МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

имени М.В. Ломоносова

ЭКОНОМИЧЕСКИЙ ФАКУЛЬТЕТ

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

«НОВОСТНАЯ АНАЛИТИКА КАК ФАКТОР ВОЛАТИЛЬНОСТИ ФИНАНСОВЫХ ПОКАЗАТЕЛЕЙ»

Выполнил студент э412:

Гаврилов Вадим

Научный руководитель:

Доцент Рощина Янина Александровна

Аннотация

Работа сфокусирована на исследовании воздействия средств массовой информации на финансовый рынок РФ. Целью работы является определение степени и направления влияния тематических новостных потоков на российский фондовый рынок. В качестве инструмента анализа используются модифицированные модели обобщенной условной гетероскедастичности (GARCH) с добавлением во вспомогательную регрессию на условную дисперсию переменных, отвечающих за новостную аналитику. В качестве таких переменных используются переменные, отражающие количество новостей на определенную дату и тему. Для обработки новостей использовались методы машинного обучения. Были подтверждены гипотезы о значимом влиянии новостей на рынок в целом, о сохранении влияния на срок до пяти дней, а также то, что новостной поток оказывает влияние на цену акции конкретного ПАО вне зависимости от рода деятельности этого ПАО.

JEL-коды: C32, C59, Z23

Введение

Актуальность. В эпоху цифровизации на людей с каждым годом оказывается все большее информационное давление. Ежедневно, ежечасно мы получаем из окружающего мира много разной информации, которая влияет на все сферы нашей жизни. Данная работа посвящена изучению воздействия новостных потоков на одну из таких сфер – на финансовый рынок, а также на характеристики его функционирования. Как показывает практика, если у инвестора в инструментарии заложен учет новостной составляющей, то это приводит к повышению качества принимаемых им инвестиционных решений. Развитие новостной аналитики и ее приложений в финансах получает все большее признание в международном инвестиционном сообществе. Увеличивается число исследований на данную тематику, разрабатываются платформы и программное обеспечение, применяются новые математические методы, которые служат инструментами на финансовом рынке (Ager, Hafez, 2011). А также, как можно убедиться в Главе 1 данной работы, новые математические методы применяются и для анализа влияния новостей на финансовый рынок. Существуют[1] также отдельные работы, которые посвящены интеллектуальному анализу данных новостного потока. Если верить в гипотезу об эффективности фондовых рынков, то котировки ценных бумаг учитывают всю имеющуюся информацию. Однако, как показано в (Tetlock, 2007), вместо мгновенной адаптации к любому всплеску новостей цены

[1] Например, Wu X., Kumar V. etc. Top 10 algorithms in data mining // Knowl Inf Syst. – 2008. – p.1-37

на акции могут реагировать на такие события с некоторым опозданием, или не реагировать вовсе. Особенно часто такое происходит с «позитивными» новостями. Как подтверждают исследования (Palfrey, Wang, 2012), (Tetlock, 2007), (Dzielinski, Rieger, Talpsepp, 2011), рынок асимметрично реагирует на «негативные» и «позитивные» новости, причем именно «негативные» оказывают большее давление на цены.

Как видно из вышесказанного, проблема оценки эффекта влияния средств массовой информации на финансовый рынок является актуальной и активно исследуемой на сегодняшний день. Но большинство исследований проводятся на данных по западным рынкам, а на российских данных такие работы практически не представлены. Поэтому **целью** данной работы является определение степени и направления влияния тематических новостных потоков на российский фондовый рынок, используя в качестве инструмента анализа методы машинного обучения, включая эконометрические методы.

Для достижения поставленной цели были сформулированы следующие задачи:

1) На основе сопоставительного обзора литературы по изучению влияния СМИ на финансовый рынок выдвинуть ряд гипотез о наличии и особенностях такого влияния на российском рынке.
2) Определить наиболее подходящие методы оценки исследуемого эффекта и эмпирические стратегии проверки выдвинутых гипотез.
3) Выбрать подходящую теоретическую модель оценки влияния СМИ на котировки акций и разработать несколько спецификаций на ее основе.
4) Разработать процедуру сбора данных по новостной аналитике на основе применения метода сингулярного разложения матрицы новостей.
5) Собрать данные по новостной аналитике и котировкам акций крупнейших российских ПАО, используя разработанную в задаче четыре процедуру
6) Провести эмпирическое исследование для проверки выдвинутых гипотез, построив предложенные в задаче три модели на собранных данных.

В рамках данного исследования выдвигаются и проверяются следующие гипотезы:

1) Существует значимое влияние новостных потоков на волатильность фондового рынка в России.
2) Значимое влияние всплеска новостей может сохраняться на срок до 5-ти дней[2].
3) Влияние всплеска новостей на котировки акций конкретной компании зависит от отрасли, в которой функционирует данная компания.

[2] 5 дней это одна рабочая неделя, так как в выходные дни рынки не торгуются

Анализ проводится на основе котировок акций 26-ти крупнейших российских публичных акционерных обществ из различных отраслей: от нефтедобывающей до IT. Полный список этих компаний представлен в Главе 3. Выборка данных относится к временному отрезку 2020 года (подробнее о выборке написано в Главе 3). Для обработки новостей используется сингулярное разложение матрицы (SVD-разложение) – один из методов решения задачи классификации в машинном обучении. Для проверки выдвинутых гипотез применяются эконометрические методы работы с временными рядами.

Работа содержит 3 главы. **Первая глава** посвящена анализу литературы на тему влияния новостей на финансовые рынки разных стран. Системный обзор таких исследований проведен на основе их сопоставления по пяти основным критериям. На основе проведенного обзора выдвигаются гипотезы исследования и решаются первая и вторая задачи.

Вторая глава является теоретической и содержит две основные части. Первая часть посвящена методам анализа новостных потоков и тематических новостных потоков, а во второй части описываются методы анализа временных рядов на основе моделей условной и обобщенной условной авторегрессионной гетероскедастичности (ARCH и GARCH). Во второй главе решаются третья и четвертая задачи.

Третья глава – эмпирическая составляющая настоящей работы. Она состоит из четырех основных частей. Первая часть – описание выборки данных и источников. Вторая часть посвящена анализу новостного потока и выделению в нем пяти основных тем. Третья часть описывает построение трех эконометрических моделей на собранных данных. В четвертой части представлены результаты исследования, обсуждение этих результатов, их интерпретация, и проверка устойчивости. В третьей главе, таким образом, решаются пятая и шестая задачи.

Глава 1. Анализ эмпирических исследований

Критерии анализа литературы

Информационный фактор давно стал одной из ключевых переменных во многих эконометрических и финансовых моделях. Существует большое количество исследований о влиянии информации (в частности, информации о финансовых рынках и макроэкономической ситуации в целом (Dzielinski, Rieger, Talpsepp, 2011)) на показатели финансовых и фондовых рынков. Несмотря на различие в инструментарии, данных, гипотезах, большинство из них в той или иной степени подтверждают значимость такого влияния. В данной главе был проведен сопоставительный анализ работ, которые исследуют качественную и количественную взаимосвязь СМИ и фондовых рынков различных стран[3]. В качестве критериев сравнения были выбраны пять основных параметров таких исследований:

- **Год публикации исследования**. В зависимости от года возможны различия, например, в выборке данных. В разные промежутки времени могут быть различные результаты в силу возможных изменений конъюнктуры рынка. Различные мировые события вроде мировых кризисов, как например рецессия 2008, могут служить причинами аномальной волатильности финансового рынка. Также от года публикации зависит актуальность и релевантность работы.
- **Основная проверяемая гипотеза**. Правильно поставленный исследователями ключевой вопрос является одним из самых важных показателей качества любой научной работы. В зависимости от того, интересен ли основной вопрос исследования обществу, был ли дан на него убедительный ответ ранее или нет, будет зависеть успех всей работы. В связи с этим, данный показатель является одним из критериев сравнения работ данной тематики.
- **Данные.** Результаты проверки одних и тех же гипотез могут различаться зависимости от данных, на которых проводится исследование – в первую очередь на них могут влиять страна, виды исследуемых финансовых активов, год, источники собираемых новостей. От репрезентативности выборки зависит не только несмещенность оценок (количественный анализ), но и то, как результат всего исследования может быть проинтерпретирован, ведь это необходимо для

[3] Большинство исследований относится к фондовому рынку США. Но некоторые исследования посвящены российскому рынку и другим. Это будет описано ниже в процессе анализа статей и исследований на данную тематику

оценивания внешней и внутренней валидностей всего исследования (качественный анализ). Вполне возможно, что в какие-то промежутки времени новостные потоки оказывали меньшее влияние, когда ситуация в экономике в целом, а также финансовый рынок были относительно стабильны, а в какие-то промежутки, наоборот, такое давление было повышенным, как во времена кризисов или других экономических шоков.

- **Метод.** Как известно, от метода зависит многое. Если метод «плохой», то и полученные результаты могут быть неоправданными. Например, результаты могут противоречить фактическим или не иметь смысла вовсе. В процессе исследования научной литературы по влиянию медиа на финансовые рынки были выделены различные методы оценок, в том числе и «неудовлетворительные». Например, в некоторых исследованиях аппроксимация временного ряда происходила с помощью полиномов, зависящих от времени. Такие модели лишены предсказательных и описательных способностей, содержательный вывод на их основе невозможен. В свою очередь, «хорошие» методы бывают разные, они могут сильно отличаться по сложности, точности, временем работы, могут давать разный результат. В связи с этим, по мнению автора, данный критерий необходим для включения в сопоставительный анализ.
- **Результат.** Вторым ключевым показателем любой научной работы (после проверяемой гипотезы) является результат исследования. Как уже говорилось ранее, результат может сильно зависеть от используемого метода и выбранных данных. На разных данных в разное время разными методами получается разный результат. Если же авторы использовали «устойчивые» методы, или даже различные методы, использовали репрезентативные выборки по которым получался один и тот же результат, то таким результатам можно доверять и интерпретировать их каузально. В связи с вышесказанным, сопоставление работ по результатам поможет лучше оценить текущие знания в этой области, выявить противоречия, понять устойчивость основных выводов.

Сопоставительный анализ литературы

Результаты проведенного анализа литературы представлены в таблице 1. В данной таблице описаны 13 основных работ, тематика которых близка к теме, исследуемой в настоящей работе.

Таблица 1 состоит из пяти столбцов. В первом столбце указан автор(-ы) и год публикации статьи. Во втором столбце – название самой статьи. В третьем – основная

проверяемая гипотеза. В четвертом столбце указана информация о выборке данных, использованной в работе. Пятый столбец - методология авторов (методы, модели и прочее). В шестом столбце описан основный результат и краткий вывод всей работы.

Конечно, это далеко не полный список использованной литературы, полный список источников и литературы можно найти в соответствующем разделе в конце данной работы. Здесь же сосредоточены только ключевые работы по данной тематике. Самая цитируемая статья Тетлока (Tetlock, 2007), она имеет более 3000[4] тысяч цитирований. Итоговая таблица 1 для основных параметров выбранных научных исследований представлена ниже:

Таблица 1.

Автор, год	***Статья***	***Основная гипотеза***	***Данные***	***Метод***	***Результат***
L. Mitra, G. Mitra, 2011	«Applications of news analytics in finance: A review»	Новости влияют на доходности акций, торговые стратегии, позволяют контролировать риск	Данные собраны по американскому рынку за 2011 год, что соответствует году издания. Новостная аналитика собиралась из новостей, препринтов и социальных блогов	Машинное обучение (ML), алгоритмы классификации (байесовский, наивный классификаторы)	Подтверждается основная гипотеза о влиянии релевантных статей на доходности и поведение участников рынка, а также положительная корреляция с рыночной капитализацией
Sanjiv R. Das, 2011	«News analytics: Framework, techniques, and metrics»	Качественное изучение новостной аналитики помогает принять верное инвестиционное решение и управлять рисками	Американские данные по фондовому индексу MSH 35 за лето 2001 года. Новостная аналитика за то же время.	Методы машинного обучения (ML), байесовский классификатор, метод опорных векторов (SVM)	Новостная аналитика помогает лучше описывать динамику доходностей рынка и его волатильность
P. Ager Hafez, 2011	«How news events impact	Новости могут предвосхищать импульсы	Американские данные по индексы S&P	ПО Raven Pack для анализа финансовых	Новостная аналитика позволяет

[4] Поисковая система по полным текстам научных публикаций// официальный сайт. URL https://google.scholar.com/ (Дата обращения 06.04.2020)

	market sentiment»	доходностей индекса S&P500	500 за 2005-2008 годы. Новостная аналитика собрана с помощью Raven Pack	новостей, корреляционный анализ индекса настроений рынка и доходности индекса S&P500	предвосхитить ценовые импульсы индекса S&P500 вплоть до 1 месяца. Анализ новостей помогает улучшить инвестиционную стратегию, вовремя занимая короткие и длинные позиции
M. Dzielinski, M. Oliver Rieger, T. Talpsepp, 2011	«Volatility asymmetry, news, and private investors»	Всплеск новостей на макроэкономическую или финансовую тематику увеличивает волатильность фондового рынка	Данные по мировому фондовому рынку (MSCI World Index) за 2004-2008 годы. Новостная аналитика – число запросов в Google.	Семейство GARCH-моделей и их модификации	Увеличение запросов в Google по макроэкономике, например «рецессия» увеличивает асимметрию и волатильность рынка.
P. S. Kalev, H. Nhan Duong, 2011	«Firm-specific news arrival and the volatility of intraday stock index and futures returns»	Скорость поступления информации влияет на волатильность доходности	Данные по австралийскому фондовому рынку: S&P/ASX 200 и SPI 200 за 2003-2007 годы. И 2007-2009 годы отдельно (мировой кризис). Новостная аналитика – система SIRCA	GARCH и EGARCH-модели волатильности рынка	Скорость поступления информации оказывает влияние на волатильность доходности индекса S&P/ASX 200 и SPI 200. Кластеризация волатильности отражает последовательную корреляцию частот поступления информации на рынок
Yu-Pin Hu, R. S. Tsay, 2014	«Principal Volatility Component Analysis»	В портфеле, состоящем из большого количества ценных бумаг, существует	Котировки валют для наиболее экономически развитых стран	Модели семейства ARCH и GARCH. Методы	По данным 2000-2011 годов, существует линейная комбинация семи

		несколько основных, которые определяют большую часть дисперсии портфеля	с 2000 по 2011 годы.	машинного обучения (ML), метод главных компонент (PCA)	основных валют, для которых дисперсия не обладает свойством условной гетероскедастичности
T.R. Palfrey, S.W. Wang, 2012	«Speculative Overpricing In Asset Markets With Information Flows»	Спекуляция на фондовом рынке возникает из-за новостных потоков. В частности, из-за разной интерпретации поступающей информации. Помимо эффекта финансового и операционного рычагов, на волатильность рынка влияют «хорошие» и «плохие» новости	Статья является больше теоретической. Нет данных о выборке кроме того, что взяты рынки 5-ти ценных бумаг, их бид (bid) и объем торгов за выбранных 9 сессий по 600 секунд.	Байесовские методы решения задачи классификации новостей по темам. Модели семейства GARCH. Двумерная EGARCH-модель, допускающая асимметричную реакцию	Асимметричная ценовая реакция на хорошие и плохие новости, что является особенностью ценовой динамики. Фактором асимметрии вторых моментов (дисперсий) являются релевантные оптимистичные или пессимистичные новости, причем реакция на плохие новости больше, чем на хорошие
J.M. Maheu, T.H. Mccurdy, 2011	«News Arrival, Jump Dynamics, and Volatility Components for Individual Stock Returns»	Новостные потоки оказывают влияние на доходность, волатильность и моменты волатильности более высокого порядка	Выборочные данные до 2001 года (были сильно не сбалансированы) по дневным котировкам акций крупнейших ПАО США: Intel, HP, Coca-Cola, Apple и другие.	GARCH, GARCH со скачками с различными спецификациями	Два новостных потока: нормальный и аномальный, которые по-разному влияют на доходность и волатильность. Первый поток новостей вызывает постепенное изменение, второй поток вызывает скачки волатильности.

