Санкт-Петербургский государственный университет

Кафедра информатики

Группа 21.М04-мм

Исследование влияния тональности инвесторов на фундаментальные показатели компаний

ДОЛАЕВА Айшат Руслановна

Отчёт по преддипломной практике
в форме «Эксперимент»

Научный руководитель:
доцент кафедры информатики, к. ф.-м. н., Д. А. Григорьев

Санкт-Петербург
2023

Оглавление

Введение

Прогнозирование фондового рынка остается популярной нерешенной исследовательской проблемой, не существует истинной экономической модели предсказания будущих цен акций.

На эффективном рынке цены на акции будут определяться главным образом фундаментальными факторами, такими как прибыль на акцию, дивиденды на акцию, коэффициент выплат, размер фирмы и т. д. Для прогнозирования будущих цен на акции фундаментальные аналитики используют коэффициенты оценки акций для получения текущей справедливой стоимости акций и прогноза будущей стоимости.

Один из таких мультипликаторов – коэффициент цена/прибыль (price-earnings ratio, P/E), который измеряет текущую рыночную стоимость акций фирмы по отношению к ее доходам. Соотношение P/E определяет ряд важных экономических характеристик компаний, например, прогнозирование будущего роста, поскольку низкий коэффициент P/E предполагает больший рост чистой прибыли в течение нескольких последующих финансовых периодов, тогда как компании с более высоким коэффициентом P/E ожидают меньшего увеличения прибыли. Таким образом P/E характеризуется как индикатор будущей эффективности инвестиций в ценные бумаги и показывает, сколько инвесторы готовы платить за доллар прибыли от акций.

Одной из задач в этом исследовании является определение факторов, влияющих на изменение соотношения P/E, с целью прогнозирования фондового рынка. Поведение инвесторов может быть нерациональным и гипотеза эффективного рынка не может объяснить нетрадиционные движения рынка, как кризисы и пузыри. Настроение инвесторов рассматривается, как один из факторов, влияющих на движение цен на фондвом рынке. Внешние ненаблюдаемые факторы могут влиять на поведение инвестора, принимающего решение инвестирования.

В этой работе в качестве измерения настроения инвесторов принимается тональность сообщений пользователей в социальной сети Twitter. Значимым преимуществом этой социальной сети является большое ко-

личество публикуемых сообщений в реальном времени, что позволяет анализировать общественное мнение большого круга пользователей практически на любую тему.

Для анализа тональности сообщений пользователей Twitter применялись, хорошо зарекомендовавших себя, модели обученных нейронных сетей BERT и FinBERT, отличающихся в первую очередь набором обучающих данных. FinBERT был дообучен на задачах, специфичных для финансовой сферы. Это позволяет FinBERT более эффективно обрабатывать финансовые данные и решать задачи, связанные с анализом финансовых рынков.

Были также отобраны ряд фундаментальных и нефундаментальных показателей, влияние которых было установлено в ряде изученнных исследований.

Собранные панельные данные представляют собой набор фундаментальных финансовых показателей и нефундаментальных оценок настроения инвестров с 2008 по 2021 год для выбранных восьми компаний, разделенных на два сектора: электроника и программное обеспечение.

Для выявления влияния настроения инвесторов на коэффициент P/E применялись регрессионные модели панельных данных с фиксированными и со случайными эффектами.

Также было проведено прогнозирование изменения соотношения P/E на 2021 год по прошлогодним данным моделью рекуррентных нейронных сетей LSTM.

Полученные результаты для двух секторов разняться. Влияние настроения инвесторов было вывлено для индустрии электроники и не было выявлено для компаний, определённых как программного обеспечения.

1. Постановка задачи

- Сбор фундаментальных экономических показателей и сообщений социальной сети Twitter для проведения исследования;
- Вычисление тональности текстов двунаправленными нейронными сетями BERT и FinBERT;
- Обработка данных сответственно их свойствам и условиям задачи;
- Прогнозирование соотношение P/E регрессионными моделями панельных данных с фиксированными и со случайными эффектами и анализование влияния тональности инвестров на коэффициент P/E;
- Анализ предсказательной способности модели на данных P/E 2021 года.

2. Обзор

В условиях совершенного рынка соотношение P/E описывается моделью Гордона-Шапиро и определяется постоянным коэффициентом выплаты дивидендов, постоянным ростом прибыли на акцию и постоянной безрисковой ставкой [4]. Помимо перечисленных факторов, ряд исследований выделяют другие фундаментальные показатели такие, как рыночная капитализация [1, 7, 17], прогнозируемый темп роста прибыли акций [1, 4, 7, 19], соотношение цена/балансовая стоимость [10], ежедневная доходность акций [8, 20, 13, 19, 21], оборачиваемость активов [15, 19], волатильность [1, 2, 19, 21], объём торгов [1, 13, 20, 21], финансовый рычаг [7, 10]. Но стандартная финансовая модель, основанная на анализе исторических фундаментальных показателей, не способна описать экономические потрясения на рынке.

Гипотеза эффективного рынка утверждает, что рыночная цена в любой момент времени отражает всю доступную информацию на рынке, включая экономические показатели и газетные публикации. Он предполагает, что инвесторы не склонны к риску и рациональны, и что даже если некоторые инвесторы вызывают шоки спроса или предложения, ведя иррациональную торговлю, рациональные инвесторы, или арбитражёры сглаживают эти шоки, чтобы вернуть стоимости ценных бумаг к их фундаментальным уровням. Тем не менее, гипотеза не объясняет непредсказуемые движения рынка [23]. Поведенческие теории обосновывают неожиданные измения на рынке тем, что инвесторы могут формировать ошибочные убеждения, и, следовательно, неправильно оценивать стоимость активов, вызывая отклонения цен активов от их внутренней стоимости. Внешние факторы могут влиять на поведение инвестора, который, в свою очередь, влияет на рынок [2, 8, 12, 15, 20]. В литературе не существует универсального определения для настроения инвесторов. Настроение может быть описано, как формирование у инвесторов представлений о будущих денежных потоках и инвестиционных рисках, которые не подтверждаются существующими фактами [1, 2, 12, 15, 17, 19, 20]. Существует различные подходы к измере-

нию настроения инвесторов. За косвенные показатели настроения инвесторов принимаются некоторые экономические характеристики, например, доходность, объём торгов и волатильность. Также процентная ставка рассматриваться как мера безрисковой ставки [1, 7, 19]. Прямыми методами являются опросы такие, как индекс потребительского доверия и опрос Американской ассоциации индивидуальных инвесторов [1, 2, 15, 19, 23].
Помимо перечисленных параметров, также для оценки настроений инвесторов может рассматриваться анализ сообщений новостных блогов или социальных сетей [13, 17].
Этот выбор опирается на концепцию «Мудрость толпы», которая утверждает, что агрегирование информации, предоставленной большим количеством людьми, часто приводит к более верным прогнозам, чем прогнозы, сделанные любым отдельным членом группы или даже экспертами [3, 11, 12, 14, 20, 21].

Twitter – сервис микроблогов, созданный в 2006 году, с кратким форматом сообщений (до 280 символов) и простотой поиска информации. Социальная сеть позволяет своим пользователям делать короткие посты и транслировать информацию в режиме реального времени, что делает социальную сеть доступным средством для сбора и анализа ежедневного общественного мнения. Однако вопрос о том, полезна ли конкретная информация о фирме из Twitter для прогнозирования прибыли фирмы и доходности её акций, так и не был решен.

В некоторых исследованиях взаимосвязь между изменением показателей на фондовом рынке и настроениями пользователей сети не была выявлена [8], в других же демострируется влияние на экономические факторы с течением времени. В небольшом количество работ настроение сообщений из социальных сетей используют для предсказания движения цен [3, 11, 12, 13, 17].

3. Данные

Доступное количество данных было собраны на промежутке с 1 января 2007 по 31 декабря 2021 для восьми компаний почти за каждый день.

3.1. Отрасли и компании

Для исследования влияния тональности инвесторов на фундаментальные показатели компаний, были отобранны восемь крупных американских компаний, разделённые на два сектора. В первый сектор «Электроника, полупроводники, информационные технологии» были помещены «Apple Inc.», «NVIDIA Corporation», «International Business Machines Corporation», «Qualcomm Inc.». Второй сектор «Программное обеспечение» состоит из «Microsoft Corporation», «Adobe Inc.», «Salesforce Inc.», «ServiceNow Inc.».

При отборе компаний было важно учитывать популярность, количество упоминаний в социальных сетях и продолжительность существования на фондовом рынке. Важной характеристикой также является их принадлежность к капиталистическому рынку и, следовательно, влияние на них рыночных механизмов [13]. Разделение компаний на сектора может оказывать влияние на прогнозирование движения цен фондового рынка, при схожести компаний между собой, но также может не оказывать никакого влияния.

3.2. Фундаментальные показатели

Выбор экономических факторов для проведения анализа – это один из сложных вопросов исследования фондового рынка.

Показатели были отобраны в соответствии с результатами эмпирических исследований и гипотез изученной литературой [1, 2, 4, 7, 8, 10, 13, 15, 17, 19, 20, 21].

Экономические данные были собраны на финансовой платформе Y-charts [22], который предоставляет практически любой экономический

показатель.

- Коэффициент цена/прибыль (Price to Earnings, P/E) представляет собой отношение цены закрытия акции к годовой разводненной (пониженной) прибыли на акцию, полученной фирмой в расчёте на одну акцию. Отношение P/E показывает текущий спрос инвесторов на акции компании.
  Несмотря на то, что математически возможно иметь отрицательное отношение PE, оно неприменимо при оценке акций, поскольку минимальная стоимость акции равна нулю. Отсутствующие значения приравнены к нулю.

- Волатильность (30-Day Rolling Volatility, VLT) – показатель, которым характеризуют изменчивость цены и используется как мера рискованности ценной бумаги. В этом исследовании выражается как стандартное отклонение последних 30 процентных изменений общей цены возврата, умноженный на квадратный корень из 252.

- Ежедневная доходность акций (1 Day Returns, RTN) определяется изменением стоимости активов за определенный период времени. Доходность предоставляет полезную информацию о вероятностном распределении цен на активы.

- Оборачиваемость активов (Total Asset Turnover, TATy) рассчитывает общий доход, полученный на каждый доллар активов, которыми владеет компания. Увеличение использования активов означает, что компания работает более эффективно с каждым долларом активов, которые у нее есть. Отсутствующие значения приравнены к нулю.

- Объём торгов (30-Day Average Daily Volume, VLM) – это среднее количество акций, торгуемых каждый день в течение торговой сессии. Объем можно использовать для измерения ликвидности акций и, следовательно, объяснить сечение ожидаемой доходности.

- Процентная ставка (Interest rate, IR3) – трехмесячных казначейских векселей, доход, полученный от инвестирования в выпущенные государством казначейские ценные бумаги со сроком погашения 3 месяца. Увеличение процентной ставки последовательно снижает чистую прибыль компании и распределение, которое она выплачивает акционерам.

- Рост прибыли на акцию (Earnings growth, GROWTH) – это годовой темп изменения чистой прибыли. Темпы роста EPS помогают инвесторам определить акции, доходность которых увеличивается или уменьшается. Отсутствующие значения приравнены к нулю.

- Рыночная капитализация (Market Capitalization, SIZE) — это измерение стоимости бизнеса, основанное на цене акций и количестве акций в обращении. Размер компании должен быть одним из основных движущих факторов коэффициента P/E.

- Соотношение цена/балансовая стоимость (Price-book ratio, P/B) – это финансовый коэффициент, используемый для сравнения балансовой стоимости компании с ее текущей рыночной ценой. Отношение цены к балансовой стоимости может выступать как индикатор ожидаемой будущей рентабельности собственного капитала. Отсутствующие значения дополнены линейной интерполяцией.

- Финансовый рычаг (Financial Leverage, FLy) – это сумма обязательств, которые компания использует для финансирования своих активов. Высокий финансовый рычаг приводит к высокой требуемой норме прибыли и более низкому соотношению P/E. Отсутствующие значения приравнены к нулю.

3.3. Нефундаментальные показатели

- Опрос Американской ассоциации индивидуальных инвесторов (American Association of Individual Investors survey, IS) – еженедельное наблюдение настроений, которое измеряет процент ин-

весторов, которые настроены медвежьи, бычьи и нейтрально на фондовом рынке в течение следующих шести месяцев.

- Индекс потребительского доверия (University of Michigan Consumer Sentiment Index, CCI) [24] Индекс потребительских настроений Мичиганского университета – ежемесячно проводимый опрос с не менее 500 телефонными интервью в выборке из США и пятьдесятью основными вопросами.

- Тональность сообщений пользователей социальной сети Twitter. Данные твитов состоят из псевдонима пользователя, текста и даты её публикации. Для каждой компании было собрано около двух миллинов сообщений, по не более 500 за один наблюдаемый день.

  Значения тональности текстов были вычисленны двумя алгоритмами BERT и FinBERT И нормализованы как отношение количества сообщений с положительным окрасом к общему количеству твитов в день.

3.4. Панельные данные

Собираемая информация в этом исследовании представляет собой трехмерный набор данных, в котором каждая строка – это уникальная единица, каждый столбец содержит данные одной из измеряемых переменных для этой единицы, а ось z содержит последовательность периодов времени, в течение которых блок был отслежен. Тем самым панельная совокупность данных представляет собой пространственную выборку объектов, наблюдаемых в течение некоторого периода времени.

Панель данных называется сбалансированной или несбалансированной панелью в зависимости от того, все ли единицы отслеживаются в течение одинакового количества периодов времени.

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. 1: Панель данных индустрии «Электроника».

На рисунке 1 представлен пример панельных данных сектора «Электроника».

4. Метод

Прогнозирование фундаментальных показателей компаний проводилось в несколько этапов:

1. Сбор всех доступных экономических данных из финансового сервиса Y-charts и сообщений пользователей из микроблога Twitter с использованием библиотеки с открытым исходным кодом «bsi-sentiment».

2. Обработка сообщений пользователей для повышения качества анализа тональности. Заполнение отсутствующих экономических данных.

3. Проведение анализа настроений пользователей Twitter с примением обученных нейронных сетей BERT и FinBERT.

4. Формирование панельных данных, делением компаний по схожести секторов.

5. Прогнозирование соотношения P/E моделью с фиксированными эффектами и моделью со случайными эффектами.

6. Предсказание коэффициента P/E моделью LSTM.

4.1. Обработка текстов

Сообщения пользователей социальной сети Twitter содержат шумы, незначащую информацию, поэтому для улучшения анализа тональности и повышения производительности вычислений данные предварительно нужно обработать, основываясь на особенностях сообщений [11, 14, 18]. Этапы преобразования текстов состоят из:

1. Удаления дублированных сообщений по имени пользователя и тексту, так как скоре всего это спам, не обладающий прогнозируемой особенностью.

2. Удаления лишних символов, URL-адресов, псевдонимов пользователей.

3. Преобразования сокращений в полные слова. Короткий формат сообщений является причиной большого количества аббревиатур, сокращений и ошибок.

4. Изменения эмотиконов и эмодзи в текстовое обозначение с помощью составленного словаря на основе частоты и значения использования того или иного изображения.

5. Замены хештегов в слова, удалением знака решетки "#" и расшифровкой слов.

6. Токенизации данных – тексты делятся на отдельные слова. Формирование списка отдельных слов для каждого твита.

7. Удаления сообщений с количеством слов меньше трёх.

4.2. Методы анализа тональности

Анализ тональности сообщений социальной сети «Twitter», в которых содержится упоминание одной из изучаемых компаний проводилось моделями нейронных сетей BERT [6] и FinBERT [5].

BERT (Bidirectional Encoder Representations from Transformers) – это модель обработки естественного языка (NLP), разработанная Google. Она представляет собой двунаправленную модель архитектуры трансформеров, обученную на больших объемах текста.

BERT стал одной из наиболее популярных моделей в обработке естественного языка благодаря своей способности эффективно обрабатывать контекст и учитывать зависимости между словами в тексте, что приводит к более точным результатам в различных задачах NLP.

FinBERT также основан на архитектуре transformer. Однако, есть несколько ключевых различий между ними.

Первое и наиболее важное отличие между BERT и FinBERT заключается в их наборе данных для обучения. BERT был обучен на боль-

ших объемах текста из разных источников, в то время как FinBERT был обучен на финансовых текстах, таких как новостные статьи, отчеты компаний и другие материалы из финансовой сферы. Это позволяет FinBERT более точно отражать особенности финансового языка, такие как специализированные термины, сокращения и другие особенности. Это позволяет FinBERT более эффективно обрабатывать финансовые данные и решать задачи, связанные с анализом финансовых рынков.

Третье различие заключается в размере моделей. FinBERT имеет оптимизированную архитектуру, чтобы улучшить скорость работы и эффективность модели в решении задач, связанных с финансами.

4.3. Модели панельных данных

Наборы панельных данных возникают в результате лонгитюдных исследований, в которых исследуется влияние измеряемых факторов на одну или несколько переменных с течением времени таких, как ежегодные инвестиции компании [16, 25, 26].

Пусть y_{it} – зависимая переменная для экономической единицы i в момент времени t x_{it} – набор объясняющих (независимых) переменных (вектор размерности k и ε_{it} – соответствующая ошибка $i = 1, \ldots, n, t = 1, \ldots, T$.

Простейшая модель – это обычная модель регрессии, которая не учитывает панельную структуру данных (pooled model).

$$y_{it} = x'_{it}\beta + \varepsilon_{it} \tag{1}$$

При этом предполагается, что ε_{it} коррелированы между собой как по i так и по t, и некоррелированы со всеми объясняющими переменными x_{it}.

Модель позволяющая учитывать индивидуальные различия между экономическими единицами:

$$y_{it} = \alpha_i + x'_{it}\beta + \varepsilon_{it}, \tag{2}$$

где α_i выражает индивидуальный эффект объекта i не зависящий от

времени t.

В зависимости от предположений относительно характера величины α_i рассматриваются две модели.

Модель с фиксированными эффектами (FE)

Экономическая единица «уникальная» и не может рассматриваться как результат случайного исхода. Предполагается, что в уравнении (2) величина α_i представляет собой специфичный эффект единицы i, компенсирующий смещение, вызванное отсутствием ненаблюдаемых переменных. Тем самым модель с фиксированными эффектами обеспечивает гарантированное получение несмещенных и состоятельных оценок.
Для моделей крупных предприятий и компания, отраслей, регионов и стран чаще используется модель с фиксированными эффектами, так как каждый из объектов такой выборки обладает своими индивидуальными особенностями.

Модель со случайными эффектами (RE)

Предположим, что объекты попали в панель случайно, тогда специфичный эффект распределяется вокруг среднего значения в соответствии с неизвестным распределением вероятности.
В уравнении (2) $\alpha_i = \mu + u_i$, где μ – параметр общий для всех единиц во все моменты времени, а u_i – ошибки, некоррелированные с ε_{it} разных i Полагается, что пропущенные переменные являются одной из составляющих ошибок.

При анализе случайной выборки большого объёма для небольших фирм или предприятий, когда интересует поведение совокупности в целом, а не отдельных объектов, используют модель со случайными эффектами.

4.4. LSTM

Долгая краткосрочная память (Long Short-Term Memory, LSTM) [9] – это тип рекуррентной нейронной сети (RNN), который был разработан для обработки последовательностей данных с долговременными зависимостями между элементами последовательности.

LSTM имеет архитектуру, которая позволяет запоминать долгосрочные зависимости в последовательности данных и учитывать их при обработке последующих элементов. Ключевой компонент LSTM – это ячейка памяти, которая может хранить информацию о предыдущих элементах последовательности и использовать её при обработке следующих элементов.

LSTM может эффективно обрабатывать последовательности с долгосрочными зависимостями, например, анализ временных рядов.

В сравнении с обычными RNN, LSTM имеет ряд преимуществ, такие как более эффективную работу с долгосрочными зависимостями, возможность избежать проблемы затухания градиента и возможность сохранять информацию в ячейке памяти для использования в будущем.

5. Эксперимент

В этом разделе приведены гипотезы и результаты проведённого исследования влияния тональности инвесторов на фундаментальные показатели компаний.

5.1. Метрики

Тестирование для проверки отсутствия индивидуальности фиксированных эффектов осуществляется с помощью критерия Фишера.
$H_0 : \beta_1 = \beta_2 = ...\beta_k$ – отсутствие индивидуальности фиксированных эффектов отвергается.
Т-критерий Стюдента применялся для проверки влияния независимых переменных.
$H_0 : \beta_k = 0$ – гипотеза об отсутствии значимого отклонения от нуля.
$H_1 : \beta_k \neq 0$ – значимое отклонение от нуля.

5.2. Результаты

Прогнозирование тональности сообщений

Тональность инвесторов определялась анализом общественного мнения пользователей Twitter. Для этой задачи были собраны сообщения пользователей, упоминавщих одну из компаний в своих твитах. Всего сообщений было собрано почти по два миллиона твитов для каждой компании. Тексты сообщений содержат много шума, поэтому они были предварительно обработаны. Затем применялись обученные нейронные сети BERT и FinBERT для прогнозирования тональности сообщений.

Результат вычисления BERT – категориальный признак: «positive» или «negative» с числовой оценкой. Значения были переведены в числовой формат заменой «positive» на 1 и «negative» на 0.
FinBERT предоставляет результат как одно из значений: «negative», «positive» или «neutral» с числовой оценкой. Трёхмерная шкала была переведена в двухмерный по значению оценки: $score >= 0 : 1; score < 0 : 0$.

Затем полученные значения были нормализованы как отношение количества позитивных сообщений к общему числу за день.

Был проведён анализ вычислений двух алгоритмов, сравнение метрик представлен на таблице 1.

Таблица 1: Сравнение результатов вычисления двух алгоритмов.

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.

FinBert демонстрирует высокую прогностическую способность для соообщений, содержащих финансовую информацию и определяет тональность текстов с точностью 99%, что выше, чем для сообщений общей тематики – 82%. BERT хуже справляется с прогнозированием финансовой информации и немного уступает FinBert в предсказании сообщений, не связанных с экономикой. Результаты почти всегда лучше для обработанных текстов, чем необработанных.

Прогнозирование соотношения P/E

Рассмотрим пару примеров изменения коэффициента P/E и тональности сообщений в течение временного отрезка.

Рис. 2: График изменения значений P/E и настроения инвесторов для компании «Microsoft».

На рисунке 2 представлено ежедневные изменения показателя P/E и тональности инвесторов для компании «Microsoft», вычисленного моделью FinBERT. Часть пиков и спусков совпадает, а резкое падение тональности ознаменует снижение фундаментального показателя.

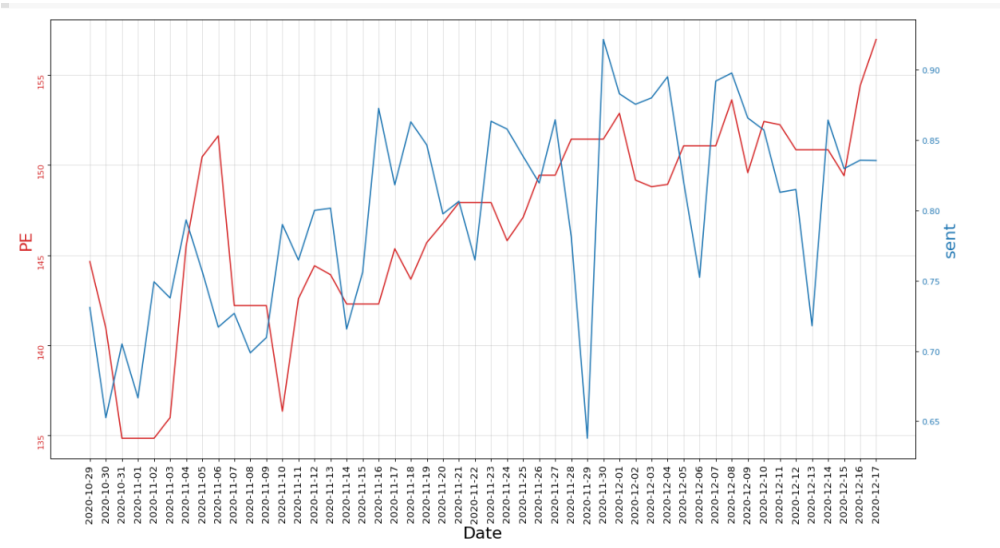

Рис. 3: График изменения значений P/E и настроения инвесторов для компании «ServiceNow».

Подъёмы и спуски схожи и для компании «ServiceNow» (рис. 3), но показатель настроения более изменчив.

Сектор «Электроника»

Компании были разделены на два сектора. Первый сектор был охарактеризован, как электроника, полупроводники, информационные технологии: «Apple Inc.», «NVIDIA Corporation», «International Business Machines Corporation», «Qualcomm Inc.». Ежемесячные экономические данные и тональности инвесторов, измеренные с 2009 по 2021 года, были сведены в панельные данные.

Для начала проанализируем описательную статистику индустрии для отобранных факторов.

Таблица 2: Описательная статистика для индустрии «Электроника».

Оценка	BERT	FinBERT	FLy	GROWTH	P/B	RTN	SIZE	TATy	VLT	VLM	CCIy	IR3m	IS	P/E
count	624.000	624.000	624.000	624.000	624.000	624.000	6.240000e+02	624.000	624.000	6.240000e+02	624.000	624.000	624.000	624.000
mean	0.344	0.775	1.377	21.541	8.826	0.123	2.794950e+11	0.728	29.266	1.697413e+07	83.919	0.489	4.245	22.021
std	0.083	0.089	2.855	41.539	9.352	0.503	4.215222e+11	0.202	14.600	1.772083e+07	9.879	0.735	11.837	15.678
min	0.144	0.560	0.000	-56.960	1.588	-1.856	4.310065e+09	0.358	8.476	2.660433e+06	66.832	0.012	-25.195	0.000
25%	0.282	0.704	0.052	-0.340	3.422	-0.147	8.486873e+10	0.604	19.270	7.209660e+06	77.600	0.051	-3.852	13.298
50%	0.343	0.801	0.488	12.770	5.749	0.116	1.274606e+11	0.723	25.803	1.186274e+07	81.600	0.117	4.623	17.395
75%	0.393	0.844	1.764	51.350	10.194	0.400	2.310090e+11	0.877	35.140	1.924290e+07	92.900	0.475	11.935	24.159
max	0.690	0.949	20.280	138.000	100.440	2.489	2.844323e+12	1.130	113.700	1.563481e+08	98.400	2.453	32.535	99.080

sd - среднеквадратическое отклонение.

Из таблицы 2 видно, что значения капитализации и объёма торгов очень отличаются от других данных, поэтому их следует сгладить с помощью логорифмирования. По отклонениям от среднего, минимальным и максимальным значениям видим, что данные некоторых показателей, как соотношение P/B и рост прибыли, могут сильно варьироваться.

Теперь проверим коррелированность данных.

Таблица 3: Таблица корреляции для индустрии «Электроника».

	BERT	FinBERT	FLy	GROWTH	P/B	RTN	SIZE	TATy	VLT	VLM	CCIy	IR3m	IS	P/E
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***
P/B	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
P/E	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***

Статистическая значимость – .05*, .01**, .001**.

Высокой статистически значимой корреляции среди данных между собой не наблюдается (таблица 3). Присутствует значимая, но невысокая корреляция между настроением инвесторов, вычисленным алгоритмом FinBERT и P/E. Соотношение P/E коррелирует почти со всеми факторами.

Проведём прогнозирование P/E двумя регрессионными алгоритмами, используя все факторы, и проанализируем результаты вычисления.

Таблица 4: Результаты регрессионного анализа с фактором FinBERT.

	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)
P/B	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)

Статистическая значимость – .05*, .01**, .001**. с 2009 по 2021 год.

Результаты прогнозирования данных с 2008 года по 2021 год выявили значимое влияние настроения инвесторов, вычисленного алгоритмом FinBERT, на соотношение P/E (таблица 4). При этом наблюдается значимое влияние ряда ключевых фундаментальных показателей, что согласуется к дркгими исследованиями.

Коэффициент детерминации (R^2), который измеряет долю общей дисперсии зависимой переменной, которая объясняется факторами после учета степеней свободы, потерянных из-за включения переменных регрессии, составляет 0.64 при анализе моделью FE или около 64% и 72% с моделью RE. Это говорит, что наши отобранные переменные неплохо объясняют изменение P/E, но всё же это немного.

Критерий Фишера или F-тест для регрессии, который измеряет совместную значимость параметров модели, дал значение статистики 246.54

моделью FE и с уровнем значимости p-value, равным 0.000. Это означает, что оценки коэффициентов модели являются совместно значимыми. Регрессионные модели лучше соответствуют данным, чем модель, основанная только на перехвате.

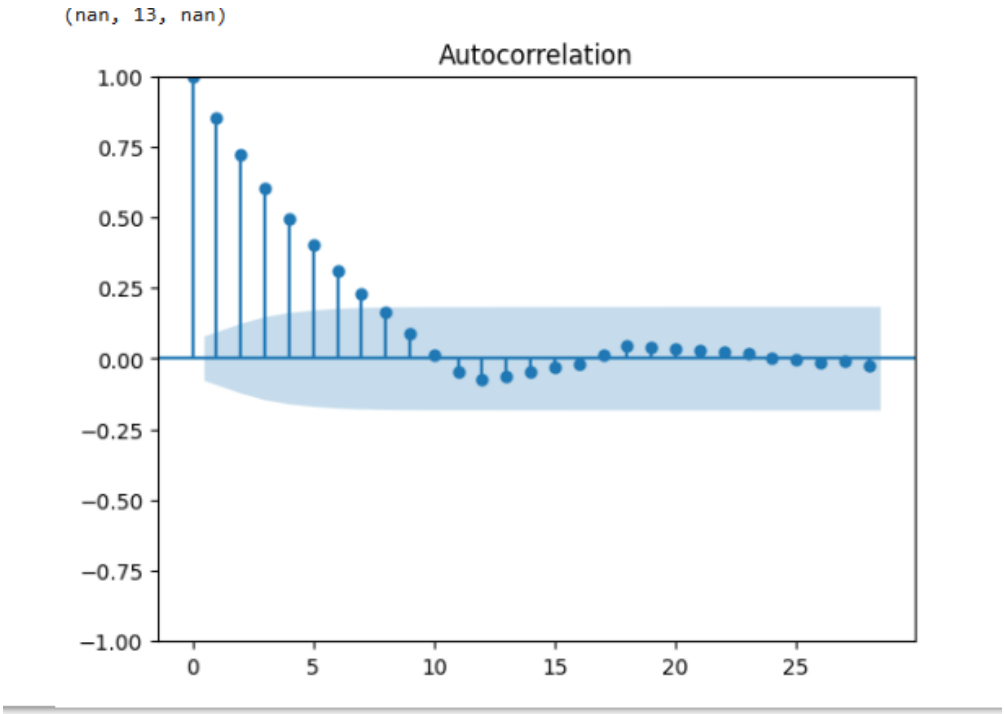

Рис. 4: График автокорреляции остаточных ошибок модели RE.

Можно наблюдать значительную автокорреляцию остаточных ошибок при лагах от одного до шести месяцев (рис. 4). То есть значения, наблюдаемые в прошлом на шестом лаге назад, влияют на текущее значение.

Автокоррелированные остаточные ошибки приводят к тому, что стандартные ошибки не указываются (занижаются), что приводит к завышению t-значений (или z-значений) и отсутствию доверительных интервалов оценок параметров. Коэффициенты, которые в действительности равны нулю, то есть незначительны, могут быть ошибочно указаны как ненулевые, то есть значимые.

Проведём анализ с фактором BERT и без фактора индекса потребительского доверия (CCI).

Таблица 5: Результаты регрессионного анализа с фактором BERT.

