МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ЭКОНОМИЧЕСКИЙ ФАКУЛЬТЕТ

Кафедра экономической информатики

РУКША Екатерина Геннадьевна

оценка влияния экономических новостей на валютный рынок методами машинного обучения

Магистерская диссертация

специальность 1-25 81 10 «Экономическая информатика»

Научный руководитель
Паньшин Борис Николаевич
доктор технических наук,
профессор

Допущена к защите
«___» _____ 2018 г.
Зав. кафедрой экономической информатики
_____ Д.А. Марушко
кандидат экономических наук, доцент

Минск, 2019

ОГЛАВЛЕНИЕ

ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ

ANN	Artificial neural network, или искусственная нейронная сеть. Это многопараметрическая модель нелинейной оптимизации.
CBOW	Continuous bag of words, или «непрерывный мешок со словами». Это модельная архитектура, которая предсказывает текущее слово, исходя из окружающего его контекста
CNN	Convolutional neural network, или сверточная нейронная сеть. Это архитектура искусственных нейронных сетей, позволяющая работать с многомерными данными (изображения, текст).
GloVe	Global vectors for word representation. Это метод обучения без учителя, применяемый с целью получения векторного представления слов.
NLP	Natural language processing, или обработка естественного языка. Это направление искусственного интеллекта, изучающее проблемы компьютерного анализа и синтеза естественных языков.
Text mining	Интеллектуальный анализ текстов. Это направление в искусственном интеллекте, целью которого является получение информации из коллекций текстовых документов, основываясь на применении эффективных в практическом плане методов машинного обучения и обработки естественного языка.
TF-IDF	Term frequency – inverse document frequency. Это статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса.
Анализ тональности	Sentiment analysis. Это класс методов в компьютерной лингвистике, предназначенный для автоматизированного выявления в текстах эмоционально окрашенной лексики и эмоциональной оценки авторов по отношению к объектам, речь о которых идёт в тексте.
Бустинг	Boosting. Это процедура последовательного построения композиции алгоритмов машинного обучения, целевая функция каждого из которых состоит в компенсации недостатков предыдущих алгоритмов.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Магистерская диссертация: 66 с., 10 рис., 20 табл., 84 источников.

Ключевые слова: КЛАССИФИКАЦИЯ НОВОСТЕЙ, NLP, TEXT MINING, ПРЕДСКАЗАНИЕ РЫНКА FOREX, ИСКУССТВЕННАЯ НЕЙРОННАЯ СЕТЬ, СВЕРТОЧНАЯ НЕЙРОННАЯ СЕТЬ.

Множество исследователей доказали влияние новостей на фондовый рынок, с помощью различных математических алгоритмов им удалось использовать текстовые данные для прогнозирования цен ценных бумаг. Однако очень небольшое число исследований было проведено в отношении валютного рынка. В данной работе используются новейшие алгоритмы обработки естественного языка, машинного и глубокого обучения наряду со статистическим анализом и разработкой на основе его собственного алгоритма, рассмотрены три подхода к влиянию информации на рынок. Основная гипотеза исследования состояла в том, что валютный рынок с запозданием реагирует на поступающую в виде новостей информацию, что позволяет использовать новости для предсказания краткосрочных изменений котировки евро/доллар. Однако в ходе исследования ни один из использованных методов не позволил подтвердить гипотезу и выявить возможность для получения прибыли.

Цель исследования: оценить возможность существования взаимосвязи между публикуемой новостной информацией и котировкой евро/доллар и использовать ее для прогнозирования.

Объект исследования: влияние новостей на валютный рынок.

Методы исследования: обработка естественного языка, статистический анализ, градиентный бустинг, искусственная нейронная сеть, сверточная нейронная сеть.

Полученные результаты и их новизна:

- не были найдены основания для отклонения гипотезы о средней эффективности рынка евро / доллар США;
- использованы новейшие алгоритмы машинного и глубокого обучения для выявления взаимосвязи между новостями и валютным рынком, в то время как большинство существующих исследования описывают фондовый рынок с помощью более простых методов;
- сопоставлены три альтернативных подхода к влиянию информации на рынок, в том числе репортажный подход, не рассмотренный ранее;
- предложен классификатор, основанный на статистическом анализе важности слов и весовых коэффициентах.

Автор работы подтверждает, что работа выполнена самостоятельно и приведенный в ней расчетно-аналитический материал правильно и объективно отражает состояние исследуемого процесса, а все заимствованные из литературных и других источников теоретические, методологические положения и концепции сопровождаются ссылками на их авторов.

_____

(*подпись*)

АГУЛЬНАЯ ХАРАКТАРЫСТЫКА РАБОТЫ

Магістарская дысертацыя: 66 л., 20 мал., 20 табл., 84 крыніц.

Ключавыя словы: КЛАСІФІКАЦЫЯ НАВІН, NLP, TEXT MINING, ПРАДКАЗАННЕ РЫНКУ FOREX, ШТУЧНАЯ НЕЙРОННАЯ СЕТКА, СВЕРТАЧНАЯ НЕЙРОННАЯ СЕТКА.

Мноства даследчыкаў даказалі ўплыў навін на фондавы рынак, з дапамогай розных матэматычных алгарытмаў ім удалося выкарыстаць тэкставыя дадзеныя для прагназавання руху цэнаў каштоўных папер. Аднак вельмі невялікі лік даследаванняў быў праведзены ў дачыненні да валютнага рынку. У дадзенай працы выкарыстоўваюцца новыя алгарытмы апрацоўкі натуральнай мовы, машыннага і глыбокага навучання нараўне са статыстычным аналізам і распрацаваны на аснове яго ўласны алгарытм, разгледжаны тры падыходы до ўплыву інфармацыі на рынак. Асноўная гіпотэза даследавання складалася ў тым, што валютны рынак з спазненнем рэагуе на інфармацыю ў выглядзе навін, што дазваляе выкарыстоўваць навіны для прадказання кароткатэрміновых змяненняў каціроўкі еўра/долар ЗША. Аднак у ходзе даследавання ні адзін з выкарыстаных метадаў не дазволіў пацвердзіць гіпотэзу і выявіць магчымасць для атрымання прыбытку.

Мэта даследавання: ацаніць магчымасць існавання ўзаемасувязі паміж публікуемай навіннай інфармацыяй і каціроўкай еўра/долар ЗША і выкарыстаць яе для прагназавання.

Аб'ект даследавання: уплыў навін на валютны рынак.

Метады даследавання: апрацоўка натуральнага мовы, статыстычны аналіз, градыентны бусцінг, нейронная сетка, згортачная нейронная сетка.

Атрыманыя вынікі і іх навізна:

- не знойдзены падставы для адхілення гіпотэзы аб сярэдней эфектыўнасці рынку еўра да даляра ЗША;
- выкарыстаны новыя алгарытмы машыннага і глыбокага навучання для выяўлення ўзаемасувязі паміж навінамі і валютным рынкам, у той час як большасць існуючых даследаванні апісваюць фондавы рынак з дапамогай больш простых метадаў;
- параўнаны тры альтэрнатыўныя падыходы да ўплыву інфармацыі на рынак, у тым ліку рэпартажны падыход, не разгледжаны ў ранейшых даследаваннях;
- прапанаваны класіфікатар, заснаваны на статыстычным аналізе важнасці слоў і вагавых каэфіцыентах.

Аўтар працы пацвярджае, што прыведзены ў ёй разлікова-аналітычны матэрыял правільна і аб'ектыўна адлюстроўвае стан доследнага працэсу, а ўсе запазычаныя з літаратурных і іншых крыніц тэарэтычныя, метадалагічныя і метадычныя палажэнні і канцэпцыі суправаджаюцца спасылкамі на іх аўтараў.

_____

(*подпіс*)

GENERAL CHARACTERISTICS OF PAPER

Master thesis: 66 p., 10 fig., 20 tables, 84 sources.

Key words: NEWS CATEGORIZATION, NLP, TEXT MINING, FORECASTING FOREX, ARTIFICIAL NEURAL NETWORK, CONVOLUTIONAL NEURAL NETWORK.

Many researchers have proven the impact of news on the stock market. With various mathematical algorithms, they were able to use text data to predict the movement of prices of securities. However, few studies were conducted in relation to the foreign exchange market. This paper uses the newest algorithms of processing natural language, machine and deep learning, along with statistical analysis, proposes a new algorithm based on it and compares three ways of how information may influence market. The main hypothesis of the study is that the foreign exchange market responds to the information received in the form of news, which makes it possible to use the news to predict short-term changes in euro/dollar ratio. However, in the course of the study, none of the methods used made it possible to confirm the hypothesis and reveal the possibility for making profit.

Objective of the research: to assess the possibility of the existence of a relationship between the published news information and the euro / dollar quote and use it for forecasting.

Object: impact of economic news on the foreign exchange market.

Methods of the research: natural language processing, statistical analysis, gradient boosting, artificial neural network, convolutional neural network.

The results obtained and their novelty:

- no grounds were found for rejecting the hypothesis about the average efficiency of the euro / US dollar market;
- the use of new algorithms for machine and in-depth training to identify the relationship between news and currency market, while most of the existing studies describe the stock market using simpler methods;
- three alternative approaches to the impact of information on the market were compared, including the reportage approach not studied before;
- a new classificator was proposed which is based on statistical analysis of words importance and weightsю

The author confirms that the work is prepared independently and all the calculations and analytical material correctly and objectively reflects the state of the process under investigation, all borrowed from the literature and other sources of theoretical, methodological information and concepts accompanied by references to their authors.

_____

(*signature*)

ВВЕДЕНИЕ

Широкое распространение Интернета в 1990-х гг. изменило скорость принятия решений на финансовом рынке. Новости, публикуемые в режиме реального времени, стали отражаться на волатильности финансовых активов немедленно. Для валютного рынка спектр новостей, имеющих значения для котировок, необычайно широк: объявления о политических решениях в различных страна, публикация макроэкономической информации, кадровые перестановки в крупных государственных органах и надгосударственных организациях, интервью с целью так называемых «вербальных интервенций», изменение цен ан сырьевые товары, падение акций крупных коммерческих компаний и так далее. Осознанно или нет люди на всех уровнях используют информацию с целью участия или воздействия на финансовый (валютный) рынок. Участниками валютного рынка являются и физические лица, и крупные компании типа банков, но, в конечном счете, она касается и главного монетарного органа страны.

Информация, публикуемая в новостях, в различной степени оказывает влияние на всех участников рынка, в особенности, если эта информация являлась неожиданной. Потому большую важность имеет максимально быстрый анализ информации, чтобы помочь участникам рынка принять решение прежде, чем рынок скорректирует цены с учетом поступившей информации. С учетом огромного объема информации, поступающей едва ли не ежесекундно, выполнить такой анализ вручную практически невозможно. Это обуславливает потребность в создании автоматической системы для анализа воздействия новостей на рынок.

В той же декаде произошло развитие вычислительных алгоритмов, статистики, искусственного интеллекта, алгоритмов обработки естественного языка, что наряду с увеличением вычислительных мощностей способствовало формированию такой ветви обработки информации как интеллектуальный анализ текста.

Развитие технологий анализа данных и интеллектуального анализа текста и одновременное увеличение скорости поступления информации в виде публикуемых новостей создали возможности для применения анализа текста с целью прогнозирования финансового рынка.

Возможность использования текстовой информации для прогнозирования финансового рынка представляет огромный интерес и для исследователей, и для участников рынка. Множество исследований было проведено с целью предсказания цен на фондовом рынке. Однако из-за сложности моделирования поведения людей, до сих пор не существует устоявшейся методологии решения этой задачи. Тем не менее, в отношении прогнозирования фондового рынка на основе новостей были достигнуты определенные успехи.

Другая сложность состоит в создании алгоритма, способного «понимать» текст, извлекать из него важные характеристики. Текстовая информация поступает из различных источников и имеет разнообразную форму: от сжатого

отчета до интервью. Сильно варьируется и лексика, одни и те же слова могут встречаться в совершенно различном контексте. С одной стороны, новостные сообщения, как правило, имеют нейтральную окраску, что влияет на характер используемой лексики. С другой стороны, информация может быть по-разному интерпретирована авторами новостей [62].

Современные исследователи рассматривают широкий спектр источников и типов информации. Наиболее традиционный из них – новости с крупного новостного портала. Другое популярное направление сейчас – анализ публикаций в социальных сетях (например, в Твиттере) с целью предсказания на основе их текста движений на фондовом рынке. Во множестве публикаций используется анализ тональности по их комментариям, чтобы понять настроение агентов рынка и предсказать рост или падение котировки [25]. Для анализа фондового рынка также используют годовые отчеты компаний, пресс-релизы.

Второй тип данных, необходимых для исследования, - рыночные данные. Данные, как правило, используются для обучения алгоритмов машинного обучения, реже их используют как объясняемую переменную в регрессии. Большинство публикаций сконцентрированы на исследовании рынка ценных бумаг [66]. Например, рассматриваются следующие индексы: Dow Jones Industrial Average [75], US NASDAQ [50], Morgan Stanley High-Tech Index [81], the Indian Sesex Index [38], S&P 500 [67], - или цена акций конкретной компании (например, BHP Billion Ltd [82], Apple [78], Google, Microsoft, Amazon [6]), или группы компаний [66].

Рынок Forex рассмотрен примерно в десяти процентах работ [66]. Например, в одной из публикаций рассмотрено изменение котировки доллара к немецкой марке в 1990-х гг. за день и получена модель, способная давать адекватный прогноз [12]. В другой интересной работе о влиянии макроэкономических новостей с использованием регрессии получен результат, что новости могут объяснить порядка 22 – 56 % скачков котировок в течение пяти минут [13]. Еще один пример использования линейной регрессии наряду с использованием алгоритмов обработки языка, кластеризации по темам и анализа тональности позволил создать систему Forex-Foreteller [22].

Тем не менее, валютный рынок изучен в значительно меньшей степени и, как правило, исследователи концентрировались на анализе локальных валют. В данном исследовании будет проведена попытка смоделировать поведение свободно конвертируемой пары – котировки евро к доллару США. Это вариант более сложного анализа, поскольку намного больший объем информации воздействует на данную котировку: от войны в странах, добывающих нефть, до референдума о Brexit. Кроме того, намного больше круг участников рынка данной валютной пары.

В случае, если будет доказана возможность предсказания краткосрочных колебаний котировки евро/доллар, полученная модель может быть использована всеми участниками рынка: от трейдеров – физических лиц, до Национального банка Республики Беларусь.

Также новизна работы в попытке использовать новейшие алгоритмы машинного и глубокого обучения, в то время как в большинстве исследований авторы ограничивались сравнительно более простыми методами: от наивного байесовского классификатора до классификации на основе опорных векторов.

Объект исследования: влияние публикуемых новостей на валютный рынок.

Предмет исследования: оценка взаимосвязи информации, поступающей в виде текстовых сообщений, на изменение котировки евро/доллар.

Основная гипотеза, рассмотренная в работе: новости влияют на краткосрочные колебания валютного рынка и возможно их использование с целью предсказания движения котировки.

Цель работы: оценить возможность существования взаимосвязи между публикуемой новостной информацией и котировкой евро/доллар и использовать ее для прогнозирования.

В соответствии с выделенными предметом, объектом и целью исследования можно сформулировать следующие задачи:

- изучить теоретические обоснования влияния информации на финансовый рынок, ознакомиться с результатами практических исследований, современными методами, используемыми для обработки текстовой информации и оценки ее влияния посредством классификации;
- собрать необходимые данные, подготовить их к исследованию, статистически оценить возможность существования взаимосвязи между новостями и котировкой евро/доллар;
- используя новейшие алгоритмы машинного и глубокого обучения, оценить модели классификации, использующие тексты новостей для предсказания направления движения курса валюты.

При подготовке работы был использован широкий спектр источников. Большое количество идей почерпнуто из публикации Kim-Georg Aase “Text mining of news articles for stock price predictions” 2011 года [2]. При описании алгоритмов глубокого обучения использовалась книга С. Николенко, А. Кадурина и Е. Архангельской «Глубокое обучение» [84], она же использовалась для адекватного перевода англоязычной терминологии или как основание для использования английского термина вместо перевода. В качестве источника текстовых данных выступил новостной портал Reuters.com [41]. Для выполнения исследования использовались языки программирования R и Python.

Автор работы подтверждает, что приведенный в ней расчетно-аналитический материал правильно и объективно отражает состояние исследуемого процесса, а все заимствованные из литературных и других источников теоретические, методологические, методические и концептуальные положения сопровождаются ссылками на их авторов.

ГЛАВА 1

ОБЗОР ЛИТЕРАТУРЫ И МЕТОДОВ ОЦЕНКИ

1.1 Теории и подходы к учету информации в прогнозировании финансовых рынков

1.1.1 Гипотеза эффективного рынка

В публикации Grace S. Thomson [73], теории, касающиеся роли информации на финансовых рынках, могут быть условно разделены на теории оценки активов и теории финансового поведения. К первой группе относятся теория эффективного рынка, ко второй – REMM теория человеческого поведения (resourceful, evaluative, maximizing model).

Долгое время превалирующими финансовыми теориями были гипотеза случайного блуждания и гипотеза эффективного рынка.

Гипотеза эффективного рынка базируется на работе Ф. Хайека 1945 года «Использование знаний в обществе» [71]. В указанной статье Хайек утверждает, что рынок – это наиболее эффективный способ агрегировать информацию, принадлежащую различным индивидуумам в обществе. При этом под знаниями (информацией) Хайек понимает не только статистически данные, а еще и «знание особых условий времени и места», той уникальной информации, использование которой позволяет любому индивиду обладать определенным преимуществом перед всеми остальными.

Хайек противопоставляет централизованную систему планирования в экономике конкуренции (децентрализованному планированию множеством отдельных лиц) с промежуточным вариантом в виде монополии. Он добавляет, что централизованная плановая экономика никогда не сможет достичь эффективности открытого рынка, поскольку информация, которой владеет один агент, является лишь малой частью информации, существующей в обществе:

«По сути, в системе, где знание значимых фактов распылено среди множества людей, цены могут координировать разрозненные действия различных лиц так же, как субъективные ценности помогают индивиду координировать части его плана» [71].

Хайек призывает смотреть на систему цен как на механизм передачи информации, ведь с ее помощью стало возможным разделение труда и «скоординированное употребление ресурсов, основанное на равномерно распределенном знании».

В дальнейшем гипотеза о влиянии информации на рынок получила развитие и была сформулирована Юджином Фама в 1965 году как гипотеза эффективного рынка в отношении рынка ценных бумаг.

По определению Фама, эффективный рынок – это рынок, характеризующийся активной конкуренцией большого числа рациональных

агентов, максимизирующих свою прибыль и пытающихся предсказать будущую рыночную цены отдельных ценных бумаг, и свободно доступной всем участникам рынка информацией [21]. Согласно гипотезе об эффективном рынке, на эффективном рынке конкуренция между рациональными участниками приводит к тому, что в любой момент времени текущие цены отдельных ценных бумаг уже отражают эффект от информации, вызванной уже произошедшими событиями и событиями, которые, по ожиданию рынка в данный момент, должны произойти в будущем.

При определенных допущениях: отсутствие транзакционных издержек по поиску информации, невозможности запаздывания или асимметрии информации – никакая активная стратегия инвестирования не позволит инвестору «обыграть» рынок.

Гипотеза случайного блуждания предполагает, что изменение цен на рынке следует процессу случайного блуждания и не может быть предсказано. Основы ее заложил Пол Кутнер (1964) в книге «Случайный характер цен фондового рынка» [11], а далее она стала тесно связана с гипотезой эффективного рынка.

В 1965 году Пол Самуэльсон опубликовал статью [48], в которой доказывал, что если рынок эффективен, то поведение цен будет описываться процессом случайного блуждания. Таким образом, модель случайного блуждания применима после того, как полностью учтены фундаментальные факторы. Он также утверждал, что гипотеза эффективного рынка лучше описывает отдельные ценные бумаги, нежели движение на рынке в целом, разделяя, таким образом, экономику на микро и макро уровень. Неэффективность гипотезы на макроуровне он доказывал финансовыми пузырями и другими кризисами.

Согласно Самуэльсону, инвестор имеет шанс превзойти рынок, если обладает частной информацией (средняя форма эффективности рынка):

«На мой взгляд, современным биржам свойственно то, что я называю ограниченной микроэффективностью. Пока небольшое число инвесторов, обладающих значительными активами, играет против упрямых неинформированных трейдеров, можно практически наблюдать ограниченную эффективность рынка. Проявление ценовых отклонений вызывает ажиотаж среди трейдеров, действия которых устраняют эти ценовые отклонения» [48].

Фактически, это означает, что рыночные цены должны учитывать изменение информации мгновенно:

«К моменту, когда многочисленные инвесторы осмыслят новую инфрмацию в будущем, все, что имеет практическую ценность, уже учтено в текущих ценах» [48].

Цитируя Самуэльсона, Питер Бернстайн делает вывод, что «направление движения цен акций сложно предсказать именно потому, что цена акций сама является прогнозом будущего» благодаря учтенной в ней информации, возможно, еще не известной всем участникам рынкам [83].

В статье 1970 года Фама ввел понятия разной степени эффективности рынка. Низкая эффективность означает, что цены отражают только

исторические данные и невозможно систематически получать выигрыши на основе прошлой ценовой информации. Во временных рядах должна отсутствовать автокорреляция, движение цен является случайным.

Средняя степень эффективности предполагает, что цены активов адаптируются к новой общедоступной информации достаточно быстро, чтобы было невозможно заработать дополнительную прибыль используя публичную информацией. Преимущество на рынке имеют те игроки, которые владеют частной информацией. Во временных рядах наблюдаются изломы после объявления публичной информации.

В случае высокой эффективности, цены на финансовом рынке отражают всю информацию – частную и общедоступную – и ни один участник рынка не может получить дополнительную прибыль.

В статье Фама приводятся доказательства случаев слабой и средней рыночной эффективности, тогда как сильная форма предлагается к рассмотрению в качестве ориентира для сравнения.

Таким образом, гипотеза эффективного рынка тесно связана с гипотезой о случайном блуждании: чем более эффективен рынок, тем более случайным является процесс, которому следуют цены. Однако наоборот это утверждение не действует: случайное блуждание в изменении цены отнюдь не означает эффективность рынка.

1.1.2 Поведенческий подход

Гипотеза эффективного рынка имеет ряд ограничений [8]:

- предположение о рациональности индивидов (инвесторов и экономических агентов);
- ограниченность информации;
- нет объяснения спекулятивным пузырям. Если цены немедленно реагируют на поступающую информацию, пузыри на рынках были бы невозможны.

Тезис о том, что рынки не полностью эффективны из-за ограниченности информации предложил еще Джон Мэйнард Кейнс в «Общей теории занятости, процента и денег»: «В совокупных капиталовложениях общества постепенно растет доля акций, которые принадлежат лицам, не принимающим участия в управлении и не обладающим специальными знаниями о вещах, имеющих отношение к настоящему или будущему данной отрасли. В результате серьезно ослаблен элемент действительного знания в оценке соответствующих предприятий или бумаг как теми, кто владеет ими, так и теми, кто намеревается их купить» [35].

Кейнс не верил в возможность долгосрочных прогнозов из-за неопределенности на рынке и сравнивал инвестиции в акции с «конкурсами красоты» [35], когда каждый инвестор-профессионал должен сделать на наилучший для себя выбор, а предугадать мнение большинства. Масса инвесторов в целом ведет себя нерационально:

«Есть еще одна характерная деталь, которая особенно заслуживает нашего внимания. Можно было бы полагать, что конкуренция между квалифицированными профессионалами, обладающими рассудительностью и знаниями выше уровня среднего частного инвестора, нейтрализует причуды неосведомленного индивидуума, предоставленного самому себе. На деле, однако, энергия и искусство профессиональных инвесторов и биржевых игроков часто направляются в иную сторону. Большинство этих лиц в действительности весьма озабочены не тем, чтобы составить наилучший долгосрочный прогноз ожидаемого дохода от инвестиций за все время их эксплуатации, а тем, чтобы предугадать немного раньше широкой публики изменения в системе взаимно разделяемых условностей как основы рыночной оценки. Их интересует не реальная стоимость какого-то объекта вложения капитала для человека, который покупает его с тем, чтобы "приберечь" его для себя, а то, как рынок будет оценивать его под влиянием массовой психологии через три месяца или через год» [35].

В самой человеческой натуре есть склонность к нерациональному поведению:

«Даже оставляя в стороне неустойчивость, связанную со спекуляцией, приходится считаться еще с неустойчивостью, проистекающей из определенного свойства человеческой природы, которое выражается в том, что заметная часть наших действий, поскольку они направлены на что-то позитивное, зависит скорее от самопроизвольного оптимизма, нежели от скрупулезных расