Исследование влияния тональности инвесторов на фундаментальные показатели компаний

Долаева Айшат Руслановна

Санкт-Петербургский государственный университет гр.21.М04-мм
Научный руководитель:
доцент кафедры информатики, к. ф.-м. н., Д. А. Григорьев

13 мая 2023 г.

Актуальность

Гипотеза эффективного рынка утверждает, что рыночная цена в любой момент времени отражает всю доступную информацию на рынке.
Тем не менее, гипотеза не объясняет нетрадиционные движения рынка, как кризисы и пузыри.
Внешние факторы могут влиять на поведение инвестора, который, в свою очередь, влияет на рынок, поэтому задача прогнозирования фондового рынка усложняется, и требуется найти альтернативные методы анализа.

Контекст задачи

Фондовые индексы сложно прогнозируемые, поскольку зависят от множества непрослеживаемых факторов.
Коэффициент Цена/Прибыль (P/E) измеряет текущую рыночную стоимость акций фирмы по отношению к её доходам и позволяет прогнозировать будущий рост стоимости акций.

Цель и задачи

Цель данной работы заключается в исследовании влияний фундаментальных и нефундаментальных показателей на коэффициент P/E.

- Сбор и обработка анализируемых данных: сообщений из Twitter'а и экономических показателей;
- Вычисление тональности текстов моделью нейронных сетей BERT и FinBERT;
- Анализ панельных данных и прогнозирование соотношения P/E.

Описание данных

Фундаментальные показатели	Нефундаментальные показатели
• Коэффициент Цена/Прибыль (P/E) • 30-дневная волотильность (VLT) • Ежедневная доходность акций (RTN) • Оборачиваемость активов (TATy) • Средний 30-дневный объём торгов (VLM) • Процентная ставка (IR3) • Рост прибыли на акцию (GROWTH) • Рыночная капитализация (SIZE) • Соотношение Цена/Балансовая стоимость (P/B) • Финансовый рычаг (FLy)	• Опрос Американской ассоциации индивидуальных инвесторов (IS). • Индекс потребительского доверия (CCI) • Тональность сообщений пользователей социальной сети «Twitter»

Система сбора данных

12 экономических данные были собраны на сайте Y-charts.
Для сбора сообщений из микроблога Twitter была применена библиотека с открытым исходным кодом «bsi-sentiment» на языке Python.
Для 8 компаний были собраны не более 500 твитов за почти каждый день с 1 января 2007 по 31 декабря 2021. Для каждой компании набралось около двух миллионов сообщений.

Предварительная обработка текстов

1. Удаление дублированных сообщений по имени пользователя и тексту.
2. Удаление лишних символов URL-адресов, псевдонимов пользователей.
3. Преобразование сокращений в полные слова.
4. Изменение эмотиконов и эмодзи в текстовое обозначение.
5. Замена хештегов в слова.
6. Токенизация данных – деление текстов на отдельные слова. Формирование списка отдельных слов для каждого твита.
7. Удаление сообщений с количеством слов меньше трёх.

№	Начальный текст	Обработанный текст
1	@united wow you even answered back! Awesome! @AmericanAir @USAirways That's customer service!!! #usairwaysfail	wow you even answered back! awesome! that is customer service!!! us airways fail
2	@SouthwestAir I ❤️ you! The only airline that understands us military families and our unpredictable changes. Pound it 👊	i love you! the airline that understands us military families and our unpredictable changes. pound it hit

Рис. 1: Пример обработки текстов

BERT и FinBERT

Двунаправленная нейронная сеть-кодировщик (BERT) предназначена для решения задач обработки естественного языка.
BERT был обучен на больших объемах текста из различных источников, включая тексты из Википедии, новостных статей, книг и других текстовых данных.
FinBERT был обучен на финансовых текстах, таких как новостные статьи, отчёты компаний и другие материалы из финансовой сферы.
FinBERT имеет оптимизированную архитектуру, что позволяет увеличить скорость работы и эффективность модели в решении задач, связанных с финансами.

LSTM

LSTM (Long Short-Term Memory) - это тип рекуррентной нейронной сети (RNN), который был разработан для обработки последовательностей данных с долговременными зависимостями между элементами последовательности.
Модель может эффективно обрабатывать последовательности с долгосрочными зависимостями, например анализ временных рядов.

Результаты тестирования BERT и FinBERT

Метрика	Финансовая тематика				Общая тематика			
	BERT	prepBERT	FinBERT	prepFinBERT	BERT	prepBERT	FinBERT	prepFinBERT
Precision	0.989	0.989	0.994	0.993	0.960	0.958	0.937	0.940
Recall	0.595	0.587	0.997	0.998	0.716	0.797	0.811	0.819
Accuracy	0.721	0.716	0.993	0.994	0.768	0.824	0.820	0.827
F1 Score	0.743	0.737	0.995	0.996	0.821	0.870	0.869	0.875
MCC	0.553	0.547	0.985	0.986	0.558	0.629	0.600	0.615

BERT и FinBERT – результаты прогнозирования необработанных текстов, prepBERT и prepFinBERT – метрики для обработанных текстов.

Модели панельных данных

Позволяют учитывать индивидуальные различия между экономическими единицами:

$$y_{it} = \alpha_i + x_{it}\beta + \varepsilon_{it}, \tag{1}$$

где α_i выражает индивидуальный эффект объекта i, не зависящий от времени t, y_{it} – зависимая переменная для экономической единицы i в момент времени t,
x_{it} – набор объясняющих (независимых) переменных (вектор размерности k),
ε_{it} – соответствующая ошибка $i = 1, \ldots, n, t = 1, \ldots, T$.

Виды моделей панельных данных

Fixed effect model

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it}, \tag{2}$$

Random effect model

$$y_{it} = x_{it}\beta + \mu + (u_i + \varepsilon_{it}), \tag{3}$$

где α_i выражает индивидуальный эффект объекта i, не зависящий от времени t, μ – общий параметр для всех единиц во все моменты времени, а u_i – ошибки, некоррелированные с ε_{it} разных i.

Пример панельных данных

company	date	finsent	BERT_sent	FLq	E_GR	PB	RTN	SIZE	TATy	VLT	VLM	CCIy	IR3m	IS	PE
apple	2008-01-01	0.570097	0.299311	0.000	72.52	8.438677	-1.425161	1.418026e+11	1.2190	43.702903	4.107710e+07	63.951613	2.810645	-30.419677	35.389677
	2008-02-01	0.604797	0.305886	0.000	72.52	6.549034	-0.324138	1.100493e+11	1.2190	53.337586	5.304448e+07	66.300000	2.202759	-11.665517	27.459655
	2008-03-01	0.588075	0.297320	0.000	72.52	6.798871	0.647419	1.150929e+11	1.2190	45.788065	4.532419e+07	66.300000	1.305484	-19.586774	28.529032
	2008-04-01	0.598860	0.329331	0.000	72.52	7.761067	1.213333	1.401130e+11	1.2190	42.981333	3.863600e+07	66.300000	1.315333	7.302667	32.770333
	2008-05-01	0.623649	0.332967	0.000	72.52	9.039710	0.673871	1.631945e+11	1.2190	37.531935	3.531710e+07	66.300000	1.768065	10.526452	38.167419
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
qualcomm	2021-08-01	0.900850	0.305439	1.582	74.12	20.122581	0.063226	1.645526e+11	0.8737	25.306129	7.168032e+06	77.600000	0.054516	4.621290	18.234839
	2021-09-01	0.912215	0.354815	1.582	74.12	18.915667	-0.496667	1.555480e+11	0.8737	19.451333	6.711633e+06	77.600000	0.043667	-5.631333	17.256333
	2021-10-01	0.863064	0.333417	1.390	74.12	14.640323	0.019032	1.456758e+11	0.8737	19.327742	7.240774e+06	77.600000	0.052903	7.167742	16.457742
	2021-11-01	0.842436	0.295088	1.390	74.12	19.182000	1.409667	1.908640e+11	0.8737	43.597000	1.048587e+07	77.600000	0.052667	6.853667	21.653333
	2021-12-01	0.856220	0.333890	1.390	74.12	20.325806	0.259355	2.038529e+11	0.8737	46.800323	1.251942e+07	77.600000	0.058065	-3.509677	22.888065

672 rows × 14 columns

Рис. 2: Показатели сектора "Электроника"

Гипотезы

Тестирование для проверки отсутствия индивидуальности фиксированных эффектов осуществляется с помощью F-теста.
$H_0 : \beta_1 = \beta_2 = ...\beta_k$ – отсутствие индивидуальности фиксированных эффектов отвергается.

Влияние независимых переменных проводилось с помощью t-теста.
$H_0 : \beta_k = 0$ – гипотеза об отсутствии значимого отклонения от нуля.
$H_0 : \beta_k \neq 0$ – значимое отклонение от нуля.

Таблица корреляции для сектора «Электроника»

В этот сектор были включены данные «Apple Inc.», «NVIDIA Corporation», «International Business Machines Corporation», «Qualcomm Inc.» с 2008 года по 2021.

	BERT	FinBERT	FLy	GROWTH	PB	RTN	SIZE	TATy	VLT	VLM	CCIy	IR3m	IS	PE
BERT	1.000***	0.160***	0.010	0.010	0.100*	0.060	0.250***	0.150***	-0.120**	-0.270***	0.250***	0.200***	0.050	0.020
FinBERT	0.160***	1.000***	-0.060	0.030	-0.090*	-0.030	-0.590***	-0.330***	0.060	-0.580***	0.090*	-0.080*	0.030	0.150***
FLy	0.010	-0.060	1.000***	-0.160***	0.080	-0.010	0.060	-0.230***	-0.090*	-0.170***	0.240***	0.320***	0.060	-0.180***
GROWTH	0.010	0.030	-0.160***	1.000***	0.180***	0.130***	0.040	0.390***	0.210***	0.290***	-0.160***	-0.140***	0.070	0.360***
PB	0.100*	-0.090*	0.080	0.180***	1.000***	0.060	0.350***	0.280***	0.130**	0.160***	-0.020	0.190***	-0.040	0.350***
RTN	0.060	-0.030	-0.010	0.130***	0.060	1.000***	0.090*	0.120**	0.040	0.050	-0.060	-0.090*	0.140***	0.160***
SIZE	0.250***	-0.590***	0.060	0.040	0.350***	0.090*	1.000***	0.250***	-0.270***	0.350***	0.110**	0.130***	0.040	0.090*
TATy	0.150***	-0.330***	-0.230***	0.390***	0.280***	0.120**	0.250***	1.000***	0.130***	0.310***	-0.230***	-0.130**	0.020	0.020
VLT	-0.120**	0.060	-0.090*	0.210***	0.130**	0.040	-0.270***	0.130***	1.000***	0.330***	-0.210***	0.030	-0.290***	0.300***
VLM	-0.270***	-0.580***	-0.170***	0.290***	0.160***	0.050	0.350***	0.310***	0.330***	1.000***	-0.110**	-0.000	-0.070	0.160***
CCIy	0.250***	0.090*	0.240***	-0.160***	-0.020	-0.060	0.110**	-0.230***	-0.210***	-0.110**	1.000***	0.350***	0.070	-0.080*
IR3m	0.200***	-0.080*	0.320***	-0.140***	0.190***	-0.090*	0.130***	-0.130**	0.030	-0.000	0.350***	1.000***	-0.040	0.030
IS	0.050	0.030	0.060	0.070	-0.040	0.140***	0.040	0.020	-0.290***	-0.070	0.070	-0.040	1.000***	0.030
PE	0.020	0.150***	-0.180***	0.360***	0.350***	0.160***	0.090*	0.020	0.300***	0.160***	-0.080*	0.030	0.030	1.000***

Результаты вычислений для сектора «Электроника»