M.Melvin, Xixi Yin, 2000	«Public Information Arrival, Exchange Rate Volatility, and Quote Frequency»	Публичная информация при прочих равных условиях добавляет доходности и волатильности валютному рынку	Данные собраны по котировкам валют США, Германии и Японии за 1993-1995. Новостная аналитика – число заголовков в Reuters за тот же промежуток времени	Модели семейства ARCH, GARCH	Публичная информация влияет на частоту котировок и волатильность обменного курса
J.R. Nofsinger, 2001	«The impact of public information on investors»	Новостная аналитика (общественная информация) значимо влияет на принятие решений инвесторами	Котировки акций случайных 144 компаний из Нью-Йоркской фондовой биржи за 3 месяца с ноября 1990 года. Число статей в журнале Wall Street Journal за тот же срок.	Методы обработки панельных данных, МНК	Инвесторы увеличивают торги с выходом новостей (особенно новостей о дивидендах и доходах), причем индивидуальные инвесторы торгуют только на хороших новостях
Kenneth R. Ahern, Denis Sosyura, 2014	«Who Writes the News? Corporate Press Releases during Merger Negotiations»	Компании играют на рынке учитывая новостные потоки, а также рыночные настроения	Цены акций крупных ПАО США за 2000-2008 годы. Новостная аналитика - база данных Factivia (включает число заголовков The Wall Street Journal, The New York	Панельные данные, метод разность-в-среднем (difference-in-mean). Модели распределенных лагов (ADL).	Крупные сделки компаний порождают новости, причем на фоне повышенного количества новостей происходит кратковременное изменение котировок акций,

			Times и другие)		которыми пользуется фирма.[5]
P.C. Tetlock, 2007	«Giving Content to Investor Sentiment: The Role of Media in the Stock Market»	Существует значимое влияние финансовых новостей из Wall Street Journal на фондовый рынок	Индекс Dow Jones (США) с 1984 по 1999 годы. За новостную аналитику отвечает число статей в журнале The Wall Street Journal (США)	Методы машинного обучения (ML), метод главных компонент (PCA). Модель векторной авторегрессии (VAR)	Медиа оказывают давление на цены (пессимистические новости – понижательное давление) в краткосрочном периоде. Сильно оптимистичные или пессимистичные новости увеличивают объемы торгов
В. Балаш, П. Дате, С. Сидоров, 2013	«Использование данных новостной аналитики в GARCH моделях»	Внешние источники информации, такие как новости и объемы торгов влияют на волатильность ценных бумаг	19 компаний из FTSE 100 с 2005 по 2008 годы. Новостная аналитика: Raven Pack и данные об объеме торгов (считается новостью)	Модели семейства ARCH, GARCH	Переменная новостной интенсивности увеличивает долю объясненной дисперсии. Это означает, что новости являются фактором волатильности. В это же время переменная «объем торгов» значимо не повлияла на волатильность котировок акций

Источник: составлено автором

Чтобы более подробно проследить основные тенденции научных исследований на тему влияния медиа на фондовый и финансовый рынки, проведем анализ по результатам сравнения для каждого из критериев

[5] Более того, авторы в статье выдвигают новую гипотезу о том, что некоторые крупные намеренно выпускают финансовые новости для извлечения прибыли в результате операций с собственными акциями

1) **Год публикации исследования**. Как видно из выборки работ таблицы 1, подавляющее большинство работ было сделано после 2010 года. Самая «старая» работа (XiXi Yin, 2000) была опубликована в экономическом журнале «The Economic Journal» в 2000 году. По результатам анализа сроков публикации работ выявились две основные закономерности. Первая закономерность заключается в том, что в выборках, с которыми работают авторы каждого исследования, избегаются временные отрезки, относящиеся к мировому экономическому кризису 2008. Вероятно, это связано с тем, что этому промежутку времени соответствует аномальная волатильность мирового финансового рынка. Данное явление не обошло и российский финансовый рынок. Это можно легко проследить на динамике различных показателей: падение российского фондового индекса ММВБ почти на 60[6] п.п., падение индекса промышленного производства России больше чем на 15[7] п.п., рост валютного курса доллара США к российскому рублю примерно на 7[8] рублей. Такие колебания фундаментальных показателей относятся к аномальным. Существует отдельное направление исследований, изучающие последствия данного кризиса. Если ставить целью исследования обобщение результатов на более долгосрочную перспективу, то короткая выборка, включающая аномальные периоды, не является валидной. Второй обнаруженной закономерностью является связь между методами, которые исследователи применяют в своих работах, и годом публикации соответствующей работы. Было отмечено, что более «старым» работам соответствуют более «старые» методы. К примеру, модели условной авторегрессионной гетероскедастичности (ARCH) и обобщенной условной авторегрессионной гетероскедастичности (GARCH) применяются в более ранних исследованиях. Сами же модели впервые были описаны в работах конца 80-х и начала 90-х годов XX века. Исключением является только российская статья (Балаш, Дате, Сидоров, 2013), опубликованная в журнале «Прикладная эконометрика» в 2013 году. Затем, со временем, применяются более новые методы машинного обучения: байесовские методы машинного обучения; байесовский, наивный и другие алгоритмы решения задачи классификации из машинного обучения; метод

[6] Федеральная Служба государственной статистики // официальный сайт. URL https://www.gks.ru/ (Дата обращения 06.04.2020)

[7] Федеральная Служба государственной статистики // официальный сайт. URL https://www.gks.ru/ (Дата обращения 06.04.2020)

[8] Федеральная Служба государственной статистики // официальный сайт. URL https://www.gks.ru/ (Дата обращения 06.04.2020)

опорных векторов (относящийся к обучению с учителем, как и эконометрические методы) для задачи классификации новостей; метод главных компонент (метод непосредственно связан с SVD-разложением) и другие. В связи с этим можно сделать вывод о том, что для исследований в данной тематике используется самый актуальный эконометрический и прикладной математический аппараты. Отсюда можно сделать еще один вывод о том, что исследования в данной области остаются востребованными, поскольку и сделано еще не все, и применяемый аппарат все время развивается, что дает исследователям возможность давать все более точную качественную и количественную оценку влияния СМИ на финансовый рынок. Первым таким исследованием стала работа (P. Tetlock, 2007), ставшая фундаментальной в своей области. В ней был применен смешанный подход: метод главных компонент PCA и модель векторной авторегрессии VAR. Также была дана количественная оценка взаимодействия СМИ на американский фондовый рынок в 1990-х годах. О результатах этой работы будет написано ниже (см. пункт 4)

2) **Основная проверяемая гипотеза**. Следующим критерием сравнения является основной исследовательский вопрос. Как видно из таблицы 1, формулировки ключевых гипотез исследований не вовсе одинаковые, но все они сводятся к тому, что оценивается влияние СМИ на финансовый рынок. К финансовому рынку здесь относятся фондовый, денежный и другие рынки. Большой акцент делается на количестве и скорости поступающей информации: количество выпускаемых новостей за определенный промежуток времени, скорость поступления информации и тд. Помимо этого, отдельное внимание уделяется качественной составляющей новостей: анализируется влияние новостей, которые связаны с финансовым рынком напрямую или хотя бы косвенно. Как, например, исследователи (Dzielinski, Oliver Rieger, Talpsepp, 2011) нашли положительную корреляцию между числом запросов «рецессия» в поисковике Google и волатильностью финансового рынка. Данная связь была проинтерпретирована причинно-следственно: увеличение числа соответствующих запросов в Google провоцирует последующую волатильность финансового рынка. Еще одним направлением исследования является отдельное изучение влияния «хороших» и «плохих» новостей. Те работы, которые делали акцент на этом, подтверждали асимметричную реакцию рынка на «хорошие» и «плохие» новости. По результатам таких исследований был сформулирован ряд новых гипотез, например, результаты в исследовании (Ahern, Sosyura, 2014) натолкнули авторов на ряд новых гипотез о том, что крупные ПАО могли намеренно производить всплеск новостей, относящихся непосредственно к их компании для

того, чтобы производить спекулятивные операции на фондовом рынке: если новости «позитивные», то стоимость акций может расти, тогда капитализация растет. Если новости «негативные», то компания может скупать свои акции по заниженным ценам, а затем продавать по более высоким, получая чистую прибыль от таких спекулятивных операций. Помимо этого, многие исследователи, например (Nofsinger, 2001) и (Palfrey, Wang, 2011) искали взаимосвязь между СМИ и фондовым рынком на микроуровне: ими была поставлена гипотеза о том, что спекулятивные действия инвесторов и объемы торгов, создаваемых частными инвесторами, представляют собой асимметричную реакцию на поступающую на рынок информацию. Поскольку этот критерий самый важный, сделаем еще одну таблицу 2. По строкам данной таблицы будут находиться те же работы, что и в таблице 1, а по столбцам соотнесение результатов этих исследований с тремя проверяемыми гипотезами, которые будут проверяться настоящей работе. Если гипотеза 1 подтвердилась, то будет стоять знак плюс «+». Если не подтвердилась, то минус «-». Если данная гипотеза не проверяется даже косвенным образом, то будет стоять «НД».

Таблица 2.

Автор, год	***Статья***	***Гипотеза 1 (О влиянии СМИ на фондовый рынок РФ)***	***Гипотеза 2 (О лаге всплеска новостей)***	***Гипотеза 3 (О связи влияния и отрасли)***
L. Mitra, G. Mitra, 2011	«Applications of news analytics in finance: A review»	НД	НД	+
Sanjiv R. Das, 2011	«News analytics: Framework, techniques, and metrics»	НД	-	+
P. Ager Hafez, 2011	«How news events impact market sentiment»	НД	+	+

M. Dzielinski, M. Oliver Rieger, T. Talpsepp, 2011	«Volatility asymmetry, news, and private investors»	+	+	+
P. S. Kalev, H. Nhan Duong, 2011	«Firm-specific news arrival and the volatility of intraday stock index and futures returns»	НД	+	+
Yu-Pin Hu, R. S. Tsay, 2014	«Principal Volatility Component Analysis»	НД	НД	НД
T.R. Palfrey, S.W. Wang, 2012	«Speculative Overpricing In Asset Markets With Information Flows»	НД	-	+
J.M. Maheu, T.H. Mccurdy, 2011	«News Arrival, Jump Dynamics, and Volatility Components for Individual Stock Returns»	НД	+	+
M.Melvin, Xixi Yin, 2000	«Public Information Arrival, Exchange Rate Volatility, and Quote Frequency»	НД	+	НД
J.R. Nofsinger, 2001	«The impact of public information on investors»	НД	+	+
Kenneth R. Ahern, Denis Sosyura, 2014	«Who Writes the News? Corporate Press Releases during Merger Negotiations»	НД	+	+
P.C. Tetlock, 2007	«Giving Content to Investor Sentiment: The Role of Media in the Stock Market»	НД	+	+

В. Балаш, П. Дате, С. Сидоров, 2013	«Использование данных новостной аналитики в GARCH моделях»	НД	-	+

Источник: составлено автором

Из таблицы 2 видно, что в данной выборке нет данных о влиянии новостей на фондовый рынок России, за исключением одной статьи (Dzielinski, Rieger, Talpsepp, 2011) в которой российский рынок анализируется косвенно, как составляющая мирового фондового рынка. В процессе анализа литературы не было найдено других исследований по России, были обнаружены косвенные работы: о влиянии какой-то одной конкретной новости на финансовый рынок или другие рынки (например, недвижимости). В связи с этим, будет особенно интересно проанализировать эффекты влияния СМИ в России. Во второй гипотезе в 8 из 13 исследованиях было подтверждено влияние потока новостей на волатильность финансового рынка. Такое соотношение результатов является спорным, поэтому данная гипотеза представляет отдельный интерес. Третья гипотеза подтверждается в 11-ти исследованиях из 13-ти. Для того чтобы проверить такую гипотезу на российских данных будут рассмотрены динамики акций *отдельных* ПАО России, а не агрегированные показатели вроде индекс РТС или индекс МосБиржи (по аналогии с индексами Dow Jones, S&P 500 в работах (Hafez, 2011), (Tetlock, 2007))

3) **Данные.** Как уже было замечено ранее, авторы избегают временных отрезков, на которых рынки характеризуются повышенной волатильностью. Еще один вывод, который можно сделать на основе анализа массивов данных, состоит в том, что большинство исследований проводятся для фондовых рынков наиболее развитых стран, преимущественно для США. Третий вывод: из сопоставительного анализа таблицы 2 и столбца «Данные» следует, что если авторы берут большой промежуток времени (более 3-х месяцев), то гипотеза о длительном лаге всплеска новостей подтверждается. В противном случае, гипотеза либо опровергается, либо не тестируется.
4) **Метод.** Данный параметр был включен в анализ как отдельный критерий по двум причинам: первая причина заключается в том, что, как уже говорилось ранее, от используемого метода может зависеть результат самого исследования: нужно уделять особое внимание технической составляющей, использовать устойчивые методы. Вторая причина включения критерия – это сопоставительный анализ методов разных авторов для подбора наилучшей эмпирической стратегии в настоящей работе. Для обоснования эмпирической стратегии, которая будет

применяться в работе, необходимо предварительно проанализировать методы и модели, которые использовали авторы существующих работ по сходной тематике. Если подавляющее большинство исследований использует модели обобщенной условной гетероскедастичности (GARCH), то логичным шагом будет попытка применить данную методологию в настоящей работе, только уже в условиях российского, а не американского фондового рынка. Помимо этого, как было сказано в пункте 1, прослеживается четкая тенденция использования современных методов машинного обучения (Machine Learning) в данном направлении. Это также повлияло на конечную эмпирическую стратегию данной работы. Итогом сравнения по этому критерию стало следующее: во всех исследованиях используются либо эконометрические модели GARCH и VAR, либо другие методы машинного обучения, либо и то, и другое, как в работе (Tetlock, 2007). Причем, как это уже замечалось ранее, в более ранних работах используются модели семейства GARCH. В связи с этим, в данной работе была предпринята стратегия смешанного подхода, как и в вышеупомянутой работе (Tetlock, 2007)

5) **Результат**. Все рассмотренные работы характеризуются наличием результата о значимом влиянии новостного потока на финансовый рынок. Качественный анализ взаимодействия СМИ и финансового рынка в каждом исследовании также подтверждает наличие такого влияния медиа и средств массовой информации на фондовый и денежный рынки. Вне зависимости от метода, среди рассматриваемых работ была продемонстрирована устойчивость в результатах, которая может свидетельствовать о наличии причинно-следственной связи между поступающей на рынок информацией и такими показателями как доходность и волатильность рынка. Несмотря на устойчивость результата, что существует значимое влияние СМИ на финансовый рынок, в рассмотренных работах предлагаются альтернативные объяснения этому явлению: одна группа авторов утверждает, что это обусловлено асимметричной реакцией участников на поступающую информацию (Palfrey, Wang, 2011), другая объясняет это спекулятивными действиями крупных ПАО (Ahern, Sosyura, 2014), третья группа выделяет в качестве основного фактора волатильности именно «негативные» новости, как, например, в работах (Tetlock, 2007), а другие наоборот – «позитивные новости» (Nofsinger, 2001). Альтернативность интерпретаций кроется в различии проверяемых гипотез, используемых методов и, самое важное, собранных данных.

