NATIONAL RESEARCH UNIVERSITY

АНАЛИТИЧЕСКИЕ ИНСТРУМЕНТЫ АНАЛИЗА ИНФОРМАЦИОННОГО ПОЛЯ ЧАСТНЫХ ИНВЕСТОРОВ: ПОИСК РЫНОЧНЫХ ПРОВАЛОВ И ПРИМЕНЕНИЕ ИИ ДЛЯ ПРОГНОЗА ДИНАМИКИ БИРЖЕВЫХ ХАРАКТЕРИСТИК АКТИВОВ

Файзулин М.С.
Центр финансовых исследований и анализа данных (НИУ ВШЭ)

Исследование выполнено в рамках гранта РНФ (№ 22-18-00276 от 13.05.2022)

Санкт-Петербург, 2024

ЦЕЛЬ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Цель исследования: оценить эффективность применения технических индикаторов и дополнительных метрик сентимента и состояния финансового рынка России по отношению к прогнозу динамики цен и волатильности по обсуждаемым акциям компаний.

Предмет: прогноз доходности и уровня риска акций отечественных эмитентов за счет анализа информационной среды частных инвесторов и биржевых характеристик акций компаний.

Задачи:

1. **Сформировать систему сбора и классификации текстовых данных.**
2. **Построение метрик сентимента для оценки настроений среди частных инвесторов.**
3. **Сформировать систему базовых технических индикаторов на основе биржевых характеристик акций.**
4. **Провести ряд тестов по отбору значимых признаков и выбора модельного решения для прогноза будущего движения цен и волатильности по наиболее обсуждаемым акциям российских компаний.**

МЕТОДОЛОГИЯ И ДАННЫЕ

Используемая выборка данных:

1) **Отобраны 5 наиболее обсуждаемых акций российских компаний, по которым было больше всего сообщений с октября 2019 г. по март 2024 г. (SBER, LKOH, MTLR, VTBR, GAZP)**
2) **По отобранным активам рассчитаны технические индикаторы и метрики сентимента частных инвесторов.**
3) **Все данные представлены с дневной частотой. Для избежания сезонных эффектов и проблемы «больших чисел» была проведена нормализация данных с целью достижения их стационарности.**

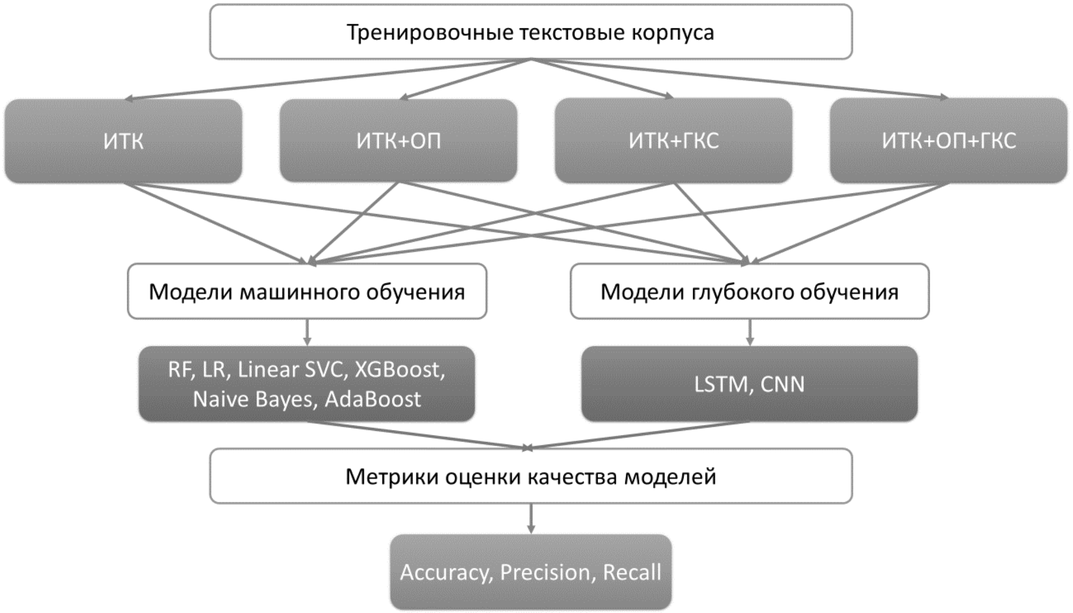
Эндогенные переменные

SO_14	Стохастический осциллятор за 14 дней
SO_5	Стохастический осциллятор за 5 дней
BB_20	Полосы Боллинджера за 20 дней
BB_10	Полосы Боллинджера за 10 дней
MACD_Close	Индикатор MACD на основе цен в момент закрытия биржи за 12 и 26 дней
MACD_Open	Индикатор MACD на основе цен в момент открытия биржи за 12 и 26 дней
RSI_14_5	Индекс относительной силы за 14 дней (пятидневная доходность)
RSI_14_1	Индекс относительной силы за 14 дней (однодневная доходность)

Экзогенные переменные

RVI	Индекс волатильности российского рынка (RVI)
AKRA	Индекса финансового стреса от агенства AKRA
All_log_Sent	Совокупный уровень сентимента по наиболее обсуждаемым акциям
MFD_Log_Sent	Сентимент по наиболее обсуждаемым акциям за счет обсуждений на форуме MFD
Pulse_Log_Sent	Сентимент по наиболее обсуждаемым акциям за счет обсуждений на Tinkoff Pulse
SL_Log_Sent	Сентимент по наиболее обсуждаемым акциям за счет обсуждений на SmartLab