	Случайные эффекты			Фиксированные эффекты		
const	-176.5700***	-8.3056	(-218.32; -134.82)	-158.87***	-7.7678	(-199.04; -118.71)
FinBERT	-18.0470**	-2.6559	(-31.391; -4.7024)	-18.047***	-2.6559	(-31.391; -4.702)
FLy	-0.3820**	-2.7847	(-0.6522; -0.1127)	-0.3824***	-2.7847	(-0.652; -0.112)
GROWTH	0.0910***	9.5195	(0.0727; 0.1105)	0.0916***	9.5195	(0.072; 0.110)
PB	0.1627***	3.3996	(0.0687; 0.2567)	0.1627***	3.3996	(0.068; 0.256)
RTN	1.4970*	2.1674	(0.1406; 2.8533)	1.4969*	2.1674	(0.1406; 2.853)
SIZE	11.2890***	22.896	(10.320; 12.257)	11.289***	22.896	(10.320; 12.257)
TATy	-25.4260***	-9.6627	(-30.593; -20.258)	-25.426***	-9.6627	(-30.593; -20.258)
VLT	0.0600	1.8829	(-0.0025; 0.1211)	0.0593	1.8829	(-0.003; 0.121)
VLM	-3.320***	-3.5695	(-5.1464; -1.4933)	-3.3199***	-3.5695	(-5.146; -1.493)
CCIy	-0.3010***	-7.3755	(-0.3812; -0.2209)	-0.3011***	-7.3755	(-0.381; -0.220)
IR3m	-1.040*	-1.9703	(-2.0929; -0.0034)	-1.0482*	-1.9703	(-2.092; -0.003)
IS	0.0009	0.0266	(-0.0594; 0.0610)	0.0008	0.0266	(-0.0594; 0.0610)

R^2 составляет 0.64 при анализе для модели с фиксированными эффектами или около 64% и 72% для модели со случайными эффектами. Оценка F-тест равна 246.54 для модели с фиксированными эффектами с p-значением 0.000.

График автокорреляции («Электроника»)