Сопоставительный анализ весьма полезен для погружения в тему и знакомства с фундаментальными понятиями, а также с основными наработками и результатами в данном направлении. Подобный анализ литературы также может служить хорошим толчком для проработки идей исследований из анализа литературы, а также формулирования собственных гипотез и эмпирических стратегий. В данной главе были описаны основные работы по изучению влияния СМИ на финансовый рынок, а также проведен их сопоставительный анализ. Это дало почву для дальнейшего исследования влияния СМИ на фондовый рынок в России в рамках данной работы. В следующей главе описывается теоретическая модель, гипотезы, которые будут проверяться, а также предполагаемые результаты применения этой модели на данных российского фондового рынка.

Глава 2. Теоретические основы влияния новостей на финансовые рынки

Настоящая глава посвящена теоретическим аспектам работы. Здесь описаны основные элементы используемых в работе теоретических моделей, а также представлен краткий обзор понятий, методов и моделей, необходимых для построения эмпирической части данного исследования (представленной в Главе 3). Глава 2 состоит из двух частей: первая часть посвящена новостной составляющей и методам обработки новостей, а вторая часть посвящена построению эконометрической модели. Также во второй части предложена новая спецификация модели, разработанная автором на основе на основе моделей из обзора научной литературы по влиянию СМИ на фондовые рынки.

Новостная аналитика

Для изучения влияния новостей на финансовый или фондовый рынок необходимо сопоставить каждой новости некоторую числовую характеристику. Существует большое множество методов машинного обучения, созданных специально для решения этой задачи. Самыми распространенными подходами на сегодняшний день являются векторные модели.

Суть их заключается в том, что каждый документ[9] рассматривается как вектор. Далее возникает вопрос о том, что будет служить координатами такого вектора: буквы, слова, словосочетания или что-то другое? Ответ на этот вопрос неоднозначен: все зависит от подхода. К примеру, если брать буквы, то введение тридцатитрёхмерного[10] векторного пространства содержательного ничего не даст, так как при простом подсчете числа того, сколько раз встретилась каждая буква возникают проблемы с тем, что такое соответствие между векторным пространством и текстовым сообщением (новостью) не будет взаимооднозначным. В случае буквенно-векторного представления разные по смыслу и набору слов новости могут иметь один и тот же вектор, тогда любые операции с такими векторами не будут нести смысловой нагрузки.

Представление новости в виде вектора, координаты которого отвечают за слова, тоже нуждается в уточнении. К примеру, в словаре Даля, составленном в XIX веке, около 200 тысяч слов. Работа с такими объемами данных потребует большой вычислительной мощности уже для нескольких предложений. В связи с этим такое векторное представление является неэффективным.

Самое распространенное векторное[11] представление документа: каждому документу ставится в соответствие ставится единственный вектор $v \in \mathbb{R}^k$:

$$\{\text{Документ}\} \to v = (v_1, \dots, v_k)$$

Где k – наперед заданное число, отвечающее за общее количество используемых терминов. О том, чем являются координаты этого вектора написано ниже. Такое число может подбираться из контекста всей базы таких документов (на практике используется число от 500 до 1000). Для того чтобы два слова с одним и тем же корнем с разными окончаниями обрабатывались не как разные слова, а также для устранения подобных проблем применяют предварительные преобразования документа (нормализацию документа):

- Лемматизация документа – приведение словоформы каждого слова документа к ее словарной (нормальной) форме. Например, в результате такой операции над документом, слово «Дипломная» поменяется на «Дипломный». Существительные – в именительный падеж, единственное число; прилагательные – в именительный

[9] Всюду в работе слова «текст», «документ», «новость» будут пониматься как синонимы

[10] Равно числу букв в кириллице, в случае работы с новостями на русском языке

[11] Хорошо описана в Sidorov G., Gelbukh A., Gómez-Adorno H., Pinto D. Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model // Computación y Sistemas.- 2014.- vol.18(3).- p.491–504.

падеж, единственное число, мужской род; причастия, деепричастия и глаголы меняют на глаголы в инфинитиве несовершенного вида.

- Стемминг документа – процесс нахождения неизменяемой части слова каждого документа, которая выражает его лексическое значение. Например, в слове «Дипломная» удаляется окончание «-ая».

Отдельно необходимо отметить, что русский язык имеет достаточно сложную морфологическую изменяемость слов. В связи с этим часто возникают ошибки I и II рода: ситуации, когда верная нулевая гипотеза (у *разных* слов *разные* неизменяемые части) не принимается, и когда неверная нулевая гипотеза принимается соответственно. *Overstemming* – ошибка I рода, разным словам ставится в соответствие одна и та же часть (к примеру, словам «Ломоносов» и «Ломать» ставится одно и то же слово «Лом). *Understemming* – ошибка II рода, морфологическим формам одного и того же слова ставятся в соответствие разные неизменяемые части (например, словам «Ломоносов» и «Ломоносова» ставятся формы «Ломоносов» и «Ломоносова» соответственно, вместо единственной верной неизменяемой части «Ломоносов»)

- Удаление стоп-слов – процесс исключения из документов общих и зависимых шумовых слов. К общим стоп-словам (общим шумовым словам) относятся предлоги, союзы, междометия, цифры и так далее. Зависимые стоп-слова (зависимые шумовые слова) происходят от контекста. К примеру, в запросе «Ломоносов Михаил Васильевич» слова «Михаил», «Васильевич» являются зависимыми стоп-словами, так как не несут дополнительной смысловой нагрузки и отображать их смысла нет.
- Если алгоритм чувствителен к регистру[12], то весь текст приводят к нижнему регистру для того, чтобы, к примеру, одно и то же слово в начале предложения и в середине воспринималось как одно и то же, а не как два разных слова.
- Токенизация документа - разбиение документа на более мелкие части, так называемые токены (слова, знаки пунктуации и т.д.).

В результате таких преобразований документа возможно его «корректное» векторное представление. Следующая задача, которая возникает после нормализации документа, это определение того, что будет являться координатами вектора. Ответ на этот вопрос также неоднозначен, но выделяют два основных подхода:

1) Булевское заполнение: простой подсчет встреченных нормальных словоформ каждого слова из набора терминов в документе

[12] В этом исследовании анализ новостей проводился в высокоуровневом языке программирования Python, в связи с чем при нормализации текста, документ новостей приводился к нижнему регистру.

2) Заполнение по статистической мере TF-IDF. Такая мера используется для оценки важности слова в документе. Является наиболее часто встречающейся в подобных исследованиях. Значение TF-IDF разбивается на 2 сомножителя: TF и IDF, где:

$$TF(t) = \frac{n_t}{\sum n_j} \tag{1.1}$$

$TF(t)$ - (term frequency) – вес слова t во всем документе.

- n_k – число, обозначающее сколько раз встретилось данное слово в документе
- $\sum n_j$ – общее число слов в данном документе

$$IDF(t) = \ln \frac{|D|}{|\{d_i \in D : t \in d_i\}|} \tag{1.2}$$

$IDF(t)$ - (inverse document frequency) – обратная частота документа.

- $|D|$ - общее число документов (новостей)
- $|\{d_i \in D : t \in d_i\}|$ – общее число документов, в которых слово t встречается хотя бы один раз

Для того чтобы данное векторное пространство стало метрическим, вводят функцию, обладающую определенными свойствами, так называемую метрику. Можно ввести стандартные метрики: октаэдрическую, евклидову, кубическую, однако на практике (для подобных векторных пространств) распространена косинусная метрика в силу своей эффективности в качестве оценочной меры. Ведь, как правило, такие векторные представления сильно разрежены (большое количество нулевых координат). Поэтому похожесть новостей определяется с помощью косинуса угла между ними. Пусть v_i, v_j - два вектора-новости. Тогда расстояние между этими векторами определяется следующим образом:

$$\rho(v_i, v_j) = 1 - \frac{(v_i, v_j)}{|v_i| * |v_j|} \tag{1.3}$$

Где

- (v_i, v_j) – скалярное произведение векторов
- $|v_i|,\ |v_j|$ - нормы соответствующих векторов

Если в качестве φ обозначить угол между векторами v_i, v_j , то (1.3) можно переписать в следующем виде:

$$\rho(v_i, v_j) = 1 - \cos\varphi \tag{1.4}$$

Из неравенства:

$$0 \leq cos\varphi \leq 1 \tag{1.5}$$

для любого угла $\varphi \in [0; \frac{\pi}{2}]$, следует другое неравенство:

$$0 \leq \rho(v_i, v_j) \leq 1 \tag{1.6}$$

Если новости «похожи», то и соответствующие им векторы похожи. Для похожих векторов угол между ними небольшой, а значит его косинус близок к единице, поэтому расстояние между новостями, заданными похожими векторами, будет небольшим, и наоборот, абсолютно разным новостям будут соответствовать ортогональные векторы. Математически это можно записать следующим образом: $\lim_{\varphi \to 0} \rho(v_i, v_j) = 0$. И наоборот. Абсолютно разным новостям будет соответствовать большое значение функции расстояния (равное единице на паре ортогональных векторов). Введение метрики необходимо для отражения взаимосвязи между новостями.

После всех преобразований и подсчета мер образуется массив векторов:

$$\begin{gathered} v_1 = (a_{11}, \dots, a_{1N_{\text{терминов}}}) \\ \dots \\ v_{N_{\text{новостей}}} = (a_{N_{\text{документов}}1}, \dots, a_{N_{\text{новостей}}N_{\text{терминов}}}) \end{gathered} \tag{1.7}$$

Где:

- $N_{\text{новостей}}$ – общее число документов в выборке
- $N_{\text{терминов}}$ – число используемых терминов. Как уже говорилось ранее, данное число определяется из контекста (в данном исследовании $N_{\text{терминов}} = 1000$)

По данным векторам строится матрица документов-слов $A(N_{\text{новостей}} \times N_{\text{терминов}})$, у которой строками является набор векторов $v_1, \dots, v_{N_{\text{новостей}}}$.

Понятно, что если не использовать каждое слово во всех новостях (документах), то матрица будет прямоугольной, причем $N_{\text{новостей}} > N_{\text{терминов}}$.

Для получения основных тем новостей можно применить методы линейной алгебры. Известно, что для любой квадратной матрицы $A(n \times n)$ можно получить следующее разложение:

$$A = T\Sigma T^{-1} \tag{1.8}$$

Где:

- Матрица $\Sigma = diag(\lambda_1, \dots, \lambda_n)$, если набор собственных векторов оператора A является полным. Если же максимальное число линейно-независимых собственных векторов меньше размерности пространства, то матрица Σ имеет жорданову форму (блочно-диагональную).
- T – матрица собственных векторов, записанных по столбцам (матрица перехода к новому базису), если существует такой набор собственных векторов. Если не существует, то к набору собственных векторов добавляются присоединенные векторы.

Напомним, что каждое число $\lambda \in \mathbb{C}$ определяемое из характеристического уравнения:

$$\chi(\lambda) = \det(A - \lambda E) = 0 \tag{1.9}$$

называется собственным значением линейного оператора A.

Вектор x называется собственным вектором, отвечающим собственному значению $\lambda \in \mathbb{C}$ для линейного оператора с матрицей A, если верно следующее:

$$\exists \lambda \in \mathbb{C}: \quad Ax = \lambda x \tag{1.10}$$

Где E – единичная матрица, того же размера, что и квадратная матрица A.

Вектор f называется присоединенным вектором высоты m, отвечающим собственному значению для линейного оператора с матрицей A, если верны следующие соотношения:

$$\begin{cases} (A - \lambda E)^{m-1} \neq 0 \\ (A - \lambda E)^m f = 0 \end{cases} \tag{1.11}$$

В случае, если алгебраическая кратность всех корней уравнения $\chi(\lambda) = 0$ равна их геометрической кратности (размерности ядра оператора $(A - \lambda E)$), то существует базис

пространства E^n из собственных векторов линейного оператора с матрицей A. Если же для каких-то собственных значений оказывается так, что их алгебраическая кратность больше их геометрической кратности, то набор собственных векторов не образует и базиса, и такой набор дополняют до базиса с помощью добавления присоединенных векторов.

Данная теория была построена для квадратных матриц. Но в последствие была обобщена для прямоугольных матриц. Оказывается, что любую прямоугольную матрицу $A(n \times m)$ можно разложить в композицию трех. Одним из самых известных разложений является разложение SVD[13]. Данная теория является «свежей» по математическим меркам, но уже успела найти большой отклик не только в математике и экономике, но и в социологии, политологии, психологии и других науках.

Для симметричной матрицы спектральное разложение дает нам исчерпывающую информацию о ней. SVD-разложение является прямым аналогом такого разложения для любых прямоугольных матриц.

Пусть дана матрица $A(n \times m)$, $n > m$, тогда ее сингулярным разложением (SVD-разложением) называется разложение вида:

$$A = U\Sigma V^* \tag{1.12}$$

Где:

- Σ – прямоугольная диагональная матрица
- $U(n \times n)$ – унитарная прямоугольная матрица: $U^*U = E(n \times n)$
- V^*- унитарная матрица, сопряженная к матрице V: $V^*V = E(m \times m)$

Числа $w_1, \dots, w_m$ являются сингулярными числами матрицы A, причем верно:

$$w_1 \geq w_2 \geq \cdots \geq w_m \geq 0 \tag{1.13}$$

Итак, мы получили разложение матрицы A и определили понятие ее сингулярных чисел. Если матрица A неполного ранга, то есть $rank(A) = r$, где $r < m < n$, (как это часто бывает на практике, особенно с векторной моделью текстов в силу разреженности матриц), то все сингулярные числа, начиная с w_{r+1} равны нулю. Сингулярные числа прямоугольной матрицы играют ту же роль, что и собственные числа для симметричной матрицы: число

[13] Singular value decomposition. Хорошо описано в Press W.H., Flannery B.P., Teukolsky S.A., Vetterling, W.T. Numerical Recipes in C// Cambridge University Press. - 1988

ненулевых таких чисел равно рангу матрицы. Определим $u_1, \dots, u_n$ и $v_1, \dots, v_m$ как векторы-столбцы матриц U и V соответственно. Тогда верны следующие соотношения:

$$Av_i = w_i u_i \tag{1.14}$$

$$A^T u_i = w_i v_i \tag{1.15}$$

Для того, чтобы проследить соотношения между сингулярными и собственными числами, можно перемножить матрицу A на саму себя транспонированную:

$$AA^T = U\Sigma V^*(U\Sigma V^*)^T = U\Sigma V^* V\Sigma U^* = U\Sigma^2 U^* \tag{1.16}$$

$$A^T A = (U\Sigma V^*)^T U\Sigma V^* = V\Sigma U^* U\Sigma V^* = V\Sigma^2 V^* \tag{1.17}$$

Размерность $AA^T(n \times n)$, а матрицы $A^T A(m \times m)$. И в том, и в другом случае видно, что собственные числа являются квадратами сингулярных для соответствующих матриц, стоящих в левой части уравнений (1.16) и (1.17):

$$A^T A(v_i) = A^T w_i u_i = w_i^2 v_i \tag{1.18}$$

$$AA^T(u_i) = A w_i v_i = w_i^2 u_i \tag{1.19}$$

То есть вектор v_i – собственный вектор матрицы $A^T A$, векторы u_i – собственный вектор матрицы AA^T с собственными значениями w_i^2.