МЕТОДОЛОГИЯ И ДАННЫЕ



ИТК – исходный текстовый корпус

ОП – генерация текстовых данных на основе обратного перевода

ГКС – генерация текстовых данных на основе матрицы косинусного сходства

Algorithm 1 Генерация новых данных для расширения текстового корпуса

Шаг 1: Обратный перевод (цикл: русский-английский-русский)
Шаг 2: Перевод текстовых данных в количественные при помощи алгоритма TF-IDF
Шаг 3: Генерация новых предложений на основе матрицы схожести предложений
Шаг 4: Отбор наиболее схожего предложения
Шаг 5: Перемешивание слов отобранных предложений
Шаг 6: Формирование n-грам на основе перемешанных слов и создание общего сообщения
Шаг 7: Добавление сгенерированного сообщения в тренировочный корпус

$$TF(t,d) = \frac{n_t}{\sum_{i=1}^{k} n_i}, (1)$$

$$IDF(t,d) = log\frac{n_d}{1+\sum_{j=1}^{h} n_j}, (2)$$

$$TF - IDF(t,d) = TF(t,d) * IDF(t), (3)$$

n_t – частота вхождения элемента (t) в отдельном сообщении (d);

n_i – общее число элементов (i), содержащихся в сообщении (d);

n_f – сумма всех сообщений (d) в выборке;

n_j – количество сообщений, в которых содержится элемент (t).

МЕТОДОЛОГИЯ И ДАННЫЕ

Результаты классификации текстовых данных тестовой выборки

class	RF	LR	Linear SVC	XGBoost	AdaBoost	Naive Bayes	LSTM	CNN	Avg. Acc.
	Basic Train								
	Accuracy								
	59,97	57,75	**58,19**	59,01	**57,23**	54,63	50,04	**48,93**	**55,72**
	Precision								
negative	66,39	61,41	62,60	59,08	65,18	56,20	39,35	41,72	56,49
neutral	58,64	57,35	56,91	60,38	52,41	57,71	59,26	52,31	56,87
positive	56,33	55,15	56,13	57,14	58,74	50,78	52,21	53,10	**54,95**
	Recall								
negative	51,83	48,60	48,60	52,47	47,10	49,68	48,03	46,97	49,16
neutral	73,84	72,22	74,31	74,07	80,56	57,18	51,20	51,83	66,90
positive	55,09	53,32	52,65	51,33	45,35	57,30	50,43	48,00	51,68
	Basic Train + Back Translation								
	Accuracy								
	59,97	**59,08**	54,48	**59,53**	57,08	**55,08**	**51,96**	49,52	55,84
	Precision								
negative	63,08	59,29	55,16	62,34	61,83	55,30	48,60	43,23	56,10
neutral	60,91	62,37	59,45	59,54	53,82	58,94	56,25	57,64	**58,62**
positive	56,02	55,56	49,49	57,05	57,60	51,63	51,33	48,23	53,36
	Recall								
negative	52,90	54,19	50,54	61,61	49,46	52,69	51,60	49,75	52,84
neutral	74,31	67,13	59,72	72,22	75,00	56,48	54,36	51,98	63,90
positive	53,54	56,42	53,54	55,53	47,79	56,19	50,00	46,78	52,47
	Basic Train + Proposed Text Generation Method								
	Accuracy								
	59,38	57,60	55,37	59,30	57,01	50,11	49,44	49,22	54,68
	Precision								
negative	63,39	58,35	55,80	62,02	66,57	46,50	56,56	44,95	**56,77**
neutral	59,12	58,20	56,14	58,83	50,73	62,44	43,75	53,25	55,31
positive	56,36	56,25	54,21	57,38	60,24	49,55	47,57	49,78	53,92
	Recall								
negative	49,89	51,83	49,68	51,61	48,39	68,60	46,88	46,65	51,69
neutral	74,31	65,74	61,34	74,77	80,32	31,94	53,39	51,22	61,63
positive	54,87	55,75	55,53	52,43	43,58	48,45	49,54	49,78	51,24
	Basic Train + Proposed Text Generation Method + Back Translation								
	Accuracy								
	60,04	57,97	54,41	57,38	56,63	50,85	51,37	49,15	54,73
	Precision								
negative	62,02	57,58	54,79	60,50	62,85	50,00	46,02	50,75	55,56
neutral	59,45	61,05	58,87	58,30	53,02	59,80	50,93	44,68	55,76
positive	58,98	55,32	50,20	53,36	56,61	46,65	57,30	51,77	53,77
	Recall								
negative	51,61	52,26	51,61	52,04	48,39	52,69	50,12	45,83	50,57
neutral	75,69	64,58	57,64	69,91	75,23	42,36	58,05	60,69	63,02
positive	53,76	57,52	54,20	50,08	47,35	57,08	47,70	45,35	51,63

Результаты тестирования стекинг-ансамбля

Архитектура модели:	Level 1: RF-LR-SVC-XGBoost-AdaBoost. Level 2: LR	
Class	Precision	Recall
negative	64,45	54,19
neutral	63,05	72,69
positive	57,89	58,85
Accuracy	61,68	

Всего с трех онлайн-платформ было получено и классифицировано 4 329 852 сообщения

(Tinkoff Pulse: 1 917 172, MFD: 1 797 739 и SmartLab: 614 941).

положительная (1) – если сообщение носит характер одобрения фондового актива;

нейтральная (0) – если сообщение не относится к выражении позиции пользователя к активу;

негативная (-1) – если сообщение характеризует явное неодобрение анализируемого актива.

Всего в предварительно размеченной выборке 6741 текстовых сообщений по 2247 сообщений каждой из трех тональностей.