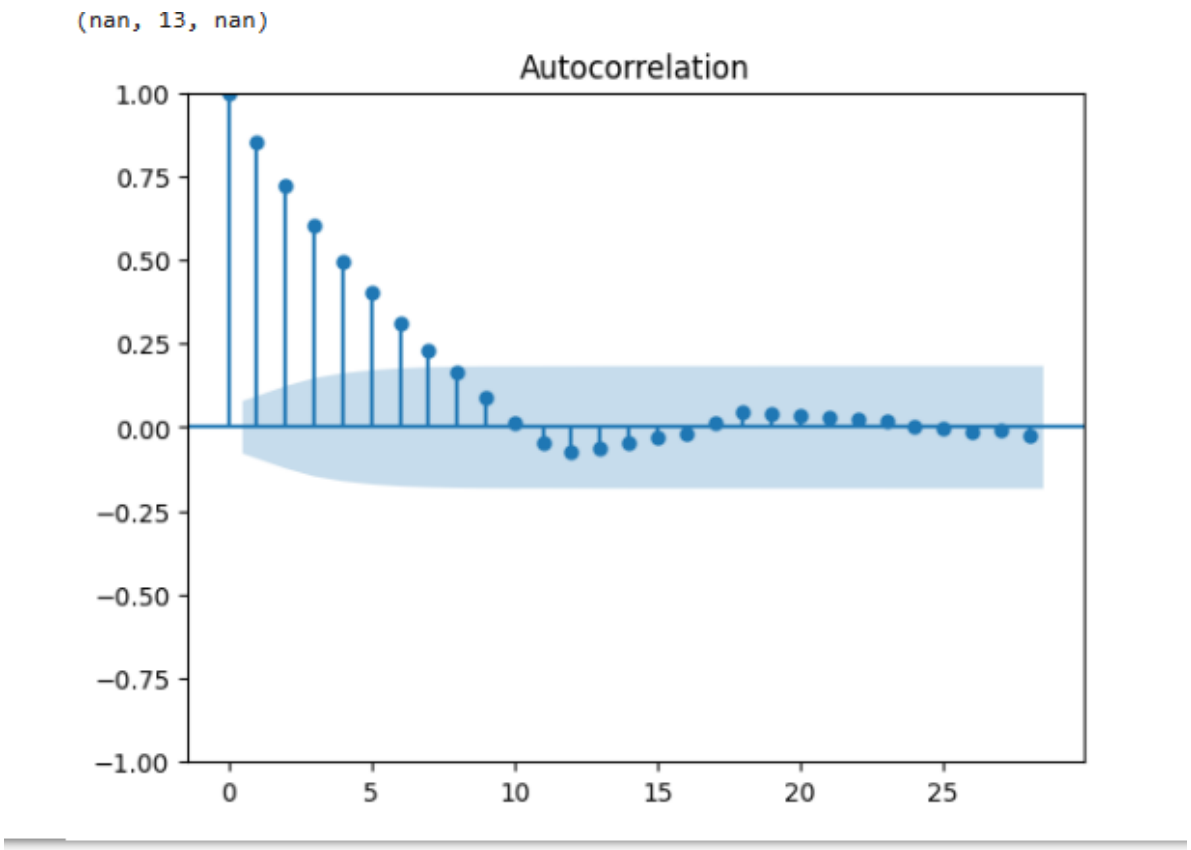


Рис. 3: График автокорреляции остаточных ошибок модели со случайными эффектами

Таблица корреляции для сектора «ПО»

В этот сектор были включены данные для компаний «Microsoft Corporation», «Adobe Inc.», «Salesforce.com Inc.», «ServiceNow Inc.» с 2016 по 2020 год.

	FinBERT	BERT	FLy	GROWTH	PB	RTN	SIZE	TATy	VLT	VLM	IR3m	IS	PE
FinBERT	1.000***	0.850***	-0.340***	0.330***	-0.110	-0.020	-0.410***	-0.180**	-0.030	-0.300***	0.050	0.070	0.080
BERT	0.850***	1.000***	-0.520***	0.330***	-0.280***	0.000	-0.160*	-0.340***	-0.040	-0.130*	0.130	0.050	0.010
FLy	-0.340***	-0.520***	1.000***	-0.270***	0.480***	-0.020	-0.140*	0.410***	-0.070	0.040	-0.120	0.060	-0.080
GROWTH	0.330***	0.330***	-0.270***	1.000***	-0.240***	0.030	0.090	-0.430***	0.280***	0.080	-0.220***	-0.150*	-0.020
PB	-0.110	-0.280***	0.480***	-0.240***	1.000***	0.030	-0.560***	0.870***	0.200**	-0.600***	0.050	0.030	0.230***
RTN	-0.020	0.000	-0.020	0.030	0.030	1.000***	0.050	-0.030	-0.040	-0.010	-0.070	0.110	-0.070
SIZE	-0.410***	-0.160*	-0.140*	0.090	-0.560***	0.050	1.000***	-0.680***	-0.130*	0.870***	0.080	-0.030	-0.080
TATy	-0.180**	-0.340***	0.410***	-0.430***	0.870***	-0.030	-0.680***	1.000***	0.160*	-0.660***	0.020	0.040	0.170**
VLT	-0.030	-0.040	-0.070	0.280***	0.200**	-0.040	-0.130*	0.160*	1.000***	-0.070	-0.170**	-0.370***	-0.000
VLM	-0.300***	-0.130*	0.040	0.080	-0.600***	-0.010	0.870***	-0.660***	-0.070	1.000***	-0.010	-0.060	-0.130
IR3m	0.050	0.130	-0.120	-0.220***	0.050	-0.070	0.080	0.020	-0.170**	-0.010	1.000***	0.270***	0.120
IS	0.070	0.050	0.060	-0.150*	0.030	0.110	-0.030	0.040	-0.370***	-0.060	0.270***	1.000***	-0.050
PE	0.080	0.010	-0.080	-0.020	0.230***	-0.070	-0.080	0.170**	-0.000	-0.130	0.120	-0.050	1.000***

Результаты вычислений для сектора «ПО»