Особый интерес SVD-разложение представляет в данном исследовании для аппроксимации, которую необходимо применить для выделения ключевых тем из выборки новостей.

Из равенств SVD-разложения матрицы следует:

$$A = \sum_{i=1}^{r} w_i u_i v_i^T \tag{1.20}$$

Где $u_i v_i^T$ $(n \times 1 \times 1 \times m) = u_i v_i^T (n \times m)$ матрицы ранга 1. Отсюда следует, что матрица A ранга r раскладывается в сумму матриц ранга 1. В силу ортонормированности U, V, верно, что для каждого i фробениусова норма матриц $u_i v_i^T$ равна 1:

$$\forall i \quad \left\| u_i v_i^T \right\|_F^2 = 1 \tag{1.21}$$

Отсюда следует, что наибольшую «значимость» в представлении (1.20) имеют слагаемые с наибольшими сингулярными числами. Это означает, что для наилучшей аппроксимации достаточно сохранять первые сингулярные значения. Поэтому определим матрицу:

$$\hat{A} = \sum_{i=1}^{k} w_i u_i v_i^T \tag{1.22}$$

Где $k < r$, как матрицу, аппроксимирующую A.
Ошибку аппроксимации определим как фробениусову норму их разностей:

$$\left\| A - \hat{A} \right\|_F^2 \tag{1.23}$$

Но такая ошибка не учитывает норму самой матрицы, и, вообще говоря, является неограниченной величиной. Поэтому определим относительную величину ошибки:

$$\Delta E_k = \frac{\left\| A - \hat{A} \right\|_F^2}{\|A\|_F^2} \tag{1.24}$$

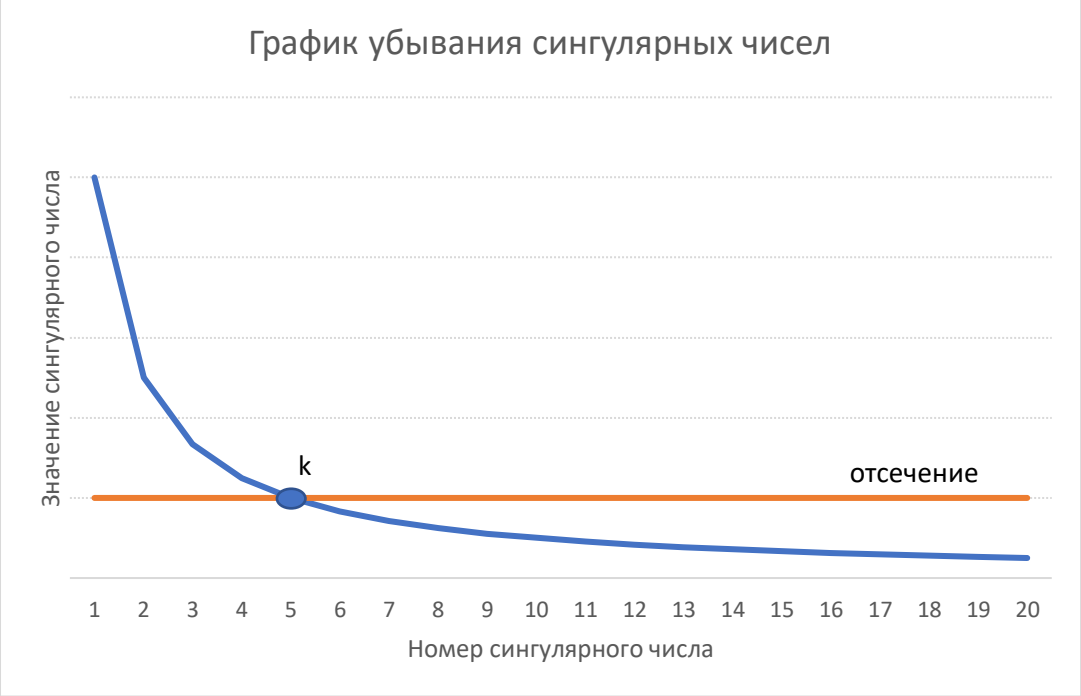
Тогда верна следующая цепочка рассуждений:

$$\begin{aligned} \Delta E_k = \frac{\left\| A - \hat{A} \right\|_F^2}{\|A\|_F^2} = \frac{\left\| \sum_{i=k+1}^{r} w_i u_i v_i^T \right\|_F^2}{\left\| \sum_{i=1}^{r} w_i u_i v_i^T \right\|_F^2} = \frac{\sum_{i=k+1}^{r} w_i u_i v_i^T \sum_{j=k+1}^{r} w_j u_j v_j^T}{\sum_{i=1}^{r} w_i u_i v_i^T \sum_{j=1}^{r} w_j u_j v_j^T} \\ = \frac{\sum_{i=k+1}^{r} \sum_{j=k+1}^{r} w_i w_j \left(u_i v_i^T, u_i v_i^T \right)}{\sum_{i=1}^{r} \sum_{j=1}^{r} w_i w_j \left(u_i v_i^T, u_i v_i^T \right)} = \frac{\sum_{j=k+1}^{r} w_j^2}{\sum_{j=1}^{r} w_j^2} \end{aligned} \tag{1.25}$$

Это означает, что ошибка аппроксимации равна доле квадратов отброшенных сингулярных чисел, начиная с $k + 1$ среди квадратов всех чисел. Для определения параметра k можно

смотреть на график убывания нормированных сингулярных чисел. На практике график может выглядеть следующим образом:

Рисунок 1.



Источник: составлено автором.

Если все сингулярные числа начиная с $k+1$ достаточно близки к нулю, то для аппроксимации они все обнуляются. После обнуления получается аппроксимирующая матрица $\hat{A}$, которая и является решением задачи наилучшей аппроксимации относительной ошибки в смысле с фробениусовой формы с наперед заданным числом k.

Если верно (1.12), то аппроксимация определяется следующим образом:

$$\hat{A} = U\hat{\Sigma}V^* \tag{1.26}$$

Где

- $\hat{\Sigma}$ – прямоугольная диагональная с первыми k сингулярными числами $w_1, \dots, w_k$ на диагонали
- U, V^* - матрицы, получаемые из SVD-разложения матрицы A

Отсюда понятна становится вся ценность SVD- разложения в прикладных задачах[14].

Итак, в данном разделе представлено краткое теоретическое обоснование теории SVD-разложения прямоугольных матриц, которое было использовано в данной работе

[14] Также сингулярное разложение может использоваться при нахождении псевдообратной матрицы для задачи МНК

применительно к матрице новостей. Более подробно об этом можно почитать, например, в (Girosi, King, 2008)[15].

После SVD-разложения матрицы новостей и ее аппроксимации для выделения ключевых тем новостей по выборке с помощью подбора параметра k, происходит качественный анализ топиков (тем). Первой теме соответствует первое сингулярное число. Из соответствующей этому топику матрицы $w_1 u_1 v_1^T$ достаются наибольшие элементы, которые означают конкретные слова из списка терминов, а затем ранжируются в порядке невозрастания. Так мы получаем топ самых важных слов в первой теме. Данная процедура осуществляется для всех k тем. Для того чтобы определить к какой из тем следует отнести конкретную новость, применяется следующая процедура: вектор, соответствующий новости, которой необходимо присвоить тему, умножается на $\hat{\Sigma}V^*$, в результате получается вектор $(a_1, \dots, a_k)$, у которого координаты соответствуют «близости» конкретной новости к каждой из тем. Простым сравнением этих чисел выбирается наибольшее, а затем данной новости присваивается конкретная тема.

Эконометрическая модель

Обработка данных новостной аналитики требует больших усилий. Немалых усилий требует и создание хорошей эконометрической модели. Как известно из литературы по финансовым исследованиям (Engle, 1982), (Bollerslev 1986), в результате наблюдений за такими временными рядами как котировки акций, валютный курс и тд, наблюдения с большими и малыми реализованными волатильностями образуют некоторую закономерность: имеют тенденцию «похожих» отклонений от ожидаемого значения в некоторой небольшой окрестности. Это означает, что периоды «спокойного» и «волатильного» рынка чередуются. Модели семейства ARCH и GARCH хорошо [16] описывают подобные явления рынка ценных бумаг.

Пусть y_t – котировка акций некоторой компании – временной ряд, подчиняющийся модели ARIMA(p,d,q):

$$\Delta^d y_t = \phi_0 + \sum_{j=1}^{p} \phi_j * \Delta^d y_{t-j} + \sum_{k=1}^{q} \theta_k * \varepsilon_{t-k} + \varepsilon_t \qquad (1.27)$$

[15] Girosi F., King G. Demographic forecasting. // Princeton. - 2008. from p.234

[16] Как было показано в работе Engle R.F., Lilien D., Robins R Estimation of time varying risk premiums in the term structure // Econometrica, Vol. 55, No. 2. – 1987. - p. 391-407

Где

- $\Delta^d y_t$ – разность порядка d по времени для y_t
- $\Delta^d y_{t-j}$- разность порядка d по времени для лага j переменной y_t
- $\phi_0, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q$ -параметры этой модели
- ε_t – ошибка модели с нулевым математическим ожиданием для каждого момента времени t
- ε_{t-k} – k-й лаг ошибки модели

Пусть $\sigma_t^2 = Var\left(\varepsilon_t \middle| \varepsilon_{t-1} \dots \varepsilon_{t-n_1}\right) = E\left(\varepsilon_t^2 \middle| \varepsilon_{t-1} \dots \varepsilon_{t-n_1}\right)$ - условная дисперсия ошибок ε_t. В таком случае подобная закономерность (в некоторых источниках именуется «кластеризацией»[17]) может быть объяснена следующей зависимостью:

$$\varepsilon_t = u_t \cdot \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \dots + \alpha_{n_1} \varepsilon_{t-n_1}^2} \tag{1.28}$$

Где u_t – белый шум с единичной дисперсией

Тогда:

$$\begin{aligned} \sigma_t^2 &= Var\left(\varepsilon_t \middle| \varepsilon_{t-1} \dots \varepsilon_{t-n_1}\right) \\ &= Var\left(u_t \cdot \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2} + \dots + \alpha_{n_1} \varepsilon_{t-n_1}} \middle| \varepsilon_{t-1} \dots \varepsilon_{t-n_1}\right) \\ &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \dots + \alpha_{n_1} \varepsilon_{t-n_1}^2 \end{aligned} \tag{1.29}$$

В случае, если условная дисперсия подчиняется соотношению (1.29), такая модель называется ARCH [18] -моделью порядка n_1 (ARCH(n_1)). Обобщенной моделью авторегрессионной условной гетероскедастичности (GARCH-моделью) порядка $(n_1; n_2)$ называется следующая модель:

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{n_1} \alpha_j \varepsilon_{t-j}^2 + \sum_{k=1}^{n_2} \beta_k \sigma_{t-k}^2 \tag{1.30}$$

Помимо использования стандартных моделей ARCH и GARCH, в данной работе предлагается модификация GARCH-модели с помощью добавления во вспомогательную регрессию следующих переменных:

- $topic_{m,l}$ - число новостей темы m в день l; $1 \le m \le k$

[17] Например, Generalized autoregressive conditional heteroskedasticity, Bollerslev, 1986

[18] Autoregressive Conditional Heteroskedasticity

Где общее число k тем определяется из задачи аппроксимации матрицы новостей A путем анализа убывания нормированных сингулярных чисел, получаемых их SVD-разложения (нормировка производится путем деления на первое из них, которое является максимальным).

Модифицированная GARCH – модель имеет следующий вид:

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{n_1} \alpha_j \varepsilon_{t-j}^2 + \sum_{k=1}^{n_2} \beta_k \sigma_{t-k}^2 + \sum_{r=1}^{n_3} \sum_{l=0}^{n_4} \gamma_{rl} \cdot topic_{r,t-l} \tag{1.31}$$

Где n_1, n_2, n_3, n_4 отвечают за максимальный порядок лага квадрата ошибки модели ARIMA, лага условной дисперсии, число новостей и максимальный порядок лага числа новостей соответственно.

Для того чтобы протестировать наличие ARCH-эффектов строится вспомогательная регрессия на квадраты остатков в основной модели, а затем тестируется значимость коэффициентов перед квадратами лагов остатков с помощью F-теста (значимость уравнения в целом) с расчетной статистикой:

$$F_{\text{расч.}} = \frac{R^2}{1 - R^2} \cdot \frac{n - k}{k - 1} \tag{1.32}$$

Где:

- R^2 – коэффициент детерминации из модели регрессии на квадраты остатков полученных из основной модели
- n – общее число наблюдений
- k – число регрессоров (включая константу) из модели регрессии на квадраты остатков полученных из основной модели. В терминах уравнения (1.30), $k = n_1 + n_2 + 1$

Либо, как альтернативу этому тесту, можно применить LM-тест (тест множителей Лагранжа). В этой работе в качестве критерия значимости использовался F-тест. В Главе 3 описаны результаты таких тестирований на выбранных данных.

После построения модели необходимо проверить на значимость все переменные, отвечающие за новостную аналитику. Сделать это можно с помощью формального теста «короткая» против «длинной» (модели (1.30) и (1.31) соответственно). Нулевая гипотеза:

$$H_0: \quad \forall\, r, l \quad \gamma_{rl} = 0 \tag{1.33}$$

И тестовая статистика:

$$F_{\text{расч.}} = \frac{R_{UR}^2 - R_R^2}{1 - R_{UR}^2} \cdot \frac{n - 2 - \sum_{j=1}^{4} n_j}{n_3 + n_4 + 1} \tag{1.34}$$

Где R_{UR}^2, R_R^2 – коэффициенты детерминации из моделей (1.31) и (1.30) соответственно.

Если нулевая гипотеза принимается, то в рамках данной модели можно сделать вывод о том, что новости не оказывают непосредственного влияния на волатильность цен акций конкретного российского ПАО. Если же гипотеза не принимается, то тогда можно сделать вывод о том, что общий поток тематических новостей оказывает значимое влияние на волатильность котировок акций соответствующего ПАО. Данный тест встроен во все стандартные эконометрические пакеты. Подробнее об этом написано в Главе 3.

Чтобы убедиться в «устойчивости» полученных результатов, строятся еще 2 альтернативные модели. Изменение спецификаций особенно полезно для проведения качественного анализа, поскольку, если, к примеру, только одна из трех моделей подтверждает значимость переменных, отвечающих за новости, то стоит задаться вопросом об ее эксплицитности.