МЕТОДОЛОГИЯ И ДАННЫЕ

Метрики сентимента инвесторов

$$S_{i,t} = ln\frac{1+N_{i,t}^{pos}}{1+N_{i,t}^{neg}}, (1)$$

$$All\_\log\_Sent_{i,t} = ln\frac{1+N_{i,t}^{pos}}{1+N_{i,t}^{neg}}, (2)$$

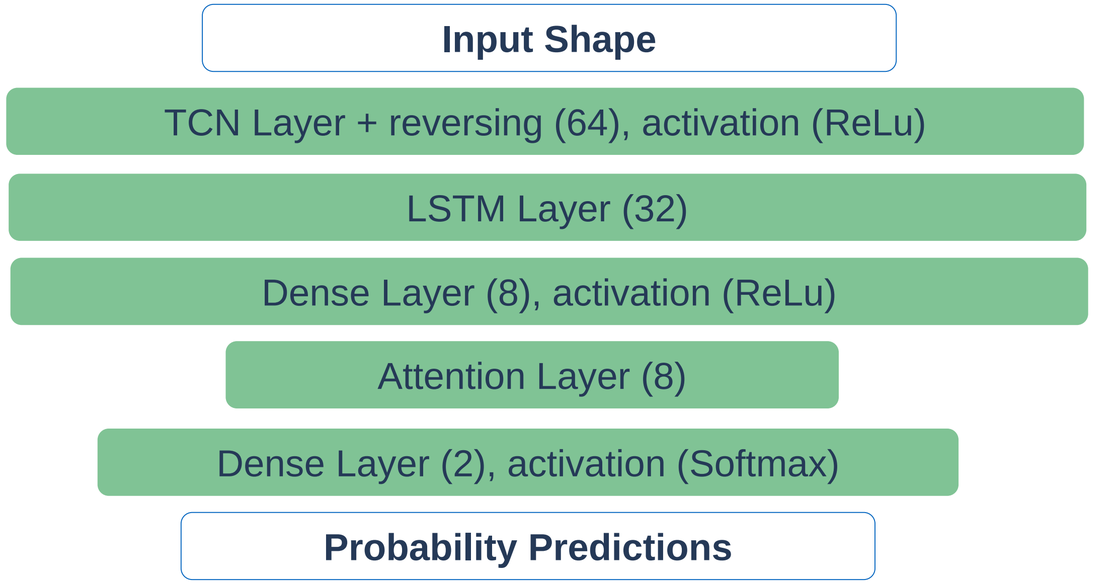
Компонент индекса	Примеры поисковых запросов
Уровень сентимента инвесторов ($S_{i,t}$) отдельно по каждой платформе: MFD, Tinkoff Pulse и SmartLab.	$S_{i,t}$ – сентимент i-ой компании в момент времени t; $N_{i,t}^{pos}$ – количество позитивных сообщений по компании i в момент времени t по отдельной платформе; $N_{i,t}^{neg}$ – количество позитивных сообщений по компании i в момент времени t по отдельной платформе.
Общий уровень сентимента по всем платформам ($All\_\log\_Sent_{i,t}$).	$N_{i,t}^{pos}$ – количество позитивных сообщений по компании i в момент времени t по всем трем платформам; $N_{i,t}^{neg}$ – количество позитивных сообщений по компании i в момент времени t по всем трем платформам.

Сентимент-метрика ($S_{i,t}$) строится по каждой паре платформ:

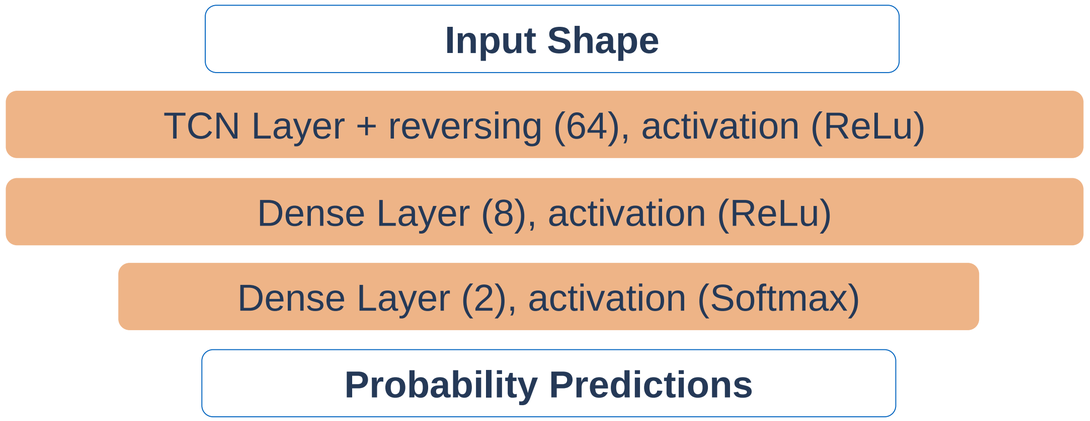
Tinkoff Pulse (Pulse_Log_Sent), Smart-Lab (SL_Log_Sent), MFD (MFD_Log_Sent).

МЕТОДОЛОГИЯ И ДАННЫЕ (МОДЕЛЬНЫЕ РЕШЕНИЯ)

Прогноз будущего движения цены акции



Прогноз будущего движения волатильности по акции



Логит-модели

$$Return\_movement_{i,t+1} = \alpha + \beta_q \delta_{i,t} + \beta_h \tau_{i,t} + \beta_z \mu_{i,t} + \varepsilon_{i,t}$$

$$Volatility\_movement_{i,t+1} = \alpha + \beta_q \delta_{i,t} + \beta_h \tau_{i,t} + \beta_z \mu_{i,t} + \varepsilon_{i,t}$$

где $\tau_{i,t}$ – вектор технических индикаторов,

$\delta_{i,t}$ – вектор метрик сентимента,

$\mu_{i,t}$ – вектор показателей состояния финансового рынка России (RVI, AKRA).

CatBoost

Параметр	Пределы значений
max_depth	Integer(1, 6)
n_estimators	Integer(100, 1000)
learning_rate	Real(0.01, 0.5, 'log-uniform')
l2_leaf_reg	Real(1e-5, 1, 'log-uniform')
bagging_temperature	Real(0.1, 1.0)
random_strength	Real(0.1, 1.0)

Random Forest

Параметр	Пределы значений
n_estimators	Integer(100, 400)
max_features	Real(0.01, 1.0, 'log-uniform')
max_depth	Integer(2, 20)
min_weight_fraction_leaf	Real(1e-6, 1e-2, 'log-uniform')
min_samples_leaf	Integer(2, 100)
min_samples_split	Integer(2, 100)

РЕЗУЛЬТАТЫ (ПРОГНОЗ РОСТА ЦЕН НА АКЦИИ КОМПАНИЙ)

Прогноз на 1 день вперед

Model	Accuracy	Test Profit	Real Profit	Формат данных
NN	71,03%	70,12%	113,77%	Все данные
RF	72,59%	74,87%		
CB	72,76%	77,12%		
NN	73,10%	73,88%		Логит фильтрация (все данные)
RF	73,79%	78,48%		
CB	73,10%	77,17%		
NN	71,90%	71,96%		Только TI и Макропоказатели
RF	73,79%	78,89%		
CB	73,45%	78,56%		
NN	75,00%	79,98%		Логит фильтрация (Только TI)
RF	72,76%	76,72%		
CB	72,59%	77,35%		

Прогноз на 1 день вперед (фильтрация)

Показатель	Коэффициент	Робастная ст. ошибка	z-значение	p-value	
const	-0,0550	0,0290	-1,8950	0,0581	*
SO_14	**0,5423**	**0,0498**	**10,8900**	**0,0000**	***
BB_20	**0,5005**	**0,0668**	**7,4910**	**0,0000**	***
MACD_Close	-0,3433	0,2243	-1,5310	0,1258	
MACD_Open	0,1577	0,2053	0,7683	0,4423	
RSI_14_5	**0,4154**	**0,1339**	**3,1030**	**0,0019**	***
RSI_14_1	**2,5868**	**0,2950**	**8,7700**	**0,0000**	***
RVI	-0,1640	0,4141	-0,3959	0,6922	
AKRA	0,0326	0,4582	0,0711	0,9433	
MFD_Log_Sent	**0,1462**	**0,0742**	**1,9710**	**0,0487**	**
Pulse_Log_Sent	**0,2252**	**0,0936**	**2,4060**	**0,0161**	**
SL_Log_Sent	**0,1434**	**0,0669**	**2,1430**	**0,0321**	**
All_log_Sent	-0,1246	0,1996	-0,6240	0,5326	
R-квадрат	0,1305				
AIC	6757,3880				
BIC	6843,5540				
Accuracy	69,80%				

Прогноз на 2 дня вперед

Model	Accuracy	Test Profit	Real Profit	Формат данных
NN	67,13%	68,47%	142,72%	Все данные
RF	67,30%	74,19%		
CB	67,65%	74,85%		
NN	65,22%	59,02%		Логит фильтрация (все данные)
RF	67,48%	74,41%		
CB	67,13%	74,18%		
NN	68,00%	69,88%		Только TI и Макропоказатели
RF	67,65%	77,10%		
CB	67,65%	75,56%		
NN	67,65%	73,23%		Логит фильтрация (Только TI)
RF	67,13%	72,69%		
CB	67,65%	74,87%		

Прогноз на 5 дней вперед

Model	Accuracy	Test Profit	Real Profit	Формат данных
NN	63,21%	69,78%	199,18%	Все данные
RF	61,61%	80,20%		
CB	61,07%	78,53%		
NN	61,07%	65,06%		Логит фильтрация (все данные)
RF	61,79%	74,37%		
CB	60,36%	70,63%		
NN	60,71%	67,88%		Только TI и Макропоказатели
RF	60,89%	75,05%		
CB	61,25%	77,58%		
NN	61,25%	67,99%		Логит фильтрация (Только TI)
RF	61,79%	79,42%		
CB	61,43%	80,04%		

РЕЗУЛЬТАТЫ (ПРОГНОЗ ВОЛАТИЛЬНОСТИ ЦЕН АКЦИЙ)

Прогноз на 1 день вперед

Model	Accuracy	Формат данных
NN	61,03%	Все данные
RF	61,03%	
CB	60,69%	
NN	54,14%	Логит фильтрация (все данные)
RF	50,52%	
CB	50,17%	
NN	58,79%	Только TI и Макропоказатели
RF	60,86%	
CB	60,52%	

Прогноз на 2 дня вперед

Model	Accuracy	Формат данных
NN	59,83%	Все данные
RF	61,04%	
CB	62,26%	
NN	49,91%	Логит фильтрация (все данные)
RF	47,83%	
CB	49,57%	
NN	61,57%	Только TI и Макропоказатели
RF	63,65%	
CB	62,43%	