	Случайные эффекты			Фиксированные эффекты		
	Parameter	t-stat	(Lower CI; Upper CI)	Parameter	t-stat	(Lower CI; Upper CI)
const	1.129e+04	0.4429	(-3.894e+04; 6.152e+04)	1.176e+04	0.4515	(-3.958e+04; 6.311e+04)
FinBERT	1.834e+04**	2.7252	(5078.0; 3.16e+04)	1.834e+04**	2.7252	(5078.0; 3.16e+04)
FLy	-4380.6***	-4.5094	(-6294.9; -2466.3)	-4380.6***	-4.5094	(-6294.9; -2466.3)
GROWTH	0.5140	1.2169	(-0.3183; 1.3462)	0.5140	1.2169	(-0.3183; 1.3462)
PB	182.31**	2.5978	(44.016; 320.60)	182.31**	2.5978	(44.016; 320.60)
RTN	-551.15	-1.2736	(-1403.9; 301.60)	-551.15	-1.2736	(-1403.9; 301.60)
SIZE	-913.71	-1.4648	(-2142.9; 315.47)	-913.71	-1.4648	(-2142.9; 315.47)
TATy	3782.7	0.6452	(-7770.2; 1.534e+04)	3782.7	0.6452	(-7770.2; 1.534e+04)
VLT	-7.6000	-0.3711	(-47.952; 32.752)	-7.6000	-0.3711	(-47.952; 32.752)
VLM	-311.12	-0.2942	(-2394.9; 1772.6)	-311.12	-0.2942	(-2394.9; 1772.6)
IR3m	222.19	0.9114	(-258.24; 702.62)	222.19	0.9114	(-258.24; 702.62)
IS	-24.346	-1.4077	(-58.428; 9.7357)	-24.346	-1.4077	(-58.428; 9.7357)

R^2 составляет 0.16 при анализе моделью с фиксированными эффектами или около 16% и 19% с моделью со случайными эффектами. Значение F-теста равно 4.02 для модели с фиксированными эффектами с p-значением 0.01.

График автокорреляции («ПО»)