Первая альтернативная модель получается из предыдущей за счет агрегации переменных, отвечающих за новостную аналитику:

$$\sum_{r=1}^{n_3}\sum_{l=0}^{n_4} \gamma_{rl} \cdot topic_{r,t-l} = \sum_{l=0}^{n_4}\sum_{r=1}^{n_3} \gamma_{rl} \cdot topic_{r,t-l} = \\ = \sum_{l=0}^{n_4} [\gamma_{1l} topic_{r,t-l} + \cdots \gamma_{n_3 l} topic_{n_3,t-l}] \tag{1.35}$$

Введем новое обозначение:

$$number\_of\_news_{t-l} = \sum_{r=1}^{n_3} \gamma_{rl} \cdot topic_{r,t-l} \tag{1.36}$$

Тогда новая переменная[19] будет обозначать общий поток новостей за день $t - l$. Данная модель позволит оценивать меньшее число параметров:

$$\sum_{l=0}^{n_4} [\gamma_{1l} topic_{r,t-l} + \cdots \gamma_{n_3 l} topic_{n_3,t-l}] \equiv \sum_{l=0}^{n_4} \gamma_l \cdot number\_of\_news_{t-l} \tag{1.37}$$

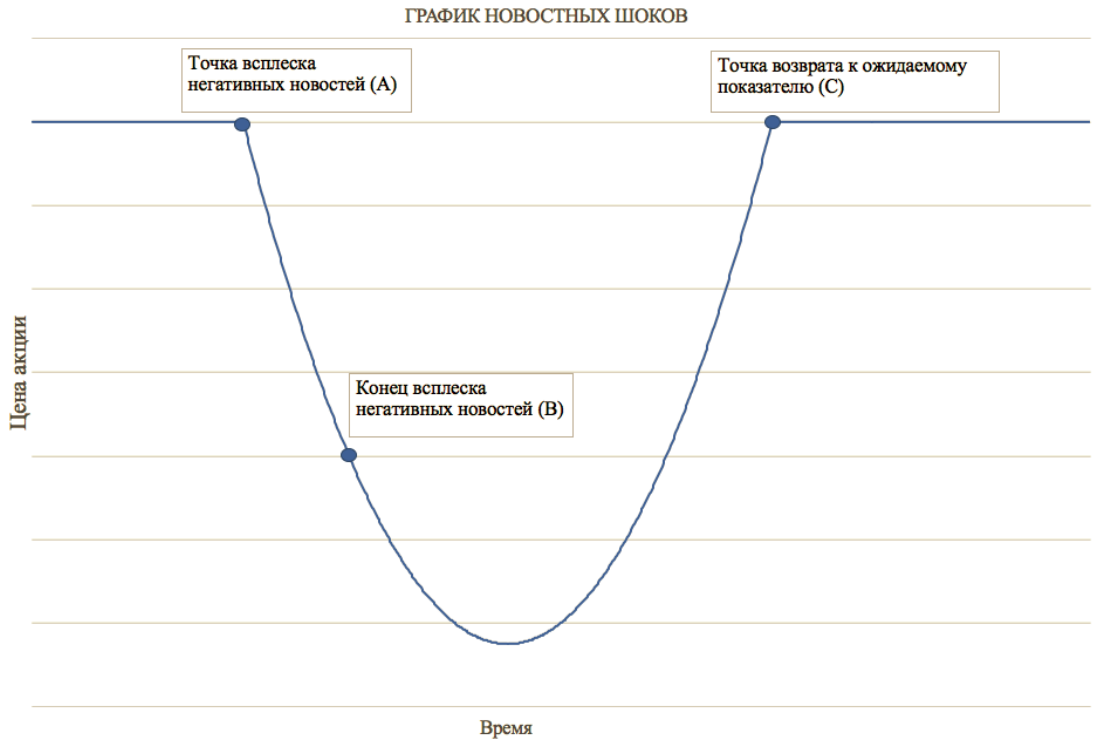
[19] Здесь и далее подразумевается, что данная также будет тестироваться на стационарность с помощью вышеизложенных тестов.

Итоговая модель имеет вид:

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{n_1} \alpha_j \varepsilon_{t-j}^2 + \sum_{k=1}^{n_2} \beta_k \sigma_{t-k}^2 + \sum_{l=0}^{n_4} \gamma_l \cdot number\_of\_news_{t-l} \quad (1.38)$$

В чем преимущества и недостатки такой модели? Преимуществом является то, что в данной модели меньше параметров, а значит можно использовать больший порядок лага. В соответствии с теорией, выведенной в работе (Tetlock, 2007), негативные новости могут оказывать понижательное давление на цены, как показано на графике, приведенном ниже на рисунке 2:

Рисунок 2.



Источник: составлено автором.[20]

Согласно теории Тетлока (Tetlock), временной отрезок, соединяющий точки A и B, может продолжаться до 5 дней. Для модели (1.38) на новостную аналитику приходится только $n_4 + 1$, что в n_3 раз меньше, чем в модели (1.31). В связи с этим возможно рассмотрение больших лагов для выявления более долгих шоков. Недостатком данной модели является то, что она учитывает лишь общий поток новостей за день, и никак не

[20] Рисунок является адаптацией оригинала из Paul C. Tetlock «Giving Content to Investor Sentiment: The Role of Media in the Stock Market» // The Journal Of Finance vol. LXII, - June 2007 p. 1139-1168

учитывает общую тематику. Модель (1.31) учитывает тематику и направлена на качественный анализ дневных новостей, в зависимости от того, что является повесткой конкретного дня, и поэтому она является более предпочтительной. Тем не менее, модель (1.38) будет полезна для анализа в качестве одной из альтернативных моделей. Также подобная модель была использована в статье (Балаш, Дате, Сидоров, 2013).

Суть второй альтернативной модели заключается в том, что теперь можно измерить не «непосредственное влияние» на волатильность, как в случае с моделью (1.31), поскольку все значимые коэффициенты интерпретируются именно таким образом, а только опосредованное, за счет изменения цен самого актива. Ее спецификация выглядит следующим образом:

$$\Delta^d y_t = \phi_0 + \sum_{j=1}^{p} \phi_j * \Delta^d y_{t-j} + \sum_{k=1}^{q} \theta_k * \varepsilon_{t-k} + \sum_{r=1}^{n_3} \sum_{l=0}^{n_4} \gamma_{rl} \cdot topic_{r,t-l} + \varepsilon_t \quad (1.39)$$

Где ошибка модели ε_t подчиняется модели (1.30), то есть имеет GARCH-эффект[21]. Единственное, что здесь нужно проверить отдельно, это стационарность переменных, отвечающих за новостную аналитику. Для этого будут использоваться два основных теста: расширенный тест Дики-Фуллера (ADF-тест) с нулевой гипотезой о том, что ряд не стационарен, и тест KPSS с нулевой гипотезой о том, что ряд является тренд-стационарным. Данные тесты также встроены в любой хороший эконометрический пакет.

Итак, в данной главе были описаны теоретические основы методов, применяемых для анализа исследуемого влияния средств массовой информации на финансовый рынок. Кроме того, были описаны и сами такие методы. Теоретические вопросы были рассмотрены подробно, так как их изучение выходит за рамки стандартных курсов по линейной алгебре. Описание эконометрических методов, напротив, было дано более кратко, с акцентом на авторскую модификацию. Про применение вышеописанных подходов на конкретных данных написано в следующей Главе 3.

[21] Тест на наличие ARCH и GARCH-эффектов также будет проводиться для данной спецификации в Главе 3.

Глава 3. Описание данных и построение модели

Настоящая глава является эмпирической составляющей данной работы. Она содержит четыре основные части:

1) Описание данных
2) Обработка данных новостей
3) Эконометрическая модель
4) Обсуждение результатов. Дискуссия

Описание данных

В данной работе использовались два основных источника данных:

1) Сайт *finanz.ru* [22] использовался в качестве источника новостной аналитики (специализируется на финансовых новостях, статьях, аналитике, оценках и прочей финансовой информации).
2) Сайт *yahoo.finance.com* [23] был взят А в качестве источника ежедневных котировок акций крупнейших российских компаний (ПАО) (сайт является одним из крупнейших провайдеров финансовой информации, в том числе и котировок акций ПАО)

Данные были взяты за период с 31 декабря 2019 года до 06 апреля 2020 года (дата обращения к данным ресурсам). Общее количество заголовков новостей составило 12300. Данные по котировкам акций собирались за тот же промежуток времени, что и новости. Таким образом, за вычетом выходных дней, по которым не представлена информация о торгах, общее наблюдений в выборке составило 2080. Были собраны данные по 26 крупнейшим российским ПАО (по 80 наблюдений на каждую компанию). Отдельные наименования всех компаний из выборки и их сфера деятельности представлены в таблице 3:

Таблица 3.

Номер	Название компании	Род деятельности
1	Лента	Торговля
2	Газпром	Нефть и газ
3	Evraz	Черная металлургия
4	Альфа-Банк	Финансы
5	Сургутнефтегаз	Нефть и газ
6	Северсталь	Черная металлургия

[22]Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)

[23]Провайдер финансовой информации// официальный сайт. URL https://finance.yahoo.com/ (Дата обращения 06.04.2020)

7	UC Rusal	Цветная металлургия
8	ПИК	Строительство
9	Новатэк	Нефть и газ
10	Норильский никель	Цветная металлургия
11	НЛМК	Цветная металлургия
12	Мвидео	Торговля
13	МТС	Телекоммуникации
14	Московский кредитный банк	Финансы
15	Мечел	Черная металлургия
16	Магнит	Торговля
17	Mail.ru Group	IT
18	ВТБ	Финансы
19	Лукойл	Нефть и газ
20	ТМК	Черная металлургия
21	Татнефть	Нефть и газ
22	Ютэйр	Транспорт
23	Уралсиб	Финансы
24	Yandex	IT
25	X5 Retail Group	Торговля
26	VEON (Vimpelcom)	Телекоммуникации

Источник: составлено автором

Как видно из таблицы 3, половина компаний относится к добывающей и обрабатывающей промышленностям, остальные компании принадлежит разным отраслям: финансы, IT, торговля и другие.

В связи с большим количеством рассматриваемых компаний, такие временные ряды оказались несбалансированными, в связи с чем общий объем выборки был незначительно подкорректирован от изначального (указанного выше) для балансировки данных временных рядов.

Обработка данных новостей

Для обработки данных новостной аналитики использовался высокоуровневый язык программирования Python. Для заполнения матрицы текста новостей использовалась статистическая мера TF-IDF (об этом подробно написано в Главе 2). Как уже было сказано, размер прямоугольной матрицы составляет $N_{\text{документов}} \times N_{\text{терминов}}$, где:

- $N_{\text{документов}}$ –12300 новостей.
- $N_{\text{терминов}}$ –1000 основных слов, имеющих наибольшую статистическую меру TF-IDF.

Для анализа новостей с их текстом были проведены следующие преобразования: текст был приведен к нижнему регистру, удалены все символы кроме букв русского алфавита, а также удалены стоп-слова (предлоги, союзы, междометия и так далее). Процедуры

лемматизации и стемминга текста новостей, описанные в Главе 2, было решено не проводить во избежание ошибок I и II рода (overstemming, understemming). Также для разных форм одного и того же слова была использована своя кластеризация, в результате которой все формы одного и того же слова попали в один и тот же кластер (topic). Итоговая матрица $A(12300 \times 1000)$ была оценена и разложена с помощью метода SVD:

$$\hat{A} = U\hat{\Sigma}V^* \tag{2.1}$$

Размеры матриц:

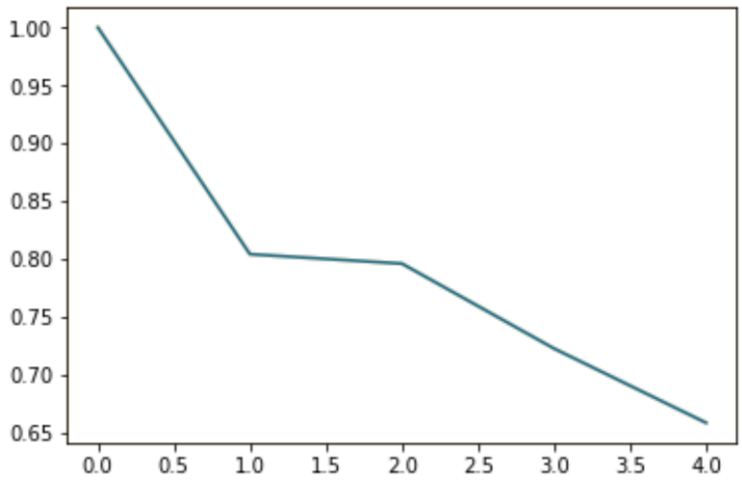
- $\hat{A}(12300 \times 1000)$ – оцененная матрица данных
- $U(12300 \times 5)$ – унитарная матрица новостей, разбитых по 5 основным темам (темы будут описаны далее)
- $\hat{\Sigma} = diag(w_0, \dots, w_4)$ – диагональная матрица с сингулярными числами, такая что:

$$w_0 \geq \cdots \geq w_4 \tag{2.2}$$

- $V^*(5 \times 1000)$ – ортонормированная матрица, сопряженная к V

Для определения размерности квадратной матрицы $\hat{\Sigma}$ (в данном случае был выбран порядок 5) использовался график сингулярных чисел, нормированных на максимальное. Соотношение (2.1) гарантирует, что график будет невозрастающим. Данный график представлен ниже в пространстве $(N; \frac{w_N}{w_0})$

Рисунок 3.



Источник: составлено автором.

Как видно по графику, отсчет начинается с $\frac{w_0}{w_0} = 1$, а затем кривая стремительно убывает, притормаживая между вторым и третьим отношением сингулярных чисел, далее кривая начинает убывать и следующие значения составляют уже становятся достаточно малыми. Ценность SVD модель для аппроксимации состоит в том, что можно обнулить остальные числа, достаточно близкие к нулю, чтобы не хранить лишнюю информацию, и при этом сохранять только самую важную. В силу того, что сингулярные числа находятся на главной диагонали матрицы $\hat{\Sigma}$ в порядке невозрастания, а также в силу наперед заданной размерности этой матрицы, такое разложение единственно, и мы будем однозначно определять интересующие нас 5 тем. Матрица $\hat{\Sigma}$ выглядит следующим образом:

$$\Sigma = diag(15{,}18; 12{,}2; 12{,}08; 10{,}96; 9{,}99) \quad (2.3)$$

Матрица A является разреженной, поэтому для нахождения сингулярных чисел использовались специальные алгоритмы из пакета scipy.sparse.linalg[24].

Результатом SVD-разложения матрицы A стало выделение пяти основных тем финансовых новостей, представленных за данный промежуток времени. Отдельно нужно сказать, что данные темы могут иметь общие слова, например «рубль», «евро», «РФ» и т.д.

Тема 1. Финансовый и фондовый рынки

Данная тема посвящена фондовому и финансовому рынкам, а также (частично) влиянию на них ограничений из-за коронавируса (COVID-19). В следующей таблице выделены содержательные слова с наибольшей TF-IDF мерой из данного топика (темы):

Таблица 4.

Номер	Слово	TF-IDF мера
1	"Индекс"	0,34
2	"Мосбиржа"	0,29
3	"РТС"	0,27
4	"Торги"	0,24
5	"РФ"	0,11
6	"Коронавирус"	0,1
7	"Данные"	0,08
8	"Пункты"	0,08
9	"Биржа"	0,07
10	"ТАСС"	0,07

Источник: составлено автором.