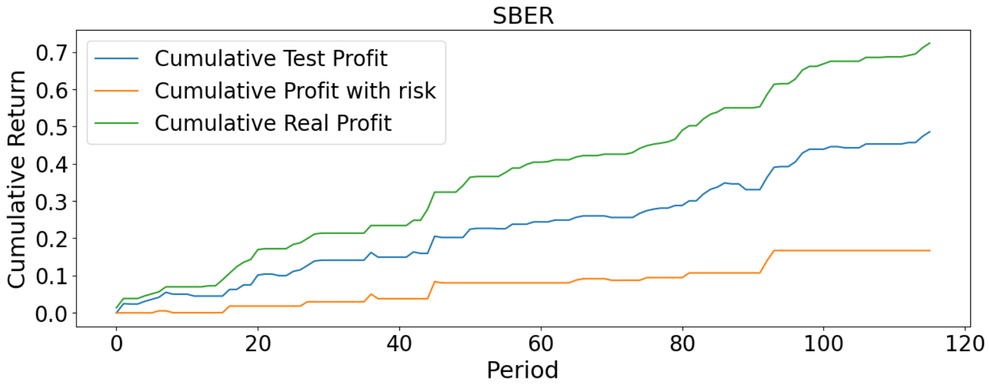
Прогноз на 5 дней вперед

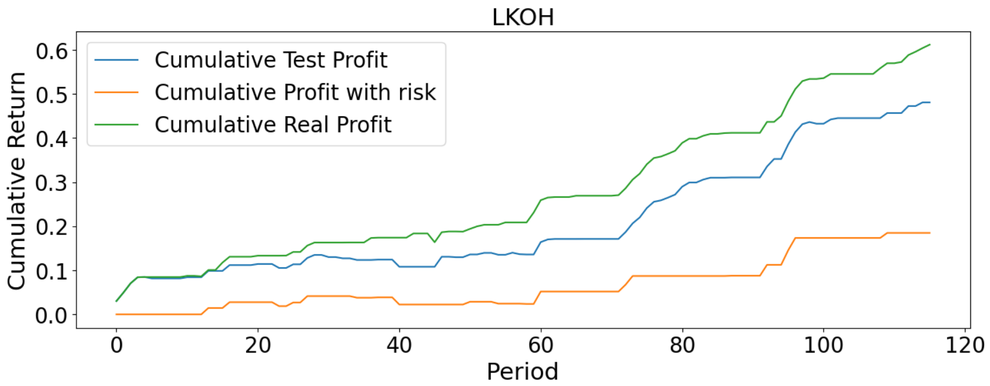
Model	Accuracy	Формат данных
NN	58,75%	Все данные
RF	61,07%	
CB	59,64%	
NN	51,43%	Логит фильтрация (все данные)
RF	48,21%	
CB	50,71%	
NN	60,36%	Только TI и Макропоказатели
RF	60,71%	
CB	62,50%	

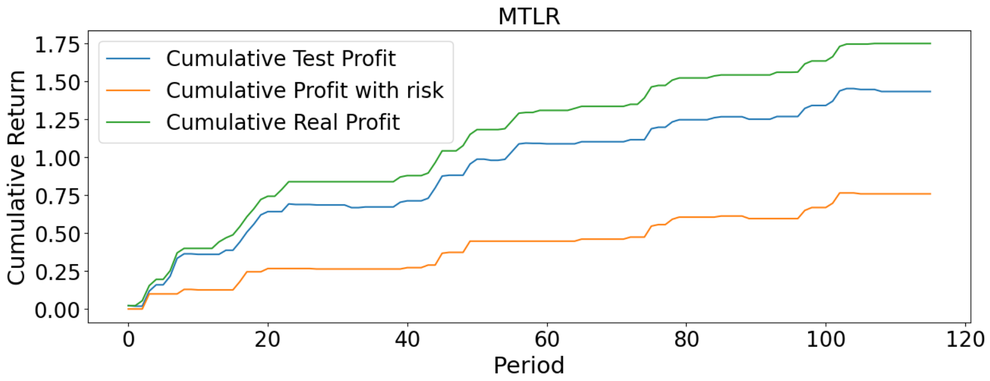
Прогноз на 1 день вперед (фильтрация)

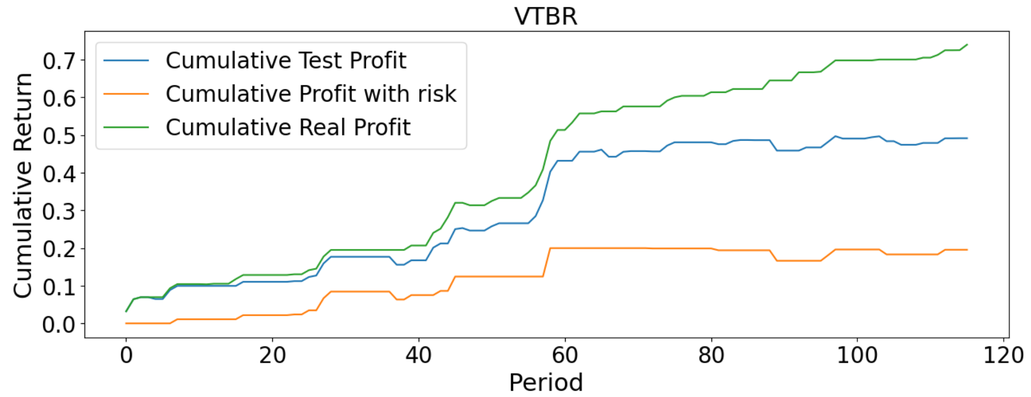
Показатель	Коэффициент	Робастная ст. ошибка	z-значение	p-value	
const	-0,0420	0,0268	-1,5660	0,1174	
SO_14	0,0321	0,0282	1,1400	0,2543	
BB_20	0,0001	0,0480	0,0025	0,9980	
MACD_Close	0,0313	0,1567	0,1996	0,8418	
MACD_Open	0,0033	0,1519	0,0218	0,9827	
RSI_14_5	0,0918	0,0999	0,9184	0,3584	
RSI_14_1	-0,0853	0,1863	-0,4580	0,6469	
RVI	-0,0915	0,3828	-0,2390	0,8111	
AKRA	**0,7729**	**0,4116**	**1,8780**	**0,0604**	*
MFD_Log_Sent	**-0,3135**	**0,1853**	**-1,6920**	**0,0906**	*
Pulse_Log_Sent	0,0598	0,0675	0,8866	0,3753	
SL_Log_Sent	-0,0546	0,0840	-0,6495	0,5160	
All_log_Sent	**0,1266**	**0,0632**	**2,0020**	**0,0453**	**
R-квадрат	0,0021				
AIC	7752,4500				
BIC	7838,6160				
Accuracy	52,30%				