	Случайные эффекты			Фиксированные эффекты		
const	-221.13***	-10.514	(-262.43; -179.82)	-204.90***	-10.360	(-243.74; -166.06)
BERT	-9.7958	-1.5149	(-22.495; 2.9031)	-9.7958	-1.5149	(-22.495; 2.9031)
FLy	-0.4998***	-3.5748	(-0.7743; -0.2252)	-0.4998***	-3.5748	(-0.7743; -0.2252)
GROWTH	0.0908***	9.0315	(0.0710; 0.1105)	0.0908***	9.0315	(0.0710; 0.1105)
P/B	0.1854***	3.6407	(0.0854; 0.2854)	0.1854***	3.6407	(0.0854; 0.2854)
RTN	1.7687**	2.4365	(0.3431; 3.1943)	1.7687**	2.4365	(0.3431; 3.1943)
SIZE	10.722***	20.733	(9.7063; 11.737)	10.722***	20.733	(9.7063; 1.737)
TATy	-21.733***	-7.7457	(-27.244; -16.223)	-21.733***	-7.7457	(-27.244; -16.223)
VLT	0.1081**	3.2773	(0.0433; 0.1729)	0.1081***	3.2773	(0.0433; 0.1729)
VLM	-2.0477	-2.0848	(-3.9766; -0.1188)	-2.0477	-2.0848	(-3.9766; -0.1188)
IR3m	-1.7119**	-3.0887	(-2.8003; -0.6234)	-1.7119***	-3.0887	(-2.8003; -0.6234)
IS	0.0068	0.2088	(-0.0569; 0.0704)	0.0068	0.2088	(-0.0569; 0.0704)

Статистическая значимость – .05*, .01**, .001**. с 2009 по 2021 год.

Из таблицы 5 можно сделать вывод, что статистически значимого влияния настроения инвесторов, вычисленного алгоритмом BERT, на соотношение P/E обнаружено не было. При этом был убран показатель индекса потребительского доверия, тем самым влияние тональности инвесторов увеличилось, но не значимо. Возможно, это говорит, что не следует исключать общественное мнение при анализировании фондового рынка.
R^2составляет 0.60 при анализе моделью FE или около 60% и 70% с моделью RE. Это говорит, что наши отобранные переменные неплохо объясняют изменение PE, но это невысокий показатель.
Значение статистики F-тест равна 213.74 с p-значением 0.000, то есть оценки коэффициентов модели являются совместно значимыми.

Сектор «Программное обеспечение»

Второй сектор был обозначен как программное обеспечение и в него вошли следующие компании: «Microsoft Corporation», «Adobe Inc.», «Salesforce Inc.», «ServiceNow Inc.»
Ежемесячные экономические данные и оценка тональности инвесторов измеренны с 2008 по 2021 год.

Результаты прогнозирования данных с 2009 по 2021 год не выявили значимого влияния настроения инвесторов, вычисленного алгоритмом BERT, на соотношение P/E. При этом наблюдается значимое влияние ряда ключевых фундаментальных показателей.

R^2 равен составляет 0.10 при анализе при FE 10% и 12% при RE. Это говорит, что наши отобранные переменные плохо объясняют изменение P/E.

Статистика F-тест составляет 1.5298 с p-значением 0.2, это означает, что оценки коэффициентов модели являются совместно не значимыми.

Данные, собранные для компании «ServiceNow» неполные, так как эта компания была зарегистрирована на фондовой бирже позже остальных, в 2012 году. Поэтому проведём еще один анализ для этих данных на период с 2016 по 2020 год.

Проанализируем описательную статистику сектора для отобранных факторов.

Таблица 6: Описательная статистика для индустрии «Программное обеспечение».

Оценка	FinBERT	BERT	FLy	GROWTH	P/B	RTN	SIZE	TATy	VLT	VLM	IR3m	IS	P/E
count	240.000	240.000	240.000	240.000	240.000	240.000	2.400000e+02	240.000	240.000	2.400000e+02	240.000	240.000	240.000
mean	0.770	0.443	0.523	154.364	15.941	0.094	2.749199e+11	0.541	29.848	1.002283e+07	1.145	1.308	433.398
std	0.056	0.106	0.388	588.926	11.877	0.416	3.826446e+11	0.102	15.374	1.224899e+07	0.836	11.461	2940.181
min	0.640	0.190	0.064	-89.510	3.943	-1.587	8.599517e+09	0.350	8.967	1.191633e+06	0.089	-20.170	0.000
25%	0.730	0.376	0.255	0.000	7.643	-0.096	5.123917e+10	0.479	19.579	2.257902e+06	0.301	-6.853	25.340
50%	0.764	0.410	0.401	13.830	10.059	0.110	1.081600e+11	0.536	25.739	3.879129e+06	1.064	3.455	49.872
75%	0.810	0.514	0.705	72.970	19.834	0.348	2.765154e+11	0.613	36.111	1.279047e+07	1.938	8.017	96.720
max	0.886	0.680	1.507	2820.000	43.523	1.684	1.651548e+12	0.724	96.280	6.567467e+07	2.453	26.358	29269.806

sd - среднеквадратическое отклонение.

Значения капитализации и объёма торгов следует прологарифмировать (таблица 6). По отклонениям от среднего, минимальных и максимальных значений можно наблюдать, что данные очень разнообразные.

Высокой статистически значимой корреляции среди данных между собой не наблюдается (таблица 7). Корреляции между настроением инвесторов и P/E нет. Присутствует корреляция соотношения P/E и соотношения цены/балансовой стоимости, а также оборачиваемостью активов.

Таблица 7: Таблица корреляции для индустрии «Программное обеспечение».

	FinBERT	BERT	FLy	GROWTH	P/B	RTN	SIZE	TATy	VLT	VLM	IR3m	IS	P/E
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
P/B	-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
P/E	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***

Статистическая значимость – .05*, .01**, .001**.

Проведём прогнозирование P/E двумя регрессионными алгоритмами и проанализуем результаты вычисления.

Таблица 8: Результаты регрессионного анализа с фактором FinBERT.

	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)
P/B	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)

Статистическая значимость – .05*, .01**, .001**. С 2016 года по 2020 год.

Результаты прогнозирования данных с 2016 по 2020 год на таблице 8 демонстрируют значимое влияние настроения инвесторов, вычисленного алгоритмом FinBERT на соотношение P/E. При этом не наблюдается значимого влияния ключевых фундаментальных показателей, кроме P/B и финансового плеча.

$R^2 = 0.16$ при анализе моделью FE или около 16% и 19% моделью RE. Это говорит, что наши отобранные переменные плохо объясняют изменение P/E.

F-тест = 4.02 с p-значением 0.01, это означает, что оценки коэффициентов модели являются совместно значимыми.

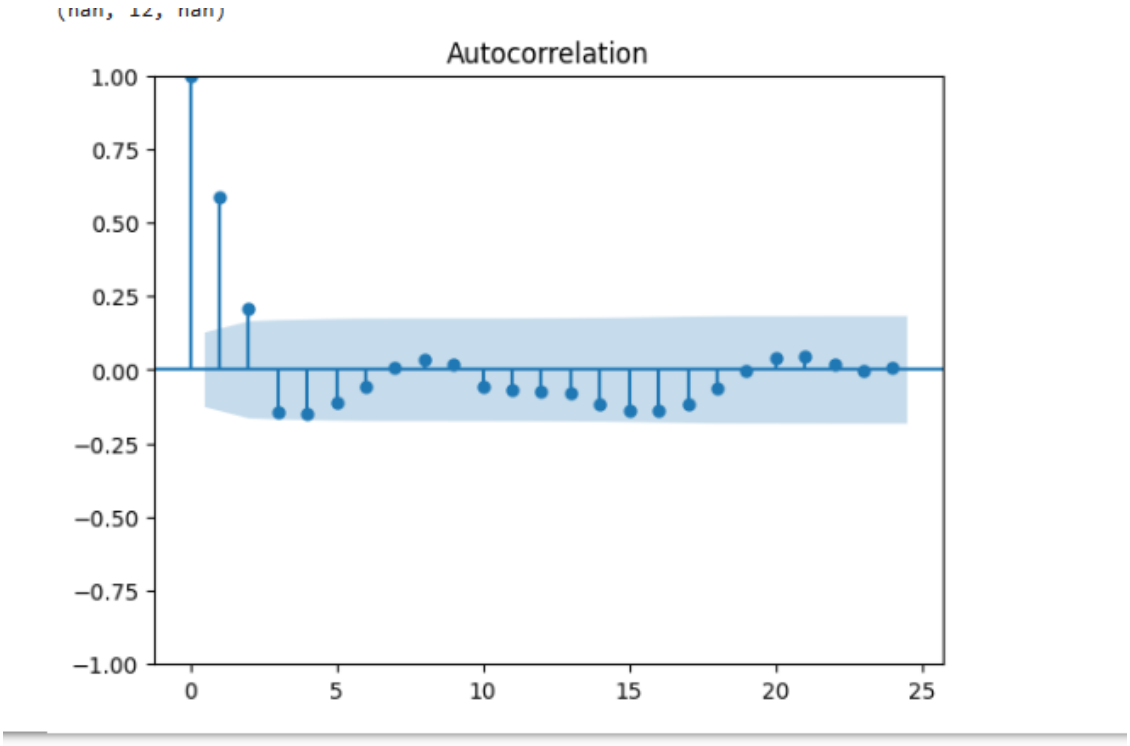

Рис. 5: График автокорреляции остаточных ошибок модели со случайными эффектами.