[24] Пакет для Python// Сайт с описанием пакета. URL https://docs.scipy.org/doc/scipy/reference/sparse.linalg.html

Наиболее типичная новость по данной теме: «Индекс Мосбиржи на открытии торгов вырос на 1,6%, индекс РТС на 2,4% - данные торгов».[25]

Помимо этого, в данной теме освещаются котировки акций (в том числе и компании Мечел из выборки), новости, связанные с мессенджером Telegram, транспортом, новости из-за рубежа, объемы торгов, выполнение мер правительства, а также новости, связанные с различными финансовыми показателями. Общее число новостей по данной теме – 755.

Тема 2. ЦБ РФ и экономика

Данная тема, по большей части, посвящена экономике РФ и кредитно-денежной политике, а также мерам, которые были предприняты за данный промежуток времени Центральным Банком Российской Федерации. Самые популярные слова по данной теме в смысле TF-IDF меры представлены в следующей таблице:

Таблица 5.

Номер	Слово	TF-IDF мера
1	"РФ"	0,44
2	"ЦБ"	0,31
3	"Курс"	0,26
4	"Доллар"	0,17
5	"Евро"	0,16
6	"Мосбиржа"	0,08
7	"Повысил"	0,05
8	"Понизил	0,04
9	"Правительство"	0,03
10	"Набиуллина"	0,03
11	"Голикова"	0,03
12	"Экономика"	0,03
13	"Регулятор"	0,02
14	"Ставки"	0,02

Источник: составлено автором.

Как видно из таблицы 5, самое большое число новостей связано с ЦБ РФ, Российской Федерацией, и курсом (валютным и прочим).

Наиболее типичная новость в данной теме выглядит следующим образом: «ЦБ РФ отозвал лицензии у страховщиков «Чувашия-Мед», «Экип» и «Симаз-Мед».»[26]

[25] Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)

[26] Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)

Помимо этого, в данной теме освещаются новости касательно Сбербанка, его руководства, новости правительства, первых лиц, а также новости про компанию Яндекс, новости касательно остальных банков России и некоторых других иностранных банков (например, новости о ЦБ КНР) и другие новости, в большей или меньшей степени, связанные с международными отношениями. Общее число новостей по этой теме – 4695. Данная тема является самой крупной по количеству новостей.

Тема 3. Финансовые операции

Данная тема касается инвестиций, кредитов и прочих денежных сделок как на внутреннем, так и на международном рынке. 10 самых популярных слов из данной темы в таблице ниже:

Таблица 6.

Номер	Слово	TF-IDF мера
1	"Млрд"	0,45
2	"Рубль"	0,3
3	"Прибыль"	0,2
4	"Чистая"	0,16
5	"Выросла"	0,14
6	"МСФО"	0,12
7	"РСБУ"	0,07
8	"Компания"	0,05
9	"Трлн"	0,01
10	"Направят"	0,01

Источник: составлено автором.

Как видно из оценок важности слов в контексте данной темы, большая часть новостей связана с денежными потоками.

Наиболее типичная новость данной темы выглядит следующим образом: «Космические аппараты «Экспресс-80» и «Экспресс-10» застрахованы на 10 млрд рублей каждый»[27].

Помимо этого, к данной теме относятся новости, касающиеся Центральных и Национальных банков других стран, а также их операций на валютном и финансовом рынках. Затрагиваются новости о финансировании и реализации национальных и частных многомиллиардных проектов. В том числе и проектов, связанных со сдерживанием распространения коронавирусной инфекции. Также освещаются новости, касающиеся

[27] Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)

международных стандартов финансовой отчетности (МСФО) и российских стандартов бухгалтерской отчетности (РСБУ). Общее число новостей по данной теме – 1221.

Тема 4. Пандемия

Тема 4 освещает влияние коронавирусной инфекции на мировую экономику. В связи с этим здесь собраны новости, касающиеся действий правительства и ЦБ РФ, прямо или косвенно касающихся действий правительства и ЦБ РФ в отношении предупредительных мероприятия, а также мероприятий, прямо или косвенно связанных с пандемией. Наиболее часто встречающиеся слова по этой темы указаны в таблице ниже вместе со своей статистической мерой:

Таблица 7.

Номер	Слово	TF-IDF мера
1	"РФ"	0,48
2	"ЦБ"	0,2
3	"Коронавирус"	0,2
4	"Россия"	0,12
5	"Нефть"	0,11
6	"США"	0,1
7	"ОПЕК"	0,09
8	"Голикова"	0,08
9	"ТАСС"	0,07
10	"Глава"	0,07
11	"Минфин"	0,05
12	"Коронавирусом"	0,05
13	"Экономики"	0,05

Источник: составлено автором.

Самая типичная новость по этой теме: «Минздрав РФ запустил горячую линию о диспансеризации и профилактических осмотрах».[28]

Помимо этого, в данной теме обсуждаются такие актуальные темы как: нужды стран Африки в борьбе с коронавирусом, меры, направленные на борьбу и предупреждение коронавирусной инфекции, гуманитарная помощь, встреча ОПЕК+, участие стран в этой встрече, переговоры об объемах нефтедобычи, изменение дивидендной политики нефтяных компаний в связи с изменением цен на нефть, падение цен на нефть, негативное влияние цен на нефть на финансовые показатели, пособия по безработице, падение спроса. Данная

[28] Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)

тема является второй по величине (количеству новостей) после темы 2 и имеет 4453 заголовка.

Тема 5. Нефть

Последняя тема имеет 1176 заголовков новостей. По результатам анализа таблицы 8 становится понятно, что она является обобщенной и может включать новости, связанные с предыдущими темами. Наиболее популярные слова:

Таблица 8.

Номер	Слово	TF-IDF мера
1	"Рубль"	0,57
2	"РФ"	0,31
3	"ТАСС"	0,27
4	"Курс"	0,24
5	"Торги"	0,16
6	"Доллар"	0,15
7	"Мосбиржа"	0,14
8	"Данные"	0,13
9	"Прибыль"	0,12
10	"Нефть"	0,11

Источник: составлено автором.

Наиболее типичной новость по данной теме: «ТАСС: ЦБ РФ поднял курс доллара на 19 марта до 77,21 руб., курс евро – до 84,89 руб.».[29]

К данной теме алгоритм отнес большое количество новостей, связанных с колебанием цен на нефть. Например, новость о том, что S&P изменили свои прогнозы относительно российских нефтяных компаний в связи с понижением цен на нефть на «негативные» вместо «стабильных». Также освещаются темы сотрудничества с Китаем и договоренности относительно нефтяных поставок, обсуждается влияние на сырьевые рынки остальных стран-экспортеров нефти вроде Саудовской Аравии. Помимо этого, представлено много новостей относительно изменений (преимущественно снижений) котировок на нефть. Помимо «нефтяной» подтемы здесь также встречаются новости относительно транспорта, туризма и прочего международного сообщение. Также в этой темы представлены новости относительно финансового и фондового рынка (в меньшей степени). Например, что глава «Алросы» продал половину своих акций в компании, а вырученные деньги он направит на борьбу с пандемией.

[29] Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)

Присвоение темы

Для присвоения каждой новости своей темы использовалась косинусная мера-сходство. Подсчет данной меры был реализован следующим образом. В силу введения ограничения на количество используемых слов в размере 1000, каждая новость представляет собой 1000-мерный вектор. Каждая координата новости-документа это TF-IDF меры соответствующих слов из новости. Тогда сходство между двумя документами (новостями) A, B считается по формуле:

$$similarity(A, B) = cos(\varphi) = \frac{(A, B)}{\|A\| * \|B\|} \quad (2.4)$$

Если новости похожи, то угол между ними стремится к нулю, а значит косинус стремится к единице. В частности, косинус равен единице на одном и том же документе. Пусть A - вектор, соответствующий конкретной новости, которой необходимо присвоить тему. Он умножается на каждый из пяти столбцов матрицы $\hat{\Sigma}V^*$. Это и есть мера схожести векторов (2.1). Выбирается наибольшая схожесть с вектором B, соответствующего теме j помощью простого сравнения чисел, а затем присваивается тема j.

Соответственно, как альтернативу, можно использовать расстояние в данном векторном пространстве, которое вводится следующим образом:

$$\rho(A, B) = 1 - similarity(A, B) = 1 - cos(\varphi) = 1 - \frac{(A, B)}{\|A\| * \|B\|} \quad (2.5)$$

Такая метрика пользуется популярностью потому, что, во-первых, достаточно эффективна для разреженных векторов, поскольку необходимо учитывать только ненулевые координаты, а во-вторых, является нормированной и лежит между нулем и единицей. Также существуют некоторые обобщения данной меры сходства, например, «мягкая» (soft) косинусная мера. Данное же исследование ограничивается только первым вариантом.

Сводное описание выделенных тем представлено ниже в таблице:

Таблица 9

Номер	**Тема**	**Число новостей**	**Примеры новостей**
1	Финансовый и фондовый рынки	755	«НМТП может выплачивать дивиденды в размере 50% от чистой прибыли по МСФО до 2029 г.»[30] «Индекс Мосбиржи на открытии торгов вырос на 0,62%, РТС - на 0,84%»[31]
2	ЦБ РФ и экономика	4695	«Банки смогут снизить ставку для МСП до 8% годовых по новой программе - власти Москвы»[32] «ЦБ предложил ввести новые условия для выявления отмывания денег через НКО»[33]
3	Финансовые операции	1221	«Расходы бюджета РФ вырастут на 163 млрд руб. в 2020 г., доходы - на 214 млрд руб. – Минфин»[34] «В Новосибирской области на 20% увеличат финансирование медпомощи по ОМС в 2020 г.»[35]
4	Пандемия	4453	«В России с подозрением на коронавирус изолированы 100 человек»[36] «Цены на нефть могут упасть на $3 из-за распространения коронавируса – Reuters»[37]

[30] Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)

[31] Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)

[32] Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)

[33] Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)

[34] Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)

[35] Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)

[36] Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)

[37] Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)

5	Нефть	1176	«Корпорация МСП оказала поддержку компании из Москвы, выпускающей системы для отбора нефти»[38] «Оператор польского участка «Дружбы» запустил обновленную систему контроля качества нефти»[39]

Источник: составлено автором

По результатам таблицы 9 видно, что самыми актуальными темами в первом квартале 2020 года были вторая и четвертая. Наибольшее количество новостей посвящено темам ЦБ РФ, пандемии, а также темам, связанным с экономикой РФ.

Эконометрическая модель

Второй этап данного исследования подразумевает построение эконометрической модели, как можно лучше описывающей динамику временных рядов по имеющейся выборке из генеральной совокупности. Как известно, в эконометрике существует целая теория, изучающие временные ряды. В результате проведенного в Главе 1 сопоставительного анализа, выбор был сделан в пользу моделей семейства ARCH и GARCH после тестирования гипотез на наличие ARCH процессов[40]. В качестве основной регрессии на котировки выступает модель ARIMA. Такой выбор был сделан по трем причинам: такие модели достаточно хорошо описывают поведение подобных временных рядов, относительно просты в интерпретации, а также имеют преимущества перед другими моделями в результате сравнения по формальным критериям (например, Байесовский критерий). Отдельно нужно отметить, что использование для моделирования цен акций ARIMA-моделей облегчает интерпретацию результатов, поскольку первая разность цен акций в этом случае представляет собой их доходность в абсолютном выражении. Для того чтобы получить доходность акций достаточно эту первую разность разделить на предыдущее значение акции. Такие модели стали популярными в литературе по финансовым и фондовым рынкам еще в конце XX века, поэтому данные предпосылки весьма типичны в подобных исследованиях. Перейдем к построению основной модели. Обозначим за y_t котировку акции некоторой компании в момент времени t . Тогда введем следующую предпосылку:

[38] Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)

[39] Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)

[40] Результаты тестирования представлены в таблице 13

$$y_t \sim ARIMA(p, d, q) \tag{2.6}$$

Это означает:

$$\Delta^d y_t = \phi_0 + \sum_{j=1}^{p} \phi_j * \Delta^d y_{t-j} + \sum_{k=1}^{q} \theta_k * \varepsilon_{t-k} + \varepsilon_t \tag{2.7}$$

Здесь ошибки модели описываются ARCH или GARCH-моделью. Например, ARCH(1) имеет вид:

$$\varepsilon_t = u_t \cdot \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2} \tag{2.8}$$

Где u_t – белый шум, ε_{t-1} – лаг ошибки из модели (2.7)

Тогда регрессия на условную дисперсию имеет вид:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \tag{2.9}$$

В данном исследовании предполагается использовать модификацию классической модели, описанную в Главе 2. В данной модификации (обозначим ее как модель I) условная дисперсия будет оцениваться с помощью следующей регрессии:

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{n_1} \alpha_j \varepsilon_{t-j}^2 + \sum_{k=1}^{n_2} \beta_k \sigma_{t-k}^2 + \sum_{r=1}^{5} \sum_{l=0}^{n_4} \gamma_{rl} \cdot topic_{r,t-l} \tag{2.10}$$

Здесь:

- σ_t^2- условная дисперсия ошибок ε_t, зависящая от времени t
- $topic_{k,l}$- число новостей темы k в день l

Для проверки устойчивости результатов предложенной модели в работе были построены две альтернативные модели II и III. Модель II (2.11) учитывает только общий поток новостей во вспомогательной регрессии:

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{n_1} \alpha_j \varepsilon_{t-j}^2 + \sum_{k=1}^{n_2} \beta_k \sigma_{t-k}^2 + \sum_{l=0}^{n_4} \gamma_l \cdot number\_of\_news_{t-l} \tag{2.11}$$

Модель III (2.12) подразумевает включение переменных новостной аналитики не во вспомогательную, как в (2.10), а в основную регрессию (2.7):

$$\Delta^d y_t = \phi_0 + \sum_{j=1}^{p} \phi_j * \Delta^d y_{t-j} + \sum_{k=1}^{q} \theta_k * \varepsilon_{t-k} + \sum_{r=1}^{n_3} \sum_{l=0}^{n_4} \gamma_{rl} \cdot topic_{r,t-l} + \varepsilon_t \quad (2.12)$$

Сформулируем гипотезу 1 в терминах построенных моделей. Нулевая гипотеза для моделей I и III с числом новостей по пяти темам:

$$H_0: \quad \forall\, r, l \quad \gamma_{rl} = 0 \quad (2.13)$$

Нулевая гипотеза для модели II с общим потоком новостей:

$$H_0: \quad \forall\, r \quad \gamma_r = 0 \quad (2.14)$$

Альтернативная гипотеза для моделей I и III имеет вид:

$$H_1: \quad \exists r, l \quad \gamma_{rl} \neq 0 \quad (2.15)$$

Для модели II альтернативная гипотеза H_1 формулируется аналогичным образом.

Альтернативная гипотеза H_1 означает, что нулевая гипотеза не принимается и хотя бы одна «новостная» переменная значимо влияет на волатильность. В случае, если принимается H_1 во всех трех моделях, взаимосвязь между поступающими новостями и волатильностью может быть проинтерпретирована каузально.

Для того чтобы протестировать данную гипотезу в моделях I и II осуществляется следующая процедура:

1) Подбирается модель ARIMA для y_t
2) Сохраняются остатки $e_1, \dots, e_T$ этой модели
3) $\widehat{\sigma_t^2} = e_t^2$. Затем оцениваются сами модели I и II:
4) Делается тест «короткая» против «длинной» с известной расчетной статистикой.

В случае, если гипотеза из теста (пункт 4) не принимается, то исследовательский вопрос о значимом влиянии новостей на стоимость акции соответствующего подтверждается ПАО.

Данная процедура проделывается для каждой из 26 котировок, а затем делается общий вывод относительно связи СМИ и фондового рынка России.