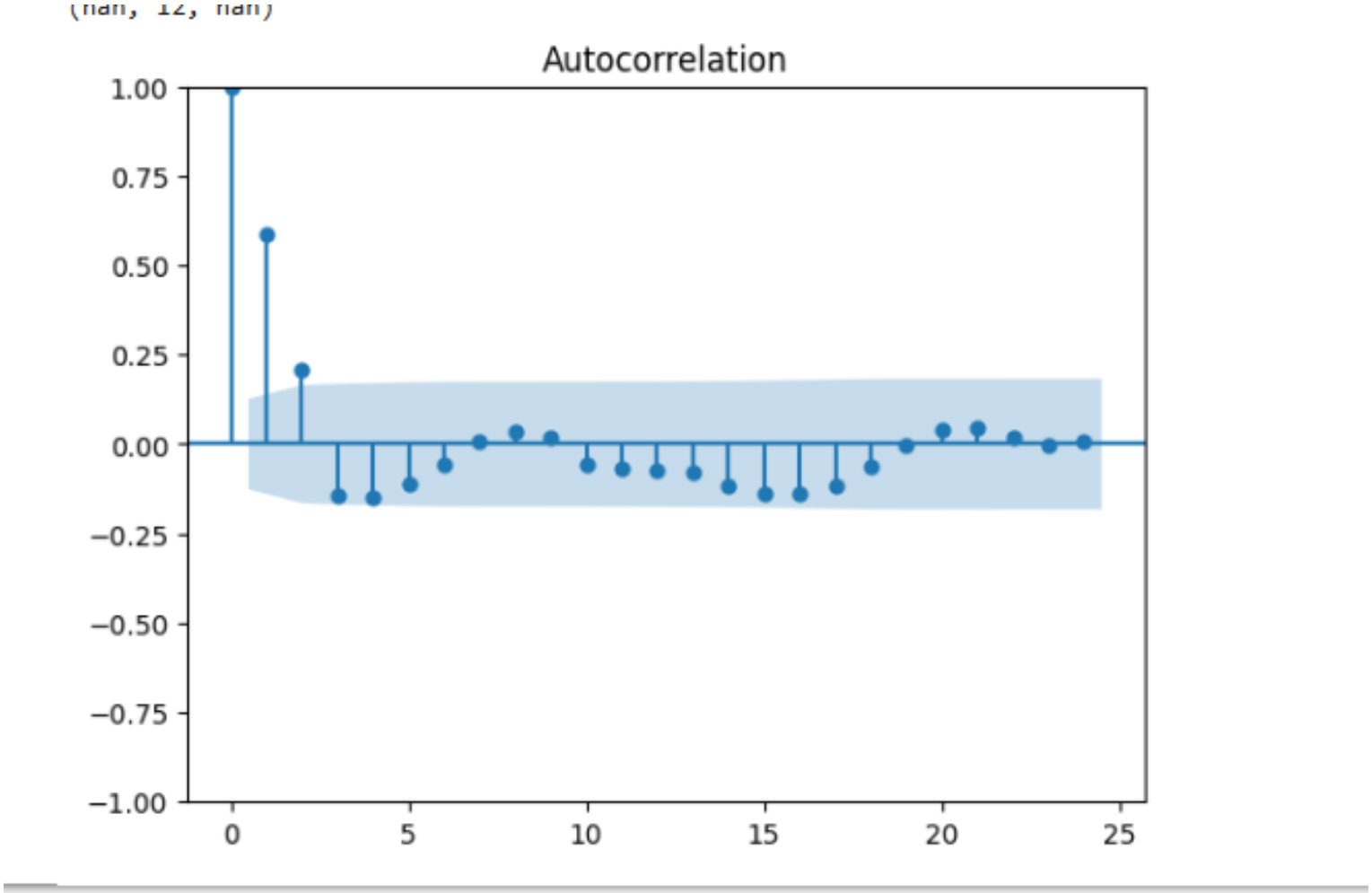
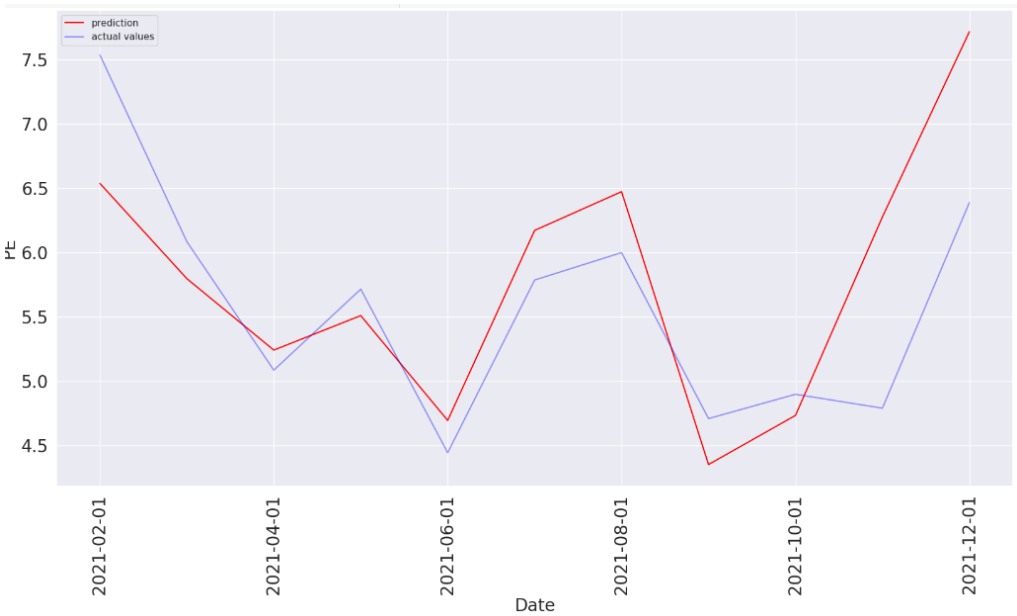
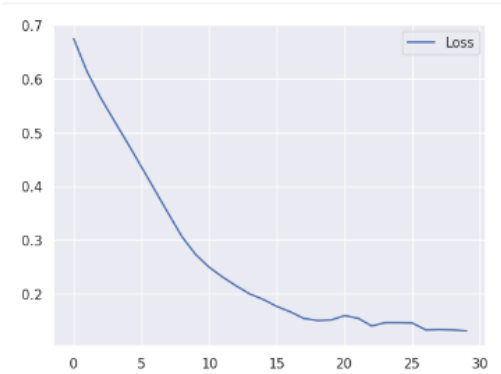