На рисунке 5 продемонстрирована автокорреляция остаточных ошибок модели RE на первом лаге. То есть значения данных прошлого месяца на текущее значение.

Таблица 9: Результаты регрессионного анализа с фактором BERT.

	Случайные эффекты			Фиксированные эффекты		
	Parameter	t-stat	(Lower CI; Upper CI)	Parameter	t-stat	(Lower CI; Upper CI)
const	3.594e+04	1.4973	(-1.136e+04; 8.324e+04)	3.68e+04	1.4985	(-1.159e+04; 8.52e+04)
BERT	3896.2	0.8232	(-5430.9; 1.322e+04)	3896.2	0.8232	(-5430.9; 1.322e+04)
FLy	-4474.1***	-4.5164	(-6426.1; -2522.0)	-4474.1***	-4.5164	(-6426.1; -2522.0)
GROWTH	0.5077	1.1765	(-0.3427; 1.3581)	0.5077	1.1765	(-0.3427; 1.3581)
P/B	177.27	2.2826	(24.234; 330.30)	177.27	2.2826	(24.234; 330.30)
RTN	-571.72	-1.2979	(-1439.7; 296.31)	-571.72	-1.2979	(-1439.7; 296.31)
SIZE	-1281.4	-2.0778	(-2496.7; -66.162)	-1281.4	-2.0778	(-2496.7; -66.162)
TATy	-1253.3	-0.2142	(-1.279e+04; 1.028e+04)	-1253.3	-0.2142	(-1.279e+04; 1.028e+04)
VLT	-11.492	-0.5544	(-52.338; 29.353)	-11.492	-0.5544	(-52.338; 29.353)
VLM	-332.93	-0.3101	(-2448.6; 1782.7)	-332.93	-0.3101	(-2448.6; 1782.7)
IR3m	260.55	1.0334	(-236.27; 757.36)	260.55	1.0334	(-236.27; 757.36)
IS	-21.984	-1.2541	(-56.527; 12.560)	-21.984	-1.2541	(-56.527; 12.560)

Статистическая значимость – .05*, .01**, .001**. с 2015 по 2020 год.

Фактор BERT не демонстрирует влияние на коэффициент P/E (таблица 9), так же как и другие факторы, кроме финансового рычага. Коэффициент R^2 равен 0.16 при FE или около 13% и 17% при RE. Это говорит, что наши отобранные переменные плохо объясняют изменение

P/E.

Статистика F-теста равена 1.6595 с p-значением 0.1766, это означает, что оценки коэффициентов модели скорее всего не являются совместно значимыми.

Предсказание соотношения P/E

Так как наши экономические данные очень разнообразные, для начала стоит нормализовать данные по тренировочному набору, затем добавить в данные лаг в один временной период назад, то есть обучение модели LSTM будет производиться по текущим значениям и по данным прошлого месяца. Производилось прогнозирование значений P/E для 2021 года по значениям прошлых лет.

time	var1(t-1)	var2(t-1)	var3(t-1)	var4(t-1)	var5(t-1)	var6(t-1)	var7(t-1)	var8(t-1)	var9(t-1)	var10(t-1)	...	var5(t)	var6(t)	var7(t)	var8(t)	var9(t)	var10(t)	var11(t)	var12(t)	var13(t)	var14(t)
1	-2.051396	-0.452193	1.302084	0.109693	-2.676006	-0.276897	2.324924	0.736878	1.697156	-1.755070	...	-0.721770	-0.380685	2.324924	1.298554	2.518826	-1.539092	2.038122	-1.144757	0	0.527014
2	-1.688447	-0.452193	1.302084	-0.124217	-0.721770	-0.380685	2.324924	1.298554	2.518826	-1.539092	...	1.002673	-0.364200	2.324924	0.858437	1.988758	-1.539092	0.901946	-1.782819	0	0.608044
3	-1.863351	-0.452193	1.302084	-0.093291	1.002673	-0.364200	2.324924	0.858437	1.988758	-1.539092	...	2.007128	-0.282419	2.324924	0.694812	1.529553	-1.539092	0.914418	0.383140	0	0.929419
4	-1.750548	-0.452193	1.302084	0.025815	2.007128	-0.282419	2.324924	0.694812	1.529553	-1.539092	...	1.049622	-0.206975	2.324924	0.377127	1.301680	-1.539092	1.487690	0.642817	0	1.338370
5	-1.491269	-0.452193	1.302084	0.184092	1.049622	-0.206975	2.324924	0.377127	1.301680	-1.539092	...	-1.181823	-0.227958	2.324924	0.078842	1.145955	-1.539092	1.655176	-1.489445	0	1.209500
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
151	1.548864	0.450531	0.163755	2.996563	1.426937	-0.397190	-0.246609	0.218633	-0.512233	-0.499842	...	1.025002	-0.324523	-0.246609	1.297040	-0.437960	-0.499842	-0.617556	-1.273947	3	2.051611
152	-0.126885	0.450531	0.163755	3.828997	1.025002	-0.324523	-0.246609	1.297040	-0.437960	-0.499842	...	-0.547583	-0.318455	-0.246609	0.785737	-0.455544	-0.499842	-0.611416	-1.583474	3	2.042948
153	1.212632	0.450531	0.163755	3.822870	-0.547583	-0.318455	-0.246609	0.785737	-0.455544	-0.499842	...	0.553788	-0.277628	0.657412	0.288345	-0.590401	-0.499842	-0.621641	-0.260440	3	0.540191
154	1.693420	0.099627	1.341160	1.948982	0.553788	-0.277628	0.657412	0.288345	-0.590401	-0.499842	...	1.275859	-0.212519	0.657412	0.867761	-0.577272	-0.499842	-0.635897	1.439372	3	0.834854
155	1.602860	0.099627	1.341160	2.354780	1.275859	-0.212519	0.657412	0.867761	-0.577272	-0.499842	...	0.123798	-0.182850	0.657412	0.742839	-0.507108	-0.499842	-0.637980	1.440995	3	0.939879

Рис. 6: Панельные данные индустрии «Электроника» с единичным лагом.

На рисунке 6 представлены преобразованные панельные данные сектора «Электроники». Помимо повторяющихся значений показателей с временой разницей в месяцах, данные также содержат категориальную переменную, определяющую компанию, и значение P/E в прошлом месяце.

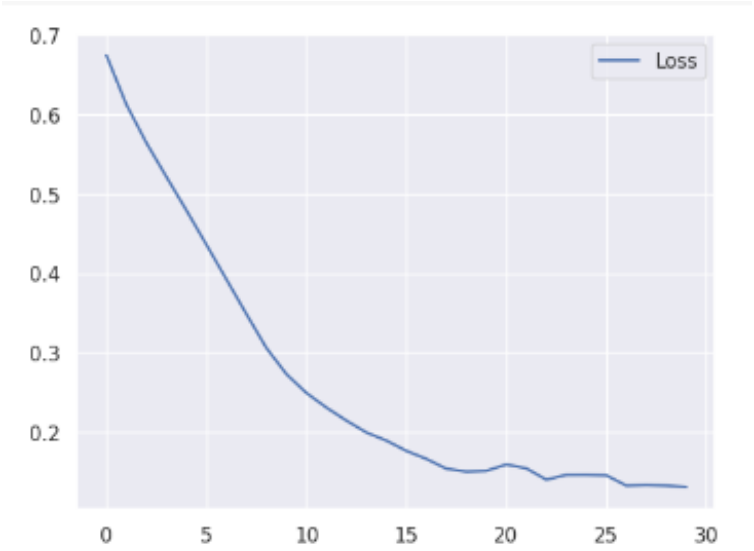

Рис. 7: Потери при обучении.

