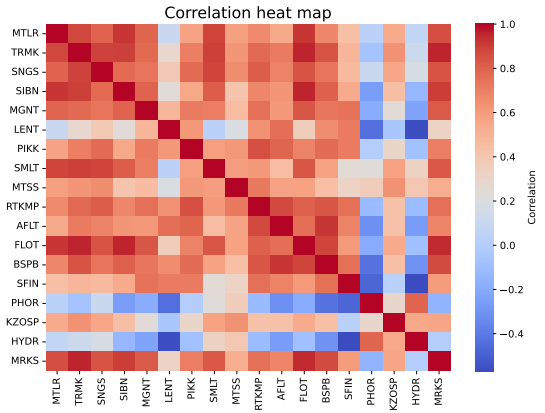

FIGURE 1. The correlations heatmap for 18 assets (close price)

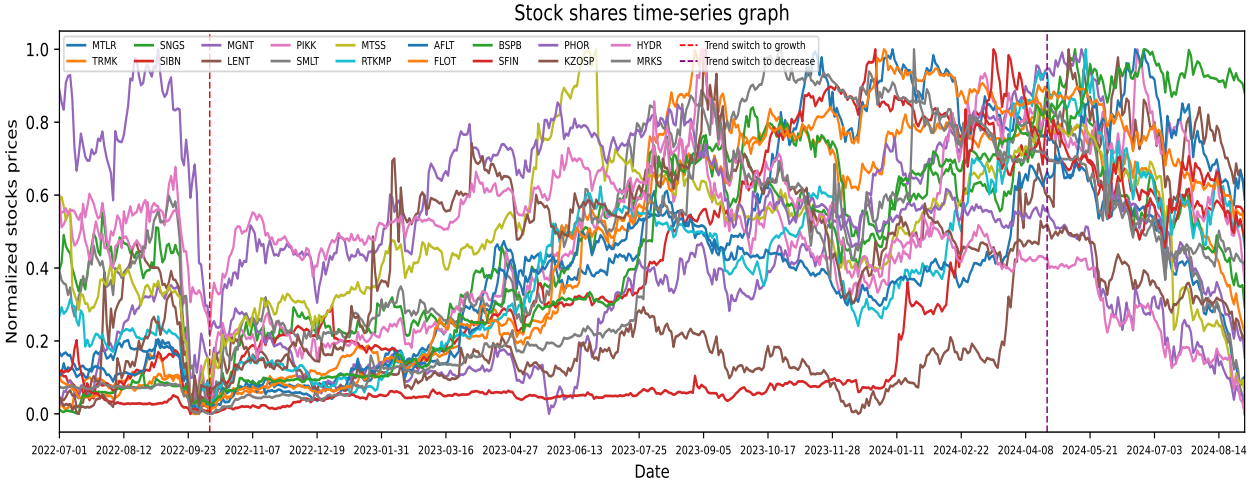

FIGURE 2. Normalized close prices of assets. Market phase transition dates denoted by vertical dashed lines

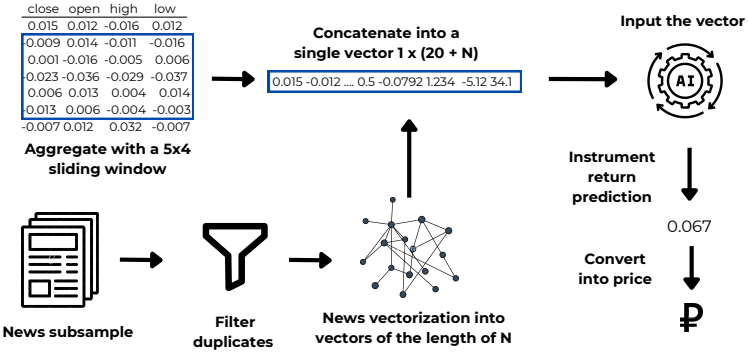

FIGURE 3. Pipeline for a single and dual modalities models

To evaluate prediction quality in the classification task, we used the *Accuracy* metric, while for regression, we employed *MAPE* (Mean Absolute Percentage Error). The choice of these metrics is justified by the nature of the tasks. In classification, the model must accurately predict the price movement direction either an increase (denoted by «+») or a decrease (denoted by «−»). The *MAPE* metric is best suited for assessing regression quality within the financial domain: it represents the average deviation from the asset's actual price in percentage terms, making it easily interpretable in monetary value.

Figure 3 illustrates the model development process for utilizing one and two modalities.

As the input parameter, the model received a return vector of the asset, calculated based on the closing price (close) over the previous five trading sessions:

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

The model's output was a prediction for the next trading session.

To assess the accuracy of predicting the price movement direction, the predicted class was determined by the sign (±) of the forecasted return

value, as the return of an asset represents the relative rate of change. Thus, a positive return indicates a price increase, while a negative return signifies a decline. To evaluate the quality of the asset price forecast, the predicted return vector was converted into price (in Russian rubles):

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

The pointwise predicted price vector, obtained through transformation, was compared to the historical price vector of assets using the *MAPE* metric.

The choice of return (rather than price) as the target variable for the predictive model is justified by the fact that when prices exceed historical highs (or fall below historical lows) during market growth (or decline), the applicability of traditional methods becomes limited.

Based on this reasoning, candlestick characteristics (close, open, high, and low prices) were considered in the form of *relative price changes*, calculated using a formula similar to (1).

Next, a rolling window of five trading days was applied to the relative price changes to form a vector-row, which was then fed into the predictive model. As a result, the model receives a vector of 20 parameters as input and predicts a single output value — the return of the instrument at the end of the next trading session.

2.2. The Dual-Modality Approach

For the experiment involving news flow, we selected news articles relevant to the analyzed assets based on keyword matching (Table 5). The keywords were chosen as the top 30 words extracted using the TF-IDF method. This method determines the importance of words in a text by considering their frequency of occurrence and uniqueness across the entire corpus. An example of keywords extracted using TF-IDF is presented in Table 8.

After obtaining the list of keywords using the TF-IDF method, we further expanded it with the help of the ChatGPT-4o model. This allowed us to increase keyword variability through permutations, letter substitutions, and modifications of word endings (Table 9). The selected news articles for each company (ticker) were converted into vectors and filtered to remove duplicates.

TABLE 8. Keywords by companies extracted from their descriptions

Ticker	Keywords
MTLR	mechel, mining, ore, raw materials, energy, ferroalloys, coal
SNGS	gas, geological exploration, oil, Surgutneftegas, petroleum products, electricity, drilling
SMLT	rent, development, developer, real estate, construction, Moscow region, residential areas
MTSS	subscriber, automation, internet, mobile communications, provider, communications
BSPB	bank, deposit, dividends, financial services, kaliningrad, spbank, saint-petersburg

TABLE 9. Complementary generated keywords

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

Figure 4 presents a distribution chart of the news articles for the companies after filtration.

As a vectorizer for the Russian language news stream, we employed two models: RuBERT [8] and Qwen [9].

While working with the news stream, we encountered two main challenges. The first challenge is the problem of news rewriting, which necessitates filtering out duplicate articles. To ensure that our model accounts for each news article only once, it is essential to implement a duplicate identification algorithm.

The second challenge is to determinate an asset on which is affected the news article. This problem can be framed as a classification task, where

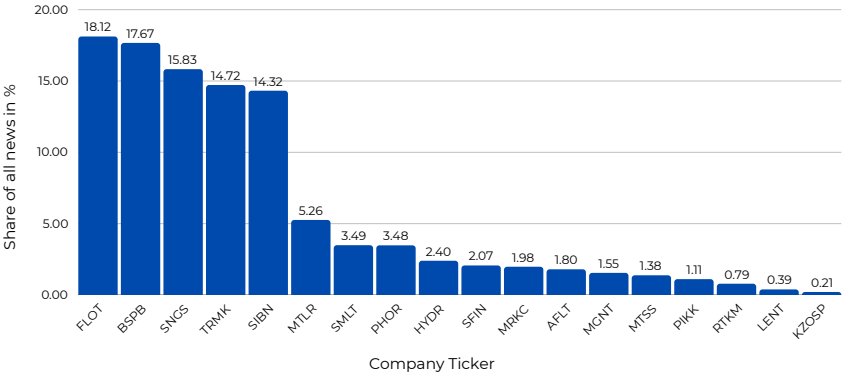

FIGURE 4. The distribution of news articles by company after filtration (Numbers on the diagram show percentage of news about the company in the dataset)

tickers serve as class labels.

To address the issue of news rewriting, we designed a Siamese neural network. We constructed a training dataset using the GigaChat API as follows: for each article, three paraphrased versions of both the title and body were generated. Then, pairs were randomly formed in equal proportion from the original and paraphrased news articles and their titles.

The Siamese neural network was designed as follows: a pair of news articles is fed as input, and vector representations of the articles are extracted using the RuBERT model [8]. The two vectors are then concatenated, and the resulting vector is passed through a fully connected neural network (MLP). To determine the optimal depth of the MLP model, we conducted a series of experiments, evaluating both prediction accuracy and news stream processing time. Based on the results, we selected the MLP architecture with three layers.

The filtered news articles are then converted into vectors so that duplicate classification can be performed in a one-shot mode when new articles arrive. This approach reduces both the processing time of the news stream and the computational resources required (in our case, a GPU V100).

To address the second challenge — matching news article samples by date and utilizing them for price forecasting — it is essential to formalize the data selection and prediction process. We assume that the closing price prediction for an asset is made for each trading day at the market opening. In this case, only news articles published before the start of the current trading day are included in the dataset.

The dataset is formed by grouping news articles based on their publication date. For predicting the price on a given day, only articles published on the previous trading day are used. For example, analytical articles such as those under the "Technical Analysis" section from the "BCS" source, which are published daily before the market opens, are included in the dataset for forecasting the prices of assets analyzed in those reports. This approach ensures that the most relevant information is considered, thereby improving prediction accuracy.

For the dual-modality approach, training sequences were formed by concatenating price return vectors from the previous five days with news stream vectors. The relative price return vectors were constructed similarly to the single-modality experiment, while news articles were selected from the previous trading day based on the chosen asset. These news articles were then transformed into vectors and aggregated.

If no publications were available on the previous day or before the market opened on the current day, a zero vector was concatenated with the relative price return vector of length 768 for the RuBERT model and 896 for the Vikhr-Qwen2.5-0.5b-Instruct (Qwen) model. Otherwise, the aggregated news vector of the same length was appended. These final vector lengths correspond to the output sizes of the pretrained RuBERT and Qwen models.

In this study, we explored two approaches for aggregating news vectors: vector summation (Sum) and averaged summation (Mean). By vector summation, we mean summing the values of corresponding vector coordinates. In the averaged summation approach, each coordinate of the aggregated vector is assigned the arithmetic mean of the corresponding coordinates across all aggregated vectors.

The baseline RuBERT model has a limited context window of 512 tokens. As a result, articles exceeding this limit were either truncated or

split for separate processing, meaning that a single news article could correspond to multiple vectors. In contrast, the Qwen model has a significantly larger context window of 32,768 tokens (64 times larger), allowing it to process entire articles without truncation. Next, we compare how different news vectorization methods impact the accuracy of price predictions.

The pointwise predicted return vectors were converted into asset prices using equation (2). The prediction quality was evaluated using two metrics: Accuracy and Mean Absolute Percentage Error (MAPE). Accuracy was measured as the proportion of correctly predicted signs of the return vector elements—either positive or negative. The MAPE metric indicates the average percentage deviation of the predicted price from the actual value. This allows us to assess the prediction quality not only in relative terms but also in absolute monetary units (rubles).

3. Computational experiment

In this section, we present the results of computational experiments for two predictive models (single- and dual-modalities). The predictive model was developed using the Transformers framework from the Hugging Face platform. All computations were performed on an NVIDIA V100 GPU.

3.1. The Single-Modality Approach Performance

The results of the experiment on predicting return vectors using only time series data for classical and deep learning models are presented in a Table 10.