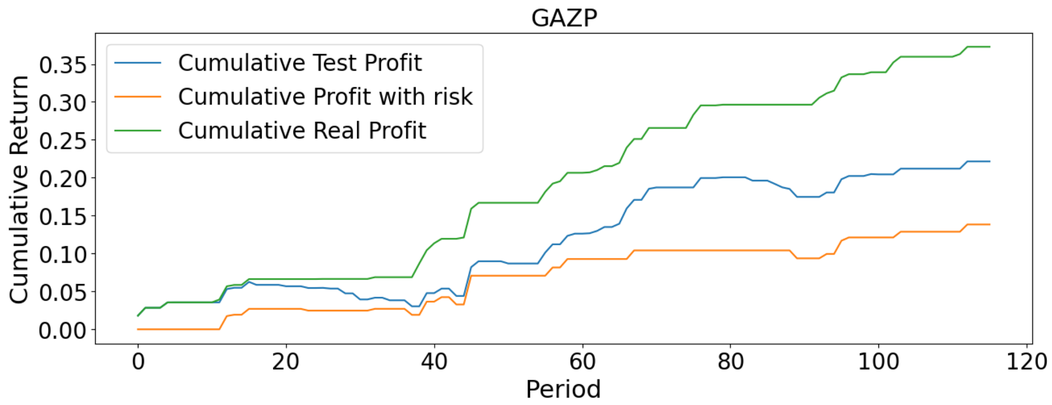
РЕЗУЛЬТАТЫ (ДОХОДНОСТЬ ТОРГОВЫХ СТРАТЕГИЙ – НА 1 ДЕНЬ ВПЕРЕД)











SBER: 48,58% (without risk correction)

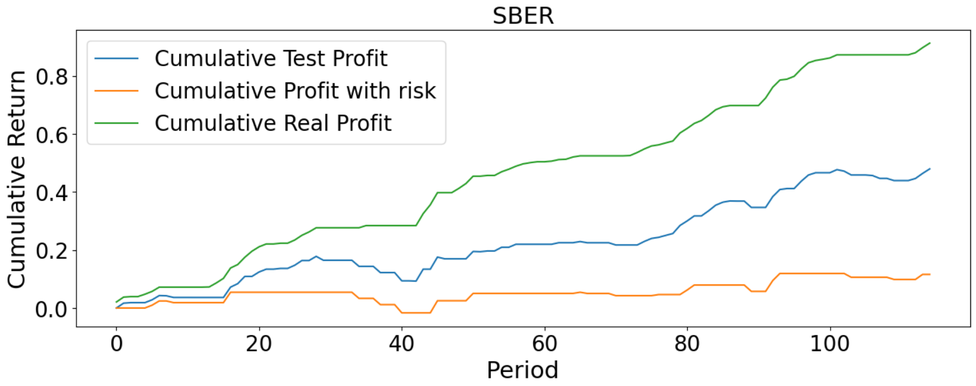
LKOH: 48.11% (without risk correction)

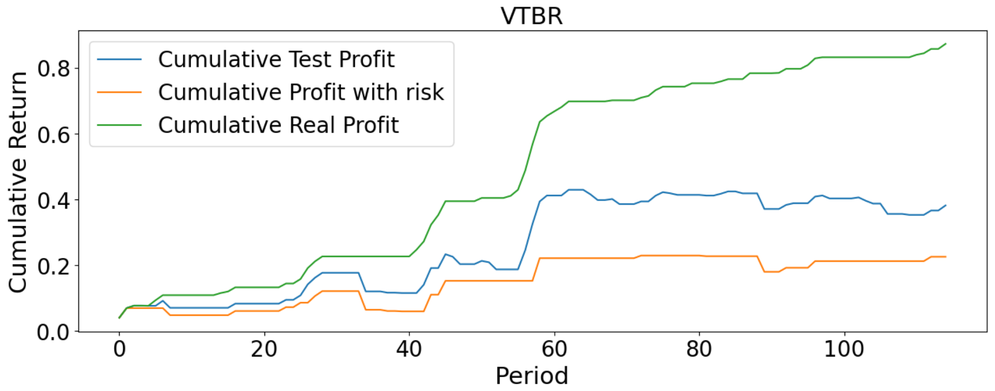
MTLR: 143,30% (without risk correction)