На рисунке 7 представлены потери на каждой эпохе обучения.

Рис. 8: Результат предсказания P/E.

На графике 8 продемонстрировано прогнозирование P/E экономическими факторами и показателем настроения инвесторов, вычисленным алгоритмом FinBERT. Модель LSTM предсказывает изменение P/E с высокой точностью. Среднеквадратичная ошибка (RMSE) составляет 1.14, а R^2 – 0.92.

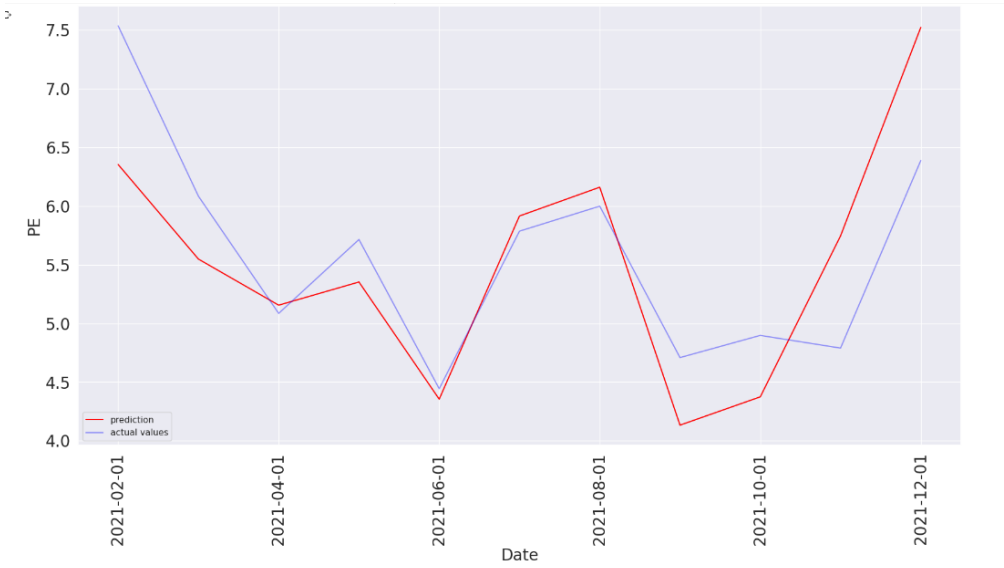

Рис. 9: Результат предсказания P/E.

На графике 9 продемонстрировано прогнозирование P/E моделью LSTM. Учитывались экономические факторы и показатель настроения инвесторов, вычисленный алгоритмом BERT. RMSE = 1.12, а R^2 – 0.94, точность предсказания высока.

Также проведем прогнозирование для сектора «Программное обеспечение».

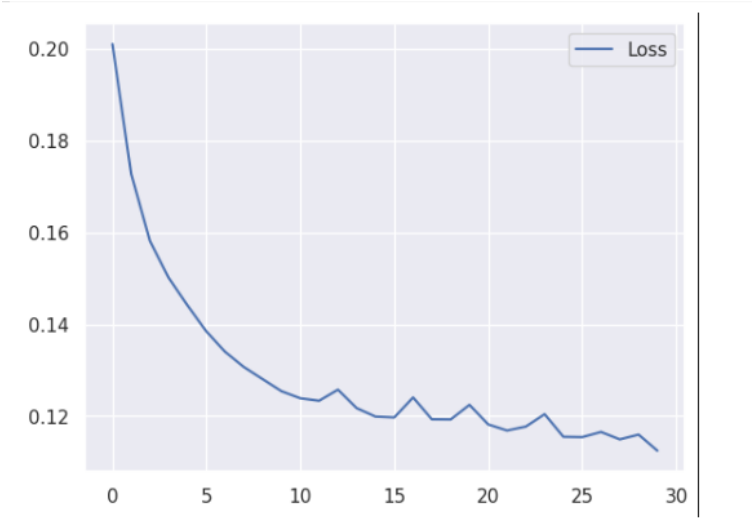

Рис. 10: Потери при обучении

На рисунке 10 представлены потери на каждой эпохе обучения.

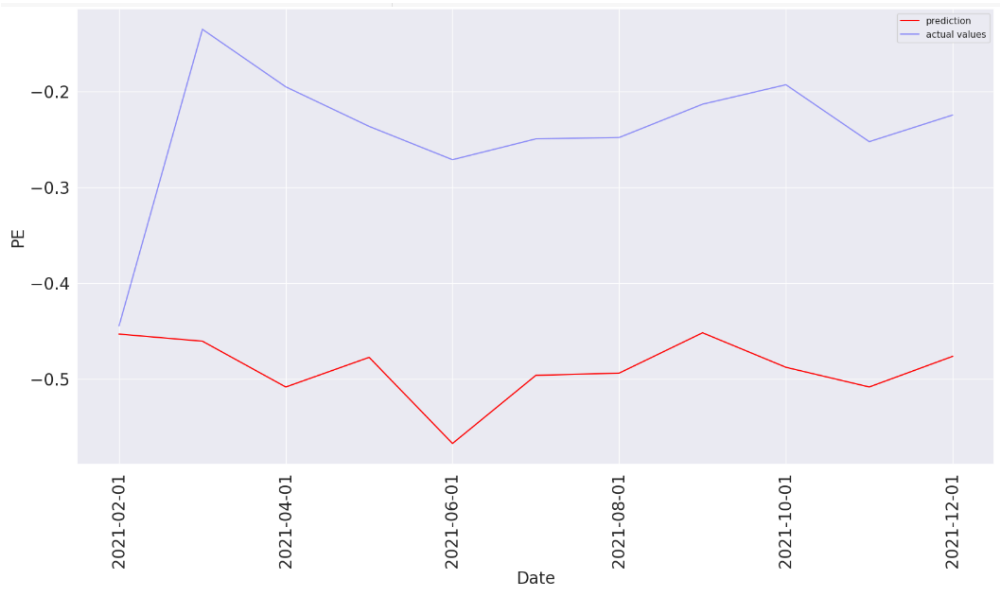

Рис. 11: Результат предсказания P/E.

На графике 11 продемонстрировано прогнозирование P/E экономическими факторами и показателем настроения инвестора, вычисленным алгоритмом FinBERT. RMSE составляет 0.09, а R^2 – 0.12.

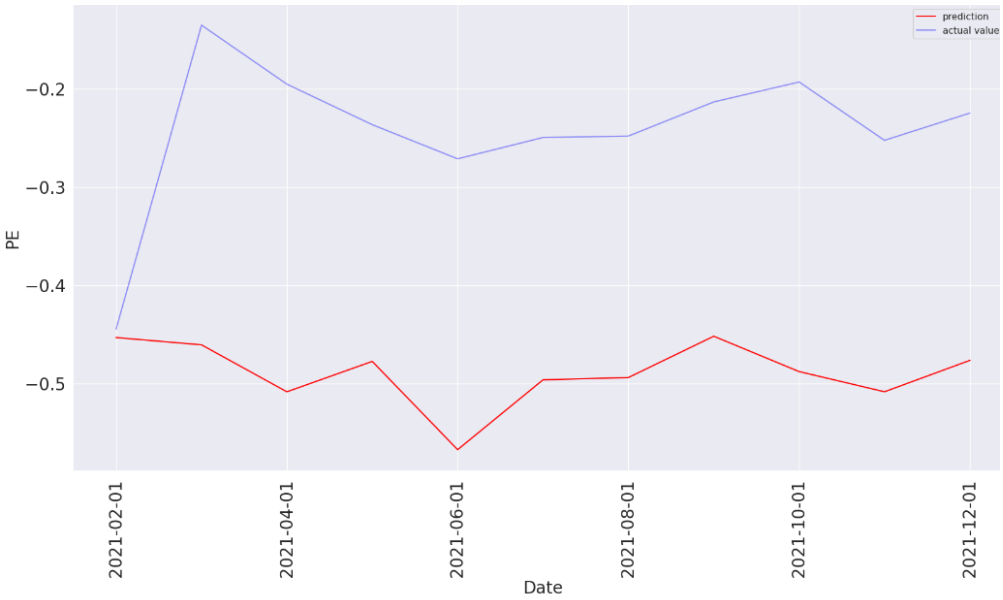

Рис. 12: Результат предсказания P/E.