Table 11 provides the averaged prediction quality metrics for all models, sorted in ascending order of the mean absolute percentage error (MAPE) (column “Deviation”).

From the experiment results, it is evident that the recurrent model LSTM achieves the best classification performance (predicting upward or downward trends) and regression accuracy (smallest deviation of the predicted price from the actual price). However, it lags slightly in terms of the mean absolute error metric.

TABLE 10. Results of forecasting return vectors using only time series. Accuracy (left) and deviation (right) in percent

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

TABLE 11. The Single-Modality approach forecast (time-series) inference metrics: Accuracy and MAPE in percentage

Model	Accuracy, %	MAPE, %
LSTM	**52.020**	**0.397**
XGB	45.000	1.627
KNN	46.010	1.631
RF	48.384	1.646
LinReg	50.152	1.669
DT	49.798	1.824

3.2. The Dual-Modality Approach Performance

The results of the second experiment, which involved merging the news stream with numerical time series data and comparing the proposed multimodal approach with a forecast based solely on candlestick time series, are presented in the Table 12.

The Table 13 provides the averaged prediction quality metrics for the considered models. The data in this table is sorted by the "Deviation" column in ascending order, reflecting the mean absolute percentage error (MAPE) of the predicted price deviations.

In this second experiment, the LSTM neural network was chosen as the baseline model. We compared different vectorization methods (RuBert, Qwen) and aggregation techniques (Sum, Mean) to evaluate their impact on prediction performance.

Figure 5 shows the dependence of the mean squared error (MSE Loss) function values on the number of training iterations for different models, based on the training set (from July 7, 2022, to March 27, 2024) and the test set (from March 28 to August 30, 2024). The graph indicates that after 30 training epochs, the curves reach a stationary value.

TABLE 12. The Dual-Modality returns vector forecasting metrics. Accuracy (the upper row), MAPE (the lower row) in percentage

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

TABLE 13. The Dual-Modality Approach forecast: Accuracy, MAPE

Model	Accuracy, %	MAPE, %
LSTM-Qwen-Mean	48.552	**0.256**
LSTM-Qwen-Sum	46.970	0.367
LSTM	**52.020**	0.397
LSTM-RuBert-Mean	49.798	0.437
LSTM-RuBert-Sum	48.148	0.445

The results from the tables implies that the forecast based on the vectorized news stream using a large language model outperforms the forecast built solely on candlestick data of assets, demonstrating the smallest deviation of the pointwise price prediction from the actual price vector. Additionally, averaging the vectors (Mean) provides the best results.

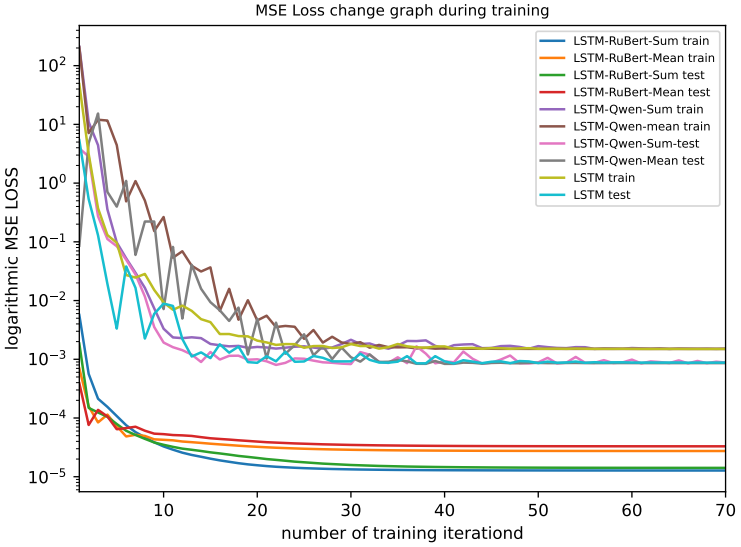

FIGURE 5. Dependence of the mean squared error function values on the number of training iterations for different models. Training and test sets

The dataset (176 stocks of Russian companies traded on the Moscow Exchange and 79,555 Russian-language financial news articles) collected for the study is available at [11].