Рис. 4: График автокорреляции остаточных ошибок модели со случайными эффектами

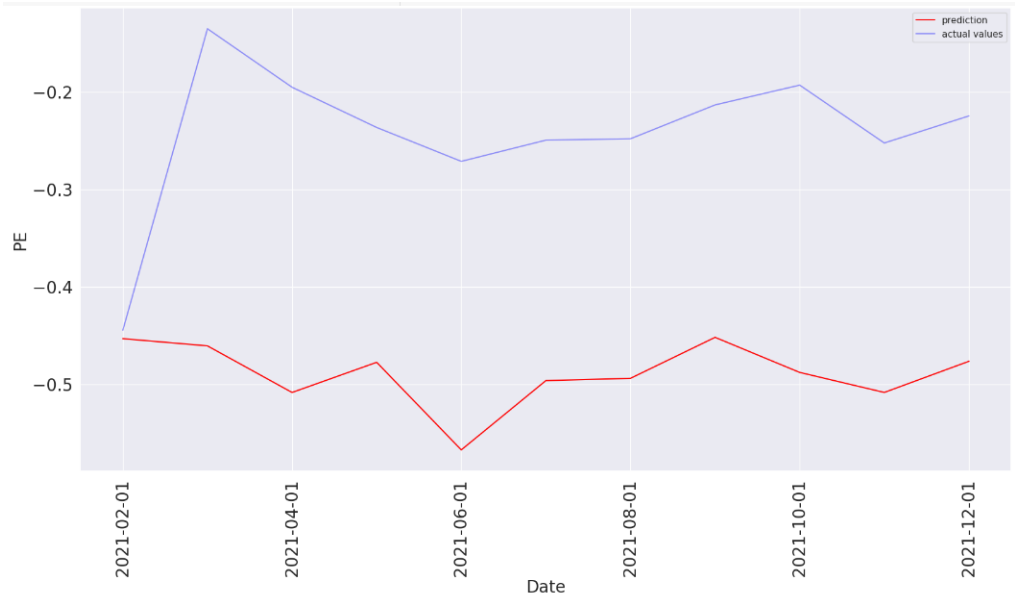
Анализ предсказательной способности модели для сектора «Электроники»

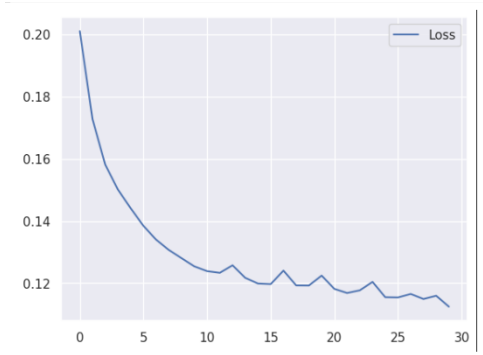




Результат предсказания P/E на 2021 год моделью LSTM и график потерь. $RMSE = 1.14, R^2 = 0.92$.

Анализ предсказательной способности модели для индустрии «ПО»





Результат предсказания P/E на 2021 год моделью LSTM и график потерь. $RMSE = 0.09, R^2 = 0.12$.

Результаты

1. Была изучена литература, связанная с исследованием фондового рынка и используемых в этой работе методов.
2. Собраны и обработаны экономические показатели и сообщения из социальной сети «Twitter» размером около двух миллионов для каждой компании.
3. Проведен анализ тональности сообщений пользователей моделями BERT и FinBERT.
4. Выявлены влияния фундаментальных экономических и нефундаментальных показателей на P/E с помощью моделей панельных данных для двух разных секторов.
5. Проведен анализ предсказательной способности модели на данных P/E 2021 года.

Выводы

1. Алгоритм FinBert лучше предсказывает тональность текстов, чем Bert.
2. Учет фундаментальных и нефундаментальных показателей повышает качество предсказания изменения показателя P/E.
3. Общественное мнение, измеренное тональностью сообщений из микроблога «Twitter», оказывает влияние на сектор электроники и показатель P/E. Влияние на сектор программного обеспечения обнаружено не было.
4. Алгоритм LSTM предсказывает изменение P/E на год с высокой точностью для сектора электроники.

Исследование влияния тональности инвесторов на фундаментальные показатели компаний

Долаева Айшат Руслановна

Санкт-Петербургский государственный университет гр.21.М04-мм
Научный руководитель:
доцент кафедры информатики, к. ф.-м. н., Д. А. Григорьев

13 мая 2023 г.