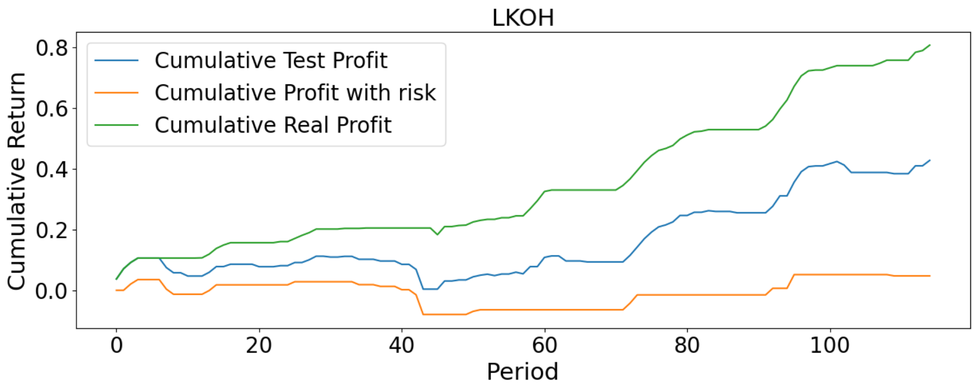
VTBR: 48,17% (without risk correction)

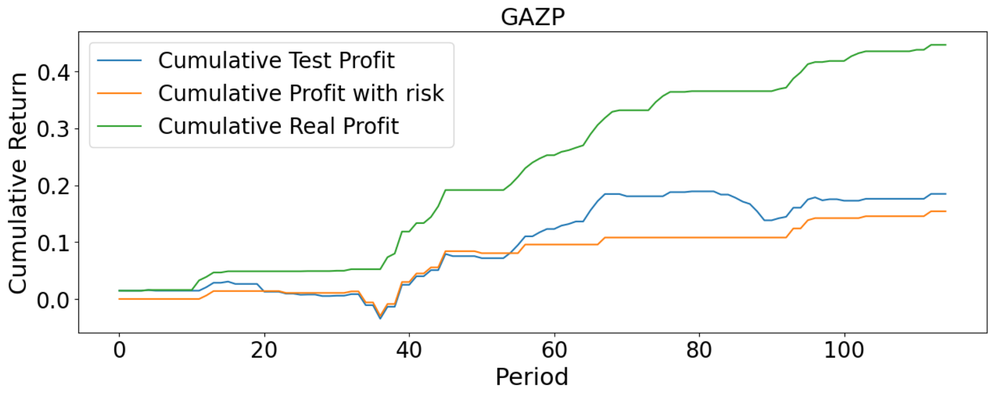
GAZP: 22,15% (without risk correction)

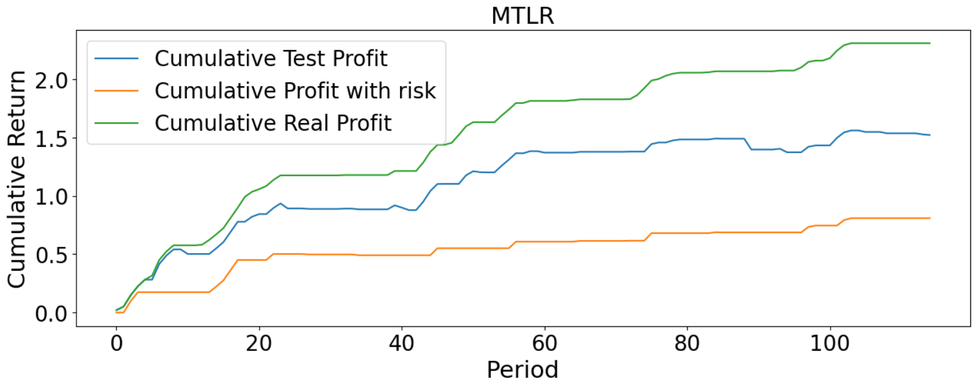
РЕЗУЛЬТАТЫ (ДОХОДНОСТЬ ТОРГОВЫХ СТРАТЕГИЙ – НА 2 ДНЯ ВПЕРЕД)











SBER: 48,38% (without risk correction)

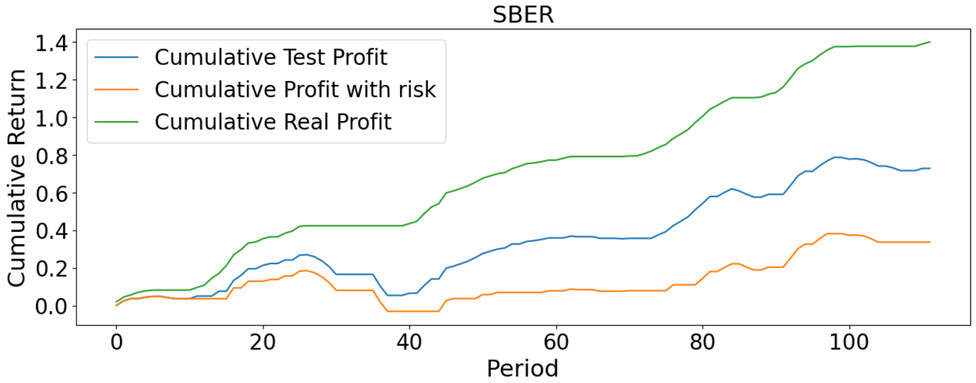
LKOH: 44.07% (without risk correction)

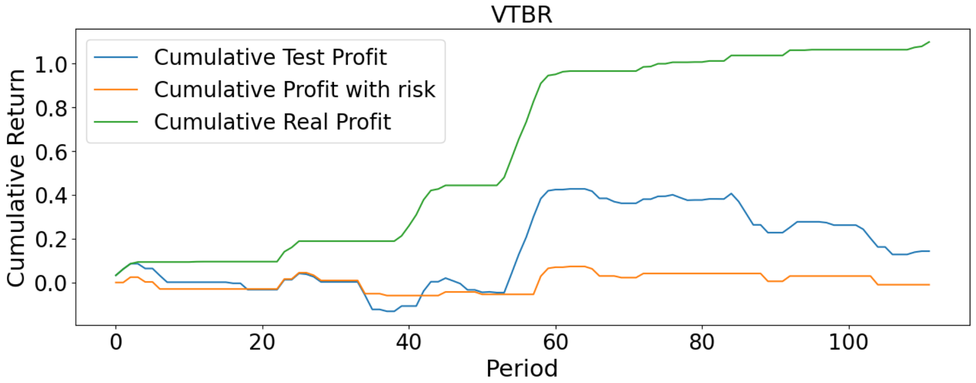
MTLR: 151,34% (without risk correction)

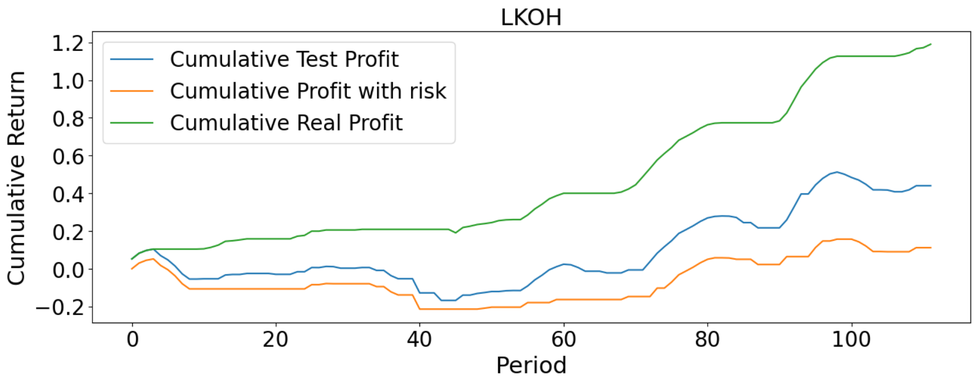
VTBR: 37,12% (without risk correction)

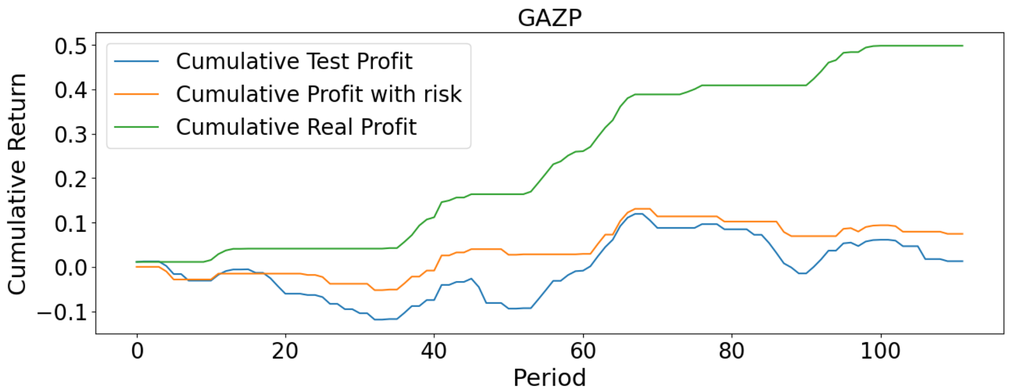
GAZP: 20,92% (without risk correction)

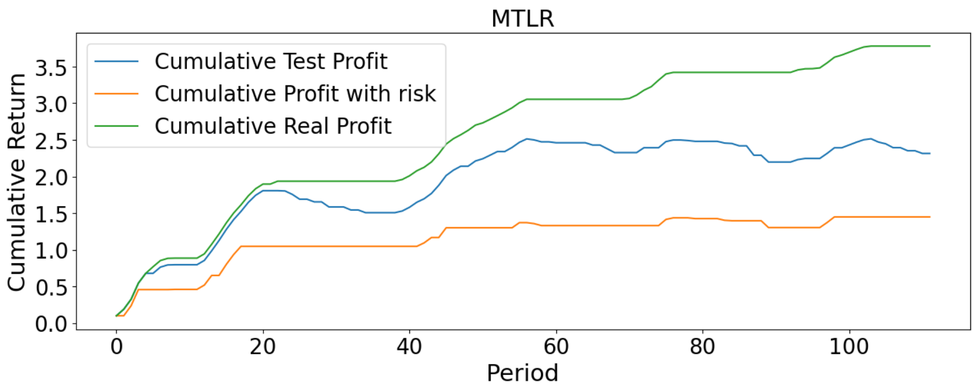
РЕЗУЛЬТАТЫ (ДОХОДНОСТЬ ТОРГОВЫХ СТРАТЕГИЙ – НА 5 ДНЕЙ ВПЕРЕД)











SBER: 71,71% (without risk correction)

LKOH: 37,27% (without risk correction)

MTLR: 247,51% (without risk correction)

VTBR: -0,07% (without risk correction)

GAZP: 8,92% (without risk correction)

ВЫВОДЫ

1. На более коротких временных горизонтах важно использовать эндогенные признаки (технические индикаторы).
2. При прогнозе движения цен на 5 дней вперед, сентимент уже не является шумом. Скорее всего, участники рынка закладывают в анализ настроение рынка на основе мнений на онлайн-платформах.
3. Прогнозирование волатильности в торговой симуляции не способствует увеличению доходности портфеля.

Ограничения

1) Анализируются только те компании, по которым идут обсуждения.
2) Есть смещения в классификации текстовых данных.
3) Не учитываются глобальные риски, например, связанные с COVID-19 и началом СВО.
4) Не учитываются Telegram каналы и чаты, в которых могут быть манипулятивные паттерны.