Conclusion

As a result of the conducted experiments, we demonstrated that adding a textual modality—analyzing the news stream—positively impacts the accuracy of price prediction. On average, the MAPE metric (the deviation of the predicted price from the actual price) decreases by 55%: from 0.397 (LSTM model) to 0.256 (LSTM-Qwen-Mean model). Additionally, predictions based on vectors obtained using the large language model Vikhr-Qwen2.5-0.5b-Instruct outperformed those based on RuBert. This can be partly attributed to the fact that the Qwen model has a significantly larger context window and is trained on a larger text corpus with support for «Chain-of-Thought» (CoT) reasoning. This enhances the model's ability to reason and capture complex semantic dependencies within the text. The experimental results indicate that the averaging method (Mean) performed better than summation (Sum) and is the preferred method for aggregating news stream vectors.

At the same time, it is important to note that the test data, on which the final metric values were calculated, covers the period from March 28 to August 30, 2024. During this period, the Russian securities market exhibited a general downward trend. The presence of a clear trend is a significant factor that simplifies the prediction task. However, even in this setting, the proposed multimodal approach proved to be the best among those considered.

The training and validation of the model for the rewriting task were conducted on news articles whose length did not exceed the context window of the RuBert model. As a result, artifacts related to the context window size only became apparent during the forecasting phase when the news dataset included articles averaging around 290 words in length. For future improvements in news filtering and classification by company, it is necessary to utilize models with a larger context window, such as Qwen.

The collected dataset [11] demonstrates good structuring and can be used for fine-tuning large language models in Russian or adapted for the Russian language for applications in the financial sector.

TABLE 14. Multimodal approach forecasting metrics in comparison with the approach based on news sentiment score (Baseline) offered by [7]

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

For a quantitative comparison of the proposed model, we conducted a computational experiment based on the approach and metrics from the study [7]. Following the methodology of [7], we used time series data of stock prices from five major American companies: AAPL, AMZN, GOOGL, NFLX, and TSLA, along with a dataset of English-language news articles labeled by company for the period from October 12, 2012 to January 31, 2020 (Table 14).

It is worth noting that the dataset used includes text data in English; therefore, we utilized the original Qwen2.5-0.5b-Instruct model [10] for news vectorization. To generate forecasts, we selected and trained the *LSTM-Qwen-Mean* model, as it demonstrated the best overall performance in our study. For evaluation, we used the coefficient of determination (*R2*), mean absolute error (*MAE*), and mean absolute percentage error (*MAPE*).

Thus, we worked with the same time series and evaluation metrics. Across all metrics, except for *MAE* on NFLX and *R2* on TSLA, the proposed multimodal approach with vector averaging outperformed the best-performing results from the approach in [7]. Based on our computational experiments, we conclude that the proposed multimodal approach demonstrated superior forecasting quality and greater adaptability to both Russian and international markets.

In the future, it is necessary to explore how to incorporate the incoming news stream into the predictive model—specifically, the optimal time window for using news data and the best approach for weighting news messages (e. g. adjusting the weight of a news article based on its chronological position in the dataset).

References

[1] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. Chitkushev, D. Trajanov. "Evaluation of sentiment analysis in finance: from lexicons to transformers", *IEEE Access*, **8** (2020), pp. 131662–131682. doi ↑84

[2] T.-T. Ho, Y. Huang. "Stock price movement prediction using sentiment analysis and CandleStick chart representation", *Sensors*, **21**:23 (2021), id. 7957, 18 pp. doi ↑84

[3] M. Jaggi, P. Mandal, S. Narang, U. Naseem, M. Khushi. "Text mining of stocktwits data for predicting stock prices", *Applied System Innovation*, **4**:1 (2021), id. 13, 22 pp. doi ↑84

[4] B. Fazlija, P. Harder. "Using financial news sentiment for stock price direction prediction", *Mathematics*, **10**:13 (2022), id. 2156, 20 pp. doi ↑84

[5] Y. Xinli, Ch. Zheng, L. Yuan, D. Shujing, L. Zongyi, L. Yanbin. *Temporal data meets LLM — Explainable financial time series forecasting*, 2023, 13 pp. arXiv 2306.11025 [cs.LG] ↑84

[6] Zh. Boyu, Y. Hongyang, X.-Y. Liu. *Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models*, 2023, 7 pp. arXiv 2306.12659 [cs.CL] ↑84

[7] T. D. Kulikova, E. Y. Kovtun, S. A. Budennyy. "Do we benefit from the categorization of the news flow in the stock price prediction problem?", *Dokl. Math.*, **108**, Suppl. 2 (2023), pp. S503–S510. doi ↑84, 104

[8] Y. Kuratov, M. Arkhipov. *Adaptation of deep bidirectional multilingual transformers for Russian language*, 2019, 8 pp. arXiv 1905.07213 [cs.CL] ↑86, 95, 96

[9] A. Nikolich, K. Korolev, A. Shelmanov, I. Kiselev. *Vikhr: The family of open-source instruction-tuned large language models for Russian*, 2024, 8 pp. arXiv 2405.13929 [cs.CL] ↑86, 95

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

X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Zh. Zhang, Zh. Guo, Zh. Fan. *Qwen2 Technical Report*, 2024, 26 pp. arXiv 2407.10671 [cs.CL] ↑104

[11] K. Khubiev. *Russian financial news dataset*, Kaggle Platform, 2025. URL doi ↑103

Received 24.12.2024;
approved after reviewing 30.01.2025;
accepted for publication 27.02.2025;
published online 11.03.2025.

Information about the authors:

Kasymkhan Usufovich Khubiyev

Researcher, Center of Social and Economic Forecasting; Master's Student of "Financial Mathematics and Financial Technologies", Sirius University of Science and Technology, Sirius, Russia. Research interests: artificial intelligence and its application in science, finance, industry, and business

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

Mikhail Evgenyevich Semenov

PhD in Physics and Mathematics, Scientific Supervisor of the "Financial Mathematics and Financial Technologies" direction, Sirius University of Science and Technology, Sirius, Russia. Research interests: Information technology, intelligent data processing and analysis technologies.

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

The authors contributed equally to this article.

The authors declare no conflicts of interests.

RU EN ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ ISSN 2079-3316
научная статья ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ, ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ, НЕЙРОННЫЕ СЕТИ

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

Мультимодальное предсказание цен акций на примере российского рынка ценных бумаг

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

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

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

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

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

Полученный мультимодальный набор данных представляет ценность для дальнейшей адаптации языковых моделей в финансовой сфере. Перспективные направления исследований включают оптимизацию параметров текстовой модальности, таких как временное окно, тональность и хронологический порядок новостных сообщений. (*Связанные тексты статьи на английском и на русском языках*)

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

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

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

Введение

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

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

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

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

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

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

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

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

Для достижения поставленной цели сформулированы следующие основные задачи.

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

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

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

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

1. Сбор и структурирование данных

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

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

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

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

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

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

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

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

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

Таблица 1. Статистические характеристики набора данных после токенизации, RuBert

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

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

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

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

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

Таблица 3. Примеры оригинального и токенизированного текста

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

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

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

(тикер, наименование компании, описание деятельности компании).

Такие данные необходимы в нашем случае для:

(*а*) извлечения ключевых слов из описания,

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

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

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

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

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

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

2. Методология

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

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

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

2.1. Использование одной модальности

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

Концептуально эксперимент состоит из двух задач:

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

($б$) предсказание цены – задача регрессии.

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

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

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

Таблица 6. Распределение новостей по компаниям после фильтрации

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

Таблица 7. Описательные характеристики для акций компаний

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

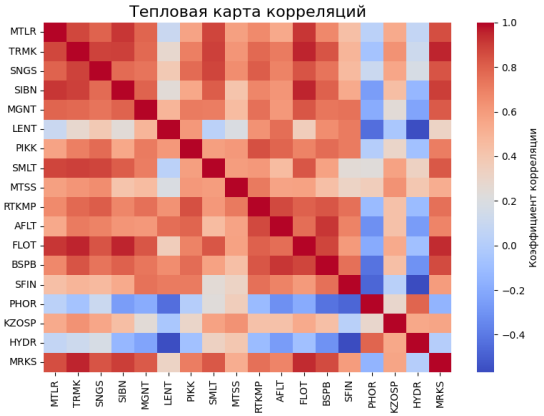

Рисунок 1. Тепловая карта корреляций цен закрытия активов

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

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

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

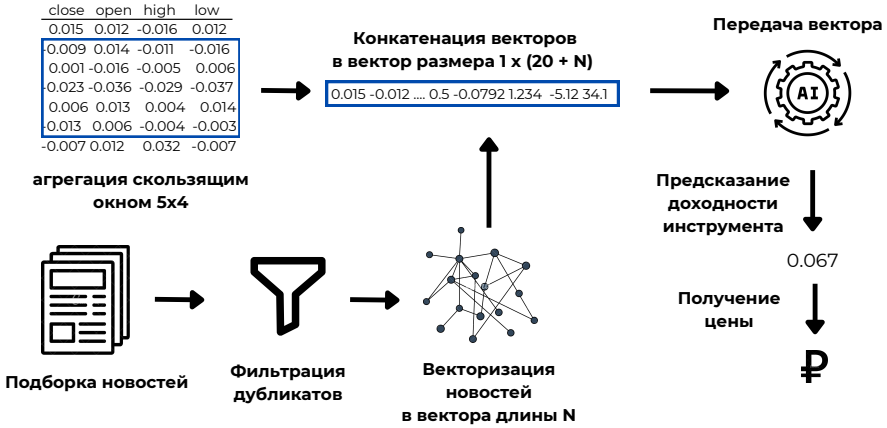

Рисунок 3. Процесс разработки модели для использования одной и двух модальностей

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

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

На выходе модели – предсказание на следующую торговую сессию.

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

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

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

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

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

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

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

2.2. Использование двух модальностей

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

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

Таблица 8. Ключевые слова, полученные из описаний компаний

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

Таблица 9. Ключевые слова, полученные дополнительно

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

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

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

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

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

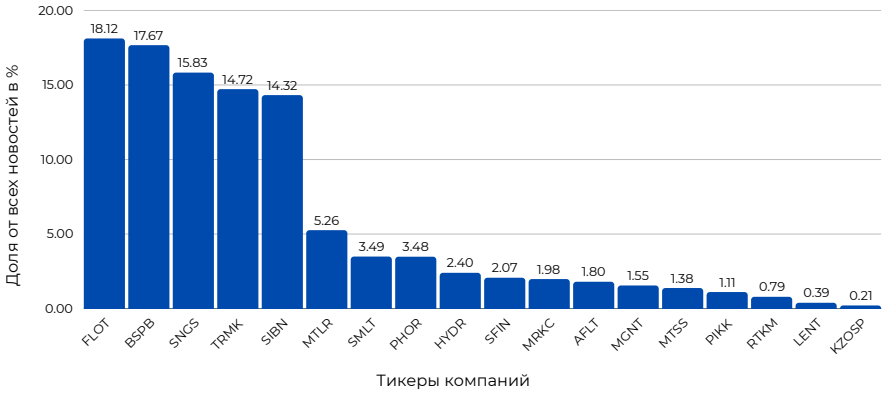

Рисунок 4. Распределение новостей по компаниям после фильтрации (цифры на диаграмме – процент новостей про компанию в наборе данных)

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

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

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

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

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

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

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

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

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

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

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

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

3. Вычислительный эксперимент

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

3.1. Результаты использования одной модальности

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

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

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

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

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

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

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

3.2. Результаты использования двух модальностей

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

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

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

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

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

Таблица 13. Результаты предсказания с использованием двух модальностей

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

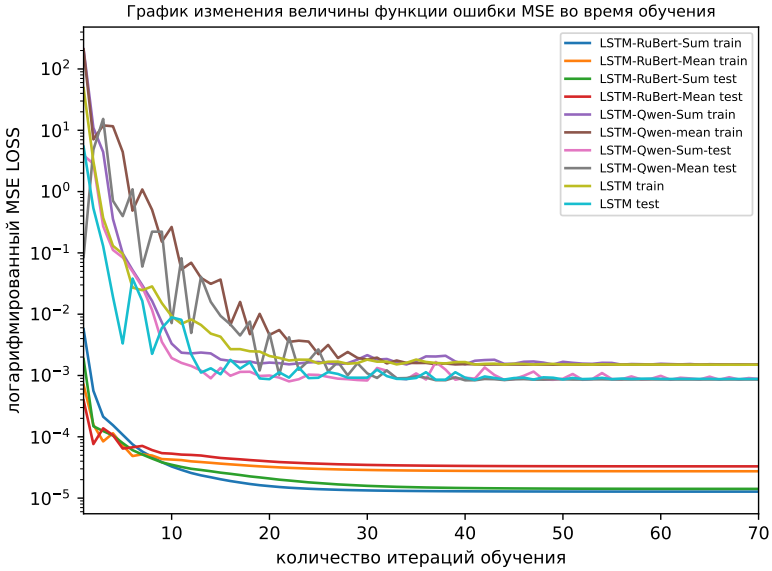

Рисунок 5. Зависимость величин функции ошибки среднеквадратичного отклонения от количества итераций обучения для различных моделей. Тренировочный и тестовой наборы

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

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

Заключение

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

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

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

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

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

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

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

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

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

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

Список использованных источников

[1] Mishev K., Gjorgjevikj A., Vodenska I., Chitkushev L., Trajanov D. *Evaluation of sentiment analysis in finance: from lexicons to transformers* // IEEE Access.– 2020.– Vol. **8**.– Pp. 131662–131682. ↑108

[2] Ho T.-T., Huang Y. *Stock price movement prediction using sentiment analysis and CandleStick chart representation* // Sensors.– 2021.– Vol. **21**.– No. 23.– id. 7957.– 18 pp. ↑108

[3] Jaggi M., Mandal P., Narang S., Naseem U., Khushi M. *Text mining of stocktwits data for predicting stock prices* // Applied System Innovation.– 2021.– Vol. **4**.– No. 1.– id. 13.– 22 pp. ↑108

[4] Fazlija B., Harder P. *Using financial news sentiment for stock price direction prediction* // Mathematics.– 2022.– Vol. **10**.– No. 13.– id. 2156.– 20 pp. ↑108

[5] Xinli Y., Zheng Ch., Yuan L., Shujing D., Zongyi L., Yanbin L. *Temporal data meets LLM — Explainable financial time series forecasting.*– 2023.– 13 pp. arXiv 2306.11025 [cs.LG] ↑108

[6] Boyu Zh., Hongyang Y., Liu X.-Y. *Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models.*– 2023.– 7 pp. arXiv 2306.12659 [cs.CL] ↑108

[7] Куликова Т. Д., Ковтун Е. Ю., Буденный С. А. *Получаем ли мы пользу от категоризации потока новостей в задаче прогнозирования цен акций?* // Доклады Российской академии наук. Математика, информатика, процессы управления.– 2023.– Т. **514**.– № 2.– С. 385–394. ↑108, 128

[8] Kuratov Y., Arkhipov M. *Adaptation of deep bidirectional multilingual transformers for Russian language.*– 2019.– 8 pp. arXiv 1905.07213 [cs.CL] ↑110, 119, 120

[9] Nikolich A., Korolev K., Shelmanov A., Kiselev I. *Vikhr: The family of open-source instruction-tuned large language models for Russian.*– 2024.– 8 pp. arXiv 2405.13929 [cs.CL] ↑110, 119

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

X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Zh. Zhang, Zh. Guo, Zh. Fan *Qwen2 Technical Report.*– 2024.– 26 pp. arXiv 2407.10671 [cs.CL] ↑128

[11] Khubiev K. *Russian financial news dataset.*– Kaggle Platform.– 2025. URL doi ↑127

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

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

Информация об авторах:

Касымхан Юсуфович Хубиев

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

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

Михаил Евгеньевич Семенов

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

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

Авторы внесли равный вклад в подготовку публикации.

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