Чтобы протестировать данную гипотезу в модели III, проделывается следующая процедура:

1) Тестируется стационарность временных рядов «новостей»
2) Если ряды из пункта (1) не стационарны, то переходим к первым разностям. Если стационарны, то переходим к пункту (3)
3) Оценивается модель ADL (ARIMA плюс набор «новостных» переменных в основной регрессии) с ошибками из GARCH(1,1)[41]-модели.
4) Тестируется значимость переменных, отвечающих за новостной поток с помощью теста на линейные ограничения.

Подбор моделей ARIMA

После выбора модели начинается задача подбора параметров. В первую очередь подбирается порядок интегрированности. Для тестирования стационарности временного ряда используются формальные тесты Дики-Фуллера (ADF-test) и KPSS-тест с нулевой гипотезой о наличии единичного корня в соответствующем характеристическом уравнении. Затем определяются параметры p, q , отвечающие за максимальный порядок лага в авторегрессии и скользящем среднем соответственно. Выбираются они, во-первых, исходя из анализа графиков автокорреляционных и частных функций, а затем используется некоторый формальный критерий. Например, критерий Шварца или критерий Акаике. Пусть SIC – критерий Шварца в модели I (2.7), тогда подбор параметров выбирается из решения формальной задачи целочисленной[42] оптимизации:

$$SIC \to \min(p, d, q) \tag{2.16}$$

Также рекомендовано использовать модели, в которых выполняется условие (условная минимизация):

$$p + q \leq 3 \tag{2.17}$$

Это ограничение нужно для того, чтобы модель не была перенасыщена параметрами, которые нужно оценивать. В машинном обучении модели с избытком параметров, хорошо

[41] Наиболее часто применимы на практике именно GARCH(1,1) особенно в финансовой литературе. Например, (Maheu, Mccurdy, 2011), (Palfrey, Wang, 2012) и другие.

[42] Вообще говоря, порядок интегрированности d может быть дробным. Такие модели называются ARFIMA. О них можно почитать, например, в *Robinson P. Time Series With Long Memory // Oxford University Press, 2003*

описывающие конкретный набор данных, но плохо работающие вне выборки (out of sample) называют «переобученными».

Результаты оптимизационной задачи (2.16) для каждой из 26-ти компаний представлены в таблице 10:

Таблица 10.

Номер	Название компании	ARIMA(p,d,q)	Дрейф (наличие константы)
1	Лента	(1,1,0)	Нет
2	Газпром	(0,1,0)	Да
3	Evraz	(0,1,0)	Да
4	Альфа-Банк	(1,1,2)	Да
5	Сургутнефтегаз	(0,1,0)	Нет
6	Северсталь	(0,1,1)	Нет
7	UC Rusal	(0,1,0)	Нет
8	ПИК	(1,1,0)	Нет
9	Новатэк	(1,1,0)	Нет
10	Норильский никель	(0,1,0)	Нет
11	НЛМК	(0,1,1)	Нет
12	Мвидео	(0,1,0)	Да
13	МТС	(1,1,0)	Нет
14	Московский кредитный банк	(0,1,0)	Нет
15	Мечел	(0,1,0)	Нет
16	Магнит	(0,1,0)	Нет
17	Mail.ru Group	(0,1,0)	Нет
18	ВТБ	(0,1,0)	Нет
19	Лукойл	(0,1,0)	Нет
20	ТМК	(0,1,0)	Нет
21	Татнефть	(0,1,0)	Нет
22	Ютэйр	(1,1,0)	Нет
23	Уралсиб	(0,1,1)	Нет
24	Yandex	(0,1,0)	Нет
25	X5 Retail Group	(0,1,0)	Нет
26	VEON (Vimpelcom)	(0,1,0)	Да

Источник: составлено автором.

Как видно из таблицы, котировки акций 6 компаний описываются авторегрессией первого порядка со скользящим средним и без него. Котировки акций еще 17 компаний описываются процессом случайного блуждания, из них 4 с дрейфом (дисперсия пропорциональна времени t, расходится). Оставшиеся 3 временных ряда описываются скользящим средним первого порядка после перехода к стационарным разностям. Динамика акций некоторых компаний из таблицы представлена в приложении Б. Тест на наличие ARCH эффектов был проведен в модели III (см. таблицу 13)

В данной работе используется теория о краткосрочном влиянии новостей из работы (Tetlock, 2007). На графике ниже показан механизм возможного влияния:

Рисунок 4.

Источник: составлено автором.

Как видно из рисунка 4, в краткосрочном периоде новости могут оказывать негативное влияние (понижающие давление) на котировки акций. Спустя короткий промежуток времени $|t_2 - t_1| \leq 5$ дней (в работе (Tetlock, 2007) он составляет 5 дней) данный показатель возвращается к своему фундаментальному ожидаемому значению.

К сожалению, нет общих рекомендаций относительно максимального порядка лага для переменной «новостной аналитики», но есть общие экономические соображения, из которых такой параметр может выбираться. Во-первых, согласно общим эконометрическим рекомендациям, нужно по возможности использовать «экономичные» модели, чтобы не возникало избытка параметров. Во-вторых, в силу того, что котировки цен акций фиксируются за пять рабочих дней из семи календарных, новости выходных дней могут находить отклик, например, в понедельник. В связи с этим, максимальный порядок лага для показателей можно выбирать из расчета «не больше двух». Если сократить число переменных (как в случае модели II), то можно полноценно использовать рекомендацию из статьи Тетлока (Tetlock, 2007) - «не больше пяти».

Модель I

Напомним, что суть данной модели заключается в оценке «непосредственного» влияния потока тематических новостей на волатильность цен акций крупнейших 26 российских ПАО:

$$\widehat{\sigma_t^2} = \alpha_0 + \alpha_1 \widehat{\sigma_{t-1}^2} + \sum_{r=1}^{5} \sum_{l=1}^{2} \gamma_{rl} \cdot topic_{r,t-l} \tag{2.18}$$

На основе тестирования равенств $\gamma_{rl} = 0$ были выделены компании, для которых подтвердилось значимое влияние тематического новостного потока. Ниже в таблице 11 приведены сводные результаты тестирования для всех компаний. Синим закрашены те ячейки последнего столбца, в котором подтвердилось значимое влияние новостного потока на волатильность.

Таблица 11.

Номер	**Название компании**	**Род деятельности**	**Значимы ли новости на уровне 5%**
1	Лента	Торговля	-
2	Газпром	Нефть и газ	+
3	Evraz	Черная металлургия	-
4	Альфа-Банк	Финансы	-
5	Сургутнефтегаз	Нефть и газ	+
6	Северсталь	Черная металлургия	+
7	UC Rusal	Цветная металлургия	-
8	ПИК	Строительство	+
9	Новатэк	Нефть и газ	+
10	Норильский никель	Цветная металлургия	+
11	НЛМК	Цветная металлургия	+
12	Мвидео	Торговля	+
13	МТС	Телекоммуникации	-
14	Московский кредитный банк	Финансы	+
15	Мечел	Черная металлургия	-
16	Магнит	Торговля	+
17	Mail.ru Group	IT	+
18	ВТБ	Финансы	+
19	Лукойл	Нефть и газ	-
20	ТМК	Черная металлургия	+
21	Татнефть	Нефть и газ	-
22	Ютэйр	Транспорт	-
23	Уралсиб	Финансы	+

24	Yandex	IT	+
25	X5 Retail Group	Торговля	+
26	VEON (Vimpelcom)	Телекоммуникации	+

Источник: составлено автором.

Итого: для 17 компаний поток тематических новостей значимо влияет на волатильность цен их акций. Для 9 компаний (5 из которых принадлежит тяжелой промышленности) такое влияние оказалось незначимо в разрезе двух дней. Интересным развитием работы стала бы проверка более поздних откликов на тематические новости, для чего необходимо собрать данные за более длинный период

По данной модели в большинстве случаев выявлено значимое влияние новостей. Можно сделать предварительные выводы об увеличении предсказательной способности моделей условной гетероскедастичности за счет добавлений переменных новостной аналитики.

Модель II

Следующим этапом является оценка модели II (2.11). Результаты представлены в таблице 12. Такая модель, как уже говорилось ранее, не учитывает тематику новостей, а лишь общий новостной поток. Зато данная модель позволяет использовать в спецификации большой порядок «лага». В соответствии с вышеизложенной теорией максимальный порядок лага равен пяти. Авторегрессия у условной дисперсии имеет порядок 1.

Таблица 12.

Номер	**Название компании**	**ARIMA(p,d,q)**	**Род деятельности**	**Значимы ли новости на уровне 5%**
1	Лента	(1,1,0)	Торговля	-
2	Газпром	(0,1,0)	Нефть и газ	+
3	Evraz	(0,1,0)	Черная металлургия	+
4	Альфа-Банк	(1,1,2)	Финансы	+
5	Сургутнефтегаз	(0,1,0)	Нефть и газ	+
6	Северсталь	(0,1,1)	Черная металлургия	+
7	UC Rusal	(0,1,0)	Цветная металлургия	+
8	ПИК	(1,1,0)	Строительство	+
9	Новатэк	(1,1,0)	Нефть и газ	+
10	Норильский никель	(0,1,0)	Цветная металлургия	+

11	НЛМК	(0,1,1)	Цветная металлургия	+
12	Мвидео	(0,1,0)	Торговля	+
13	МТС	(1,1,0)	Телекоммуникации	+
14	Московский кредитный банк	(0,1,0)	Финансы	+
15	Мечел	(0,1,0)	Черная металлургия	+
16	Магнит	(0,1,0)	Торговля	+
17	Mail.ru Group	(0,1,0)	IT	-
18	ВТБ	(0,1,0)	Финансы	+
19	Лукойл	(0,1,0)	Нефть и газ	+
20	ТМК	(0,1,0)	Черная металлургия	-
21	Татнефть	(0,1,0)	Нефть и газ	-
22	Ютэйр	(1,1,0)	Транспорт	+
23	Уралсиб	(0,1,1)	Финансы	+
24	Yandex	(0,1,0)	IT	+
25	X5 Retail Group	(0,1,0)	Торговля	+
26	VEON (Vimpelcom)	(0,1,0)	Телекоммуникации	+

Источник: составлено автором.

Как можно видеть по таблице 12, данная модель показывает следующие результаты: у 22 компаний подтвердилось значимое влияние новостного потока на условную волатильность показателя. У четырех компаний не подтвердилось влияние (2 из которых принадлежит тяжелой промышленности). Подобная модель использовалась в (Балаш, Сидоров, Дате). В их выборке значимость на уровне 5% подтвердилась у 19 из 20 рассматриваемых компаний.

Как и для модели I, можно сделать предварительный вывод о том, что добавление новостных переменных увеличивает объяснительную способность GARCH-моделей.

Модель III

Как уже описывалось ранее, суть этой модели (2.12) заключается в том, что переменные, отвечающие за новостные потоки, будут включены не во вспомогательную, а в основную регрессию. Но перед этим необходимо проверить все такие переменные на стационарность, поскольку если они не стационарны, то не будет выполнена предпосылка, и значит теорема об эффективности линейных оценок, вообще говоря, выполнена не будет. Результаты тестирования на стационарность представлены в следующей таблице:

Таблица 13.

Название темы	Соответствующая переменная стационарна[43]	
	ADF-тест	KPSS-тест
Финансовый и фондовый рынки	ДА	ДА
ЦБ РФ и экономика	ДА	ДА
Финансовые операции	ДА	ДА($\alpha = 5\%$)
Пандемия	ДА	ДА
Нефть	ДА	ДА

Источник: составлено автором

Как можно убедиться из таблицы 13, все переменные являются стационарными на уровне 5%. Это означает, что можно добавлять их в основную регрессию. Перейдем к непосредственной оценке модели (2.12) и тестированию гипотезы (2.13). Результаты удобно представить в таблице 14. Для наглядности, в пятом столбце желтым закрашены те ячейки, когда GARCH-эффект оказался незначимым. В последнем шестом столбце, по аналогии с таблицами 10 и 11, синим закрашены те ячейки, у которых искомое влияние новостей оказалось значимым.

Таблица 14.

Номер	Название компании	Порядок авторегрессионного лага для котировок акций	Род деятельности	P-значение для GARCH-эффектов	Значимы ли новости на уровне 5%
1	Лента	1	Торговля	0,00	+
2	Газпром	0	Нефть и газ	0,02	+
3	Evraz	0	Черная металлургия	0,00	+
4	Альфа-Банк	1	Финансы	0,54	+
5	Сургутнефтегаз	0	Нефть и газ	-	-
6	Северсталь	0	Черная металлургия	0,02	+
7	UC Rusal	0	Цветная металлургия	0,00	+
8	ПИК	1	Строительство	0,00	+
9	Новатэк	1	Нефть и газ	0,00	-
10	Норильский никель	0	Цветная металлургия	0,00	+
11	НЛМК	0	Цветная металлургия	0,00	+

[43] Уровень значимости $\alpha = 1\%$

12	Мвидео	0	Торговля	0,03	+
13	МТС	1	Телекоммуникации	0,35	+
14	Московский кредитный банк	0	Финансы	0,00	+
15	Мечел	0	Черная металлургия	0,00	-
16	Магнит	0	Торговля	-	+
17	Mail.ru Group	0	IT	0,03	+
18	ВТБ	0	Финансы	0,39	+
19	Лукойл	0	Нефть и газ	0,00	-
20	ТМК	0	Черная металлургия	0,00	+
21	Татнефть	0	Нефть и газ	0,00	+
22	Ютэйр	1	Транспорт	0,79	-
23	Уралсиб	0	Финансы	0,41	+
24	Yandex	0	IT	0,10	+
25	X5 Retail Group	0	Торговля	0,01	-
26	VEON (Vimpelcom)	0	Телекоммуникации	0,00	+

Источник: составлено автором

Данная спецификация показала следующие результаты: значимое влияние тематического новостного потока обнаружилась у 20 компаний. Для 6 компаний такая взаимосвязь не обнаружена. Отметим, что повышение уровня значимости до 10% приводит к значимости новостей для котировок еще двух ПАО: «Сургутнефтегаз» и «Мечел».

Обсуждение результатов. Дискуссия

По результатам применения трех основных спецификаций эконометрических моделей получились «похожие» результаты. Для удобства сравнения они представлены в таблице 15:

Таблица 15.

Название компании	Род деятельности	Значимы ли GARCH эффекты	Значимы ли новости на уровне 5%			Общее число плюсов	Средняя дневная доходность по группам	Дисперсия средней доходности
			Модель I	Модель II	Модель III			
Газпром	Нефть и газ	+	+	+	+	3	-0,26%	3,05%
Северсталь	Черная металлургия	+	+	+	+	3		
ПИК	Строительство	+	+	+	+	3		
Норильский никель	Цветная металлургия	+	+	+	+	3		
НЛМК	Цветная металлургия	+	+	+	+	3		
Мвидео	Торговля	+	+	+	+	3		
Московский кредитный банк	Финансы	+	+	+	+	3		
Магнит	Торговля	-	+	+	+	3		
ВТБ	Финансы	-	+	+	+	3		
Уралсиб	Финансы	-	+	+	+	3		
Yandex	IT	+	+	+	+	3		
VEON (Vimpelcom)	Телекоммуникации	+	+	+	+	3		
Evraz	Черная металлургия	+	-	+	+	2	-0,59%	4,93%
Альфа-Банк	Финансы	-	-	+	+	2		
Сургутнефтегаз	Нефть и газ	-	+	+	-	2		
UC Rusal	Цветная металлургия	+	-	+	+	2		
Новатэк	Нефть и газ	+	+	+	-	2		
МТС	Телекоммуникации	-	-	+	+	2		
Mail.ru Group	IT	+	+	-	+	2		
ТМК	Черная металлургия	+	+	-	+	2		
X5 Retail Group	Торговля	+	+	+	-	2		
Лента	Торговля	+	-	-	+	1	-0,26%	4,47%
Мечел	Черная металлургия	+	-	+	-	1		
Лукойл	Нефть и газ	+	-	+	-	1		
Татнефть	Нефть и газ	+	-	-	+	1		
Ютэйр	Транспорт	-	-	+	-	1		

Источник: составлено автором

Как видно из таблицы 15, для двадцати одной российской компании из выборки получен устойчивый результат (не менее двух плюсов «+») о значимом влиянии тематических новостей и общего новостного потока на котировки их акций. Среди них компании из совершенно разных отраслей: от нефтеперерабатывающей до финансовой. Это свидетельствует о том, что новости действительно могут оказывать влияние на котировки ПАО РФ из различных отраслей, вне зависимости от отрасли и рода их деятельности. Как следствие, новостные потоки могут оказывать значимое влияние на российский фондовый рынок в целом. Для оставшихся пяти компаний только по результатам одной из моделей такое влияние СМИ было выявлено. К этим компаниям относятся: Лента, Мечел, Лукойл, Татнефть и Ютэйр. Как можно видеть из графика в приложении Б, у данных компаний действительно замечена слабая волатильность котировок, кроме ПАО «Татнефть». По результатам применения модели III для этой компании, было выявлено, что на уровне значимости 5%, переменные, отражающие количество новостей «вчера» и «позавчера» (1-й и 2-й лаги) на следующие темы: *экономика и ЦБ (тема 2), финансовые операции (тема 3)* и *нефть (тема 5)* оказались значимыми. Данные результаты являются интуитивно понятными и легко интерпретируемыми, если учитывать род деятельности компании «Татнефть»: на компанию из нефтеперерабатывающей отрасли значимо оказывают новости об общей конъюнктуре рынка, а также новости о котировках цен на нефть. В связи с этим, из содержательных соображений, для данной компании можно остановиться на модели III.

Если модель I является стартовой и демонстрирует относительно слабые результаты, то остальные две модели могут быть обе применимы для более тонкого анализа, например, для краткосрочного[44] прогнозирования или оценки влияния новостей на ту или иную тему для конкретной компании, или рынка в целом.

В результате обзора литературы на данную тематику было обнаружено, что подобных исследований по российским данным очень мало. Таким образом, в России данная область исследований только начинает развиваться, в то время как на американских рынках первые такие исследования начали проводиться еще в 90-х годах XX века, и с тех пор набирают популярность. Об этом свидетельствует большое количество новых исследований, и их большая цитируемость. Выводом данного исследования стало то, что включение переменных, отвечающих за новости, увеличивает объяснительную силу моделей GARCH, причем лучшей моделью оказалась модель с включенными переменными общего новостного потока и их лагами. Модель II выявляет значимое влияние СМИ на наибольшее количество котировок акций торгующихся российских фирм. Похожие

[44] Известно, что для подобных моделей долгосрочное прогнозирование неприменимо в силу очень быстрорастущего доверительного интервала.

исследования, например (Sanjiv, Ras, 2011) приходят к аналогичному выводу: действительно, новостная аналитика помогает лучше описать волатильность рынка (модель II) и доходности рынка (модель III). Еще одно исследование (Ahern, Sosyura, 2014), хотя и использует другие методы, приходит к тому же выводу в результате анализа рынка акций крупных американских публичный акционерных обществ за период с 2000 по 2008 год.

Заключение

В Главе 1 был освещен вопрос влияния новостей на финансовые рынки разных стран. В остальных главах такое влияние было исследовано для российского фондового рынка. Было подтверждено значимое влияние средств массовой информации на российский фондовый рынок на примере крупнейших 26 публичных акционерных обществ. В исследовании был использован смешанный подход, основанный на методах машинного обучения, включая эконометрические модели.

Гипотеза 1 о влиянии СМИ на волатильность российского рынка акций *подтвердилась*: по результатам исследования было обнаружено устойчивое влияние новостного потока на волатильность котировок акций исследуемых компаний. Это означает, что новости могут служить причиной повышенных колебаний котировок акций. Важный вывод заключается в том, что новостная аналитика должна находиться в инструментарии каждого инвестора, поскольку учет новостного фактора, как демонстрирует одна из моделей исследования, поможет предсказать колебания цен в краткосрочном периоде.

Модель II подтверждает, что реакция рынка может сохраняться до пяти дней. Это означает, что *подтверждается Гипотеза 2*.

Все три построенных в работе модели модели не выявили закономерности между сферой деятельности компании и значимостью влияния новостей на ее котировки. Отсюда следует, что *Гипотеза 3* о том, что влияние новостей на котировки акций компании зависит от отрасли, в которой она функционирует, не принимается.

Одним из возможных направлений дальнейшей работы может быть более тонкий анализ влияния тематических новостных потоков на котировки акций. В частности, можно решать задачу краткосрочного прогнозирования котировок акций конкретной компании на основе анализа влияния новостей по определенным темам, связанным с деятельностью именно этой компании. Другим возможным направлением развития работы является анализ долгосрочного влияния новостей (более 5 дней), например в работе (Ahern, Sosyura, 2014) исследуются поздние отклики фондового рынка в ответ на всплеск новостей длиной до 120 часов (5 дней или 1 рабочая неделя). Для такого анализа необходимы данные за большие промежутки времени, и поэтому подобные исследования требуют мощного технического оснащения. В данной работе для наблюдений был выбран трехмесячный период, и это привело к выборке из 12300 новостей. При анализе более длительных временных периодов (например, длиной в несколько лет) число новостей возрастет на порядок. Разложение матриц таких размеров, несмотря на их разреженность, требует больших вычислительных способностей, например, средний ноутбук будет обрабатывать такие данные несколько

дней. Подобные исследования проведены в работе Тетлока (Tetlock, 2007), в которой автор анализирует данные за 80-е и 90-е годы. В своем исследовании он использовал ПО Гарварда.

Еще одним разумным развитием работы, на взгляд автора, является разделение новостей на *позитивные* и *негативные*, как это сделано в ряде работ из обзора литературы.

Для анализа тематических новостей и их влияния на финансовые рынки используются все более новые методы, на эту тему публикуется все большее количество исследований. Есть все предпосылки для того, что новостная аналитика станет одним из самостоятельных разделов финансовой аналитики, поскольку подавляющее большинство исследований демонстрирует устойчивое влияние новостей не только на фондовые, но и на финансовые рынки в целом (Melvin, Yin, 2000), (Dzielinski, Rieger, Talpsepp, 2011). Также исследования в данной области порождают новые теории и интересные факты. Например, Ahern и Sosyura исследуют двустороннее взаимодействие, а именно формулируют новую гипотезу о том, что ПАО могут намеренно делать всплески новостей для влияния на рынок с целью спекулятивных действий с собственными акциями. Для проверки подобной гипотезы необходимо включать в анализ не только данные о доходности компаний, но и проводить качественный анализ основных новостей и их влияния на различные отрасли. Все вышеперечисленное говорит об актуальности дальнейших исследований по данной теме.

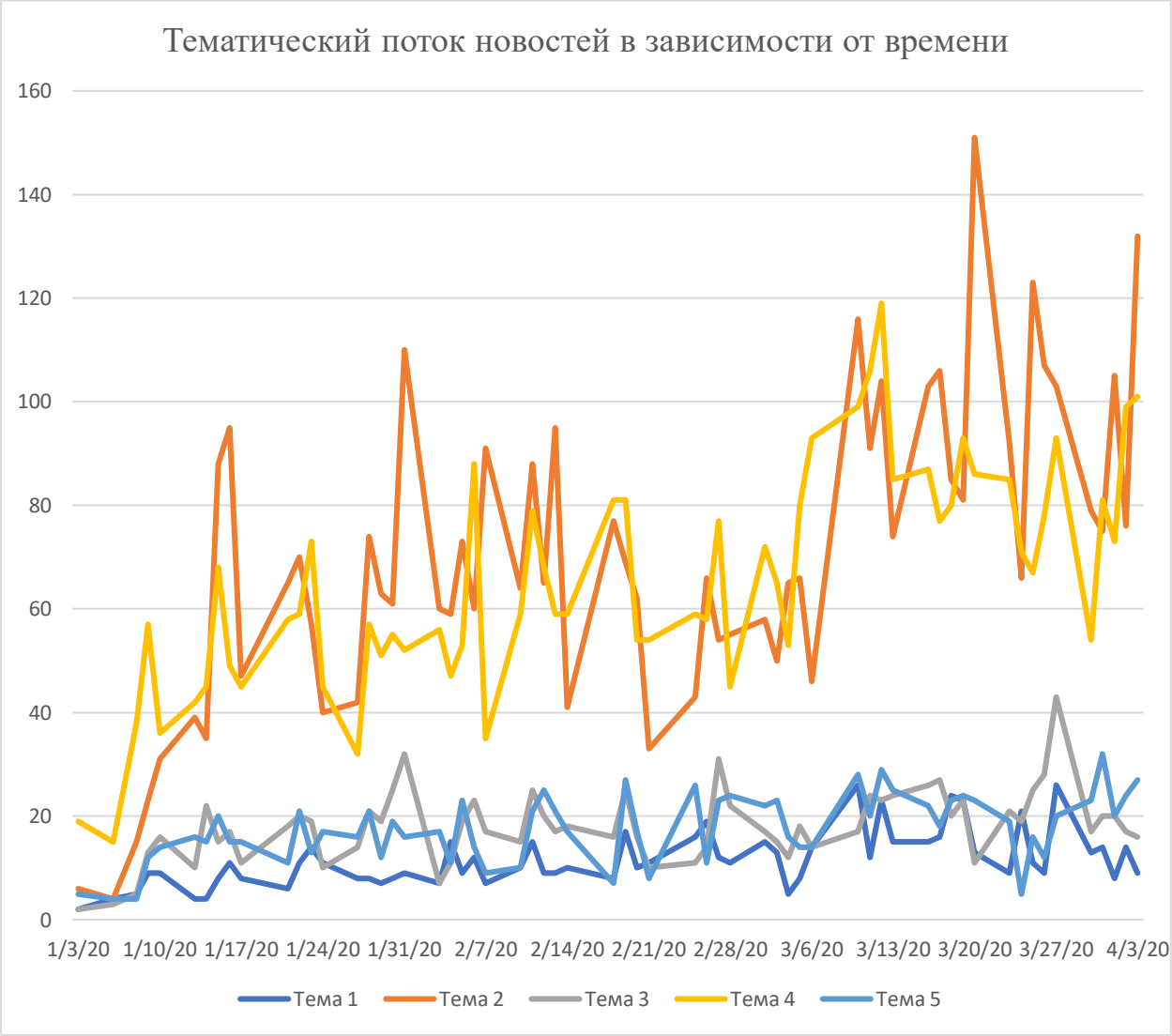
Список использованных источников и литературы

1) Балаев А. Моделирование многомерных параметрических плотностей финансовых доходностей // Квантиль №9. – 2011. – стр. 39-75
2) Белоусов С. «Моделирование волатильности со скачками: применение к российскому и американскому фондовым рынкам» // Квантиль №1. – 2006. – стр. 101-110
3) Крицкий О., Лисок Е. Асимптотическое оценивание коэффициентов модели стохастической волатильности» // Прикладная эконометрика №2(6). - 2007. – стр. 3-12
4) Мельников Р. Влияние динамики цен на нефть на макроэкономические показатели российской экономики // Прикладная эконометрика №1(17). - 2010. – стр. 20-29
5) Сидоров С., Дате П., Балаш В. Использование данных новостной аналитики в GARCH моделях // Прикладная эконометрика №29(1). - 2013. – стр. 82-96
6) Paul C. Tetlock «Giving Content To Investor Sentiment: The Role Of Media In The Stock Market» // The Journal Of Finance Vol. Lxii, - June 2007 P. 1139-1168
7) Mitra L., Mitra G. Applications of news analytics in finance: A review // The Handbook of News analytics in finance. – 2011. – p. 1-36
8) Sanjiv R. News analytics: Framework, techniques, and metrics» // The Handbook of News analytics in finance. – 2011. – p. 41-69
9) P. Ager Hafez, How news events impact market sentiment // The Handbook of News analytics in finance. – 2011. – p. 129-145
10) Dzielinski M., Rieger M., Talpsepp T. How news events impact market sentiment // The Handbook of News analytics in finance. – 2011. – p. 255-269
11) Kalev S., Nhan Duong H. Firm-specific news arrival and the volatility of intraday stock index and futures returns // The Handbook of News analytics in finance. – 2011. – p. 271-286
12) Yu-Pin Hu, Tsay R.S. Principal Volatility Component Analysis // Journal of Business &Economic Statistics. - Vol. 32 No. 2, - 2014. - p. 153-164
13) Palfrey T.R., Wang S.W. Speculative Overpricing In Asset Markets With Information Flows // Econometrica. - Vol. 80, No. 5. – 2012. - p. 1937–1976
14) Maheu J.M., Mccurdy T.H. News Arrival, Jump Dynamics, and Volatility Components for Individual Stock Returns // The Journal of Finance. – Vol. No. 2, - 2004. - p.755-703
15) M.Melvin, Xixi Yin, 2000 «Public Information Arrival, Exchange Rate Volatility, and Quote Frequency» The Economic Journal, Vol. 110, No. 465 (Jul., 2000), pp. 644-661

16) Nofsinger J.R. The impact of public information on investors // Journal of Banking & Finance No 25. – 2001. - p.1339-1366
17) Ahern K. R., Sosyura D. Who Writes the News? Corporate Press Releases during Merger Negotiations // The Journal of Finance Vol. LXIX, No. 1. - 2014. – p.241-291
18) Wang, D. Adjustable Robust Singular Value Decomposition: Design, Analysis and Application to Finance // Institute for Financial Services Analytics, University of Delaware, Newark. – 2017
19) Binder A. Jadhav O. Mehrmann V. Model order reduction for parametric high dimensional models in the analysis of financial risk // European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant. – 2020.
20) Wu X., Kumar V. etc. Top 10 algorithms in data mining // Knowl Inf Syst. – 2008. – p.1-37
21) Sidorov G., Gelbukh A., Gómez-Adorno H., Pinto D. Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model // Computación y Sistemas.- 2014.- vol.18(3).- p.491–504.
22) Robinson P. Time Series With Long Memory // Oxford University Press, 2003
23) Портал о личных финансах и частных инвестициях // официальный сайт. URL https://www.finanz.ru/novosti (Дата обращения 06.04.2020)
24) Провайдер финансовой информации// официальный сайт. URL https://finance.yahoo.com/ (Дата обращения 06.04.2020)
25) Поисковая система по полным текстам научных публикаций// официальный сайт. URL https://google.scholar.com/ (Дата обращения 06.04.2020)
26) Федеральная Служба Государственной Статистики // официальный сайт. URL https://www.gks.ru/ (Дата обращения 06.04.2020)

Приложение

Приложение А



Приложение Б