На графике 12 продемонстрировано прогнозирование P/E экономическими факторами и показателем настроения инвестора, вычисленным алгоритмом BERT. Среднеквадратичная ошибка равна 0.09, а R^2 – 0.1.

Для сектора «Программное обеспечние» точность прогноза очень низкая. Как и в случае с прогнозированием регрессионными моделями, подобранные параметры плохо описывают изменение отношения P/E. Также причиной плохой прогнозируемости может быть некорректно выбранные компании.

Заключение

По итогам проведённого исследования были сделаны несколько выводов в отношении вычислений тональности текстов и влиянии настроения инвесторов на соотношение P/E.

Было продемонстрировано, что алгоритм FinBERT лучше предсказывает тональность текстов, чем модель BERT.

Учет фундаментальных и нефундаментальных показателей повышает качество предсказания изменения показателя P/E.

Было выявлно статистически значимое влияние общественного мнения, измеренное тональностью сообщений Twitter, на сектор электроники и на P/E.

Обоснованного влияния тональности инвесторов на сектор программного обеспечения выявлено не было.

Алгоритм LSTM предсказал изменение P/E на 2021 год с высокой точностью для сектор электроники.

Список литературы

[1] B. Jitmaneeroj. Does investor sentiment affect price-earnings ratios? // Studies in Economics and Finance. — 2017. — Vol. 34, no. 2. — P. 183–193. — URL: https://www.emerald.com/insight/content/doi/10.1108/SEF-09-2015-0229/full/html.

[2] Baker M. P. Wurgler J. Investor Sentiment in the Stock Market // Journal of Economic Perspectives. — 2007. — Vol. 21. — P. 29–151. — URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=962706.

[3] Bartov E. Faurel L. Mohanram P. Can Twitter Help Predict Firm-Level Earnings and Stock Returns? // Asset Price Forecasts. — 2017. — Vol. 93, no. 3. — 66 p. — URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2631421.

[4] Beaver W. Morse D. What Determines Price-Earnings Ratios? // Financial Analysts Journal. — 1978. — Vol. 34, no. 4. — P. 65–76. — URL: https://www.jstor.org/stable/4478160.

[5] D. Araci. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models // ArXiv. — 2019. — August. — 11 p. — URL: https://arxiv.org/abs/1908.10063.

[6] Devlin J. Chang M. Lee K. Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // ArXiv. — 2019. — May. — 16 p. — URL: https://arxiv.org/abs/1810.04805.

[7] Freihat AR Farah. Factors affecting price to earnings ratio (P/E): Evidence from the emerging market // Risk Governance and Control: Financial Markets Institutions. — 2019. — Vol. 9, no. 2. — P. 47–56. — URL: https://web.archive.org/web/20220319005736id_/https://virtusinterpress.org/IMG/pdf/rgcv9i2p.4.pdf.

[8] Gan B. Alexeev V. Bird R. Yeung D. Sensitivity to sentiment: News vs social media // International Review of Financial Analysis. — 2030. —

Vol. 67. — 55 p. — URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1468-036X.2013.12007.x.

[9] Hochreiter S. Schmidhuber J. «Long short-term memory». Neural Computation. — 1997. — November. — Vol. 9, no. 8. — P. 1735–1780. — URL: https://direct.mit.edu/neco/article-abstract/9/8/1735/6109/Long-Short-Term-Memory?redirectedFrom=fulltext.

[10] Itemgenova A. Sikveland M. The determinants of the price-earnings ratio in the Norwegian aquaculture industry // Journal of Commodity Markets. — 2020. — Vol. 17, no. C. — URL: https://doi.org/10.1016/j.jcomm.2019.04.001.

[11] Kordonis J. Symeonidis S. Arampatzis A. Stock Price Forecasting via Sentiment Analysis on Twitter // The 20th Panhellenic Conference on Informatics. — 2016. — Vol. 36. — P. 1–6. — URL: https://dl.acm.org/doi/10.1145/3003733.3003787.

[12] Li B. Chan K. C. Ou C. Sun R. Discovering public sentiment in social media for predicting stock movement of publicly listed companies // Information Systems. — 2017. — Vol. 69. — P. 81–92. — URL: https://www.sciencedirect.com/science/article/abs/pii/S0306437916304860?via%3Dihub.

[13] Oliveira N. Cortez P. Areal N. On the Predictability of Stock Market Behavior Using StockTwits Sentiment and Posting Volume // Portuguese Conference on Artificial Intelligence. — 2013. — Vol. 17. — P. 355–365. — URL: https://link.springer.com/chapter/10.1007/978-3-642-40669-0_31.

[14] Pagolu V. S. Challa K. Panda G. Majhi B. Sentiment analysis of Twitter data for predicting stock market movements // International Conference on Signal Processing, Communication, Power and Embedded System. — 2016. — October. — 6 p. — URL: https://ieeexplore.ieee.org/document/7955659.

[15] Pandey P. Sehgal S. Investor sentiment and its role in asset pricing: An empirical study for India // IIMB Management Review. — 2019. — Vol. 31, no. 2. — P. 127–144. — URL: https://www.sciencedirect.com/science/article/pii/S0970389619301594?via%3Dihub.

[16] Panel Data Analysis Fixed and Random Effects using Stata. — URL: www.princeton.edu/~otorres/Panel101.pdf (дата обращения: 2023-05-10).

[17] Piñeiro J. López-Cabarcos M. A. Pérez-Pico A. M. Examining the influence of stock market variables on microblogging sentiment // Journal of Business Research. — 2015. — Vol. 69, no. 6. — P. 2087–2092. — URL: https://www.sciencedirect.com/science/article/abs/pii/S0148296315006384?via%3Dihub.

[18] Pota M. Ventura M. Catelli R. Catelli M. Esposito M. An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian // Sensors. — 2020. — December. — Vol. 21, no. 1. — 21 p. — URL: https://www.mdpi.com/1424-8220/21/1/133.

[19] Rahman Md L. Abul Shamsuddin A. Investor sentiment and the price-earnings ratio in the G7 stock markets // Pacific-Basin Finance Journal. — 2019. — Vol. 55, no. C. — P. 46–62. — URL: https://www.sciencedirect.com/science/article/abs/pii/S0927538X18305560?via%3Dihub.

[20] Shen D. Urquhart A. Wang P. Does twitter predict Bitcoin? // Economics Letters. — 2019. — Vol. 174. — P. 118–122. — URL: https://www.sciencedirect.com/science/article/abs/pii/S0165176518304634?via%3Dihub.

[21] Sprenger T. Welpe I. Tweets and Trades: The Information Content of Stock Microblogs // Innovation Finance Accounting eJournal. — 2010. — Vol. 20, no. 5. — P. 926–957. — URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1468-036X.2013.12007.x.

[22] YCharts - Financial Research and Proposal Platform. — URL: https://ycharts.com/ (дата обращения: 2023-05-10).

[23] Zouaoui M. Nouyrigat G. Beer F. How Does Investor Sentiment Affect Stock Market Crises? Evidence from Panel Data // Financial Review. — 2011. — Vol. 46, no. 4. — P. 723–747. — URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6288.2011.00318.x.

[24] of Michigan University. Surveys of Consumers - Data. — URL: https://data.sca.isr.umich.edu/data-archive/mine.php (дата обращения: 2023-05-10).

[25] Любимцев О.В. Любимцева О.Л. Линейные регрессионные модели в эконометрике. Методическое пособие. — Нижний Новгород : ННГАСУ, 2016. — 45 р.

[26] Магнус Я.Р. Катышев П.К. Пересецкий А.А. Эконометрика. Начальный курс: Учеб. 6-е изд. — М. : Дело, 2004. — 576 р. — ISBN: 978-5-85006-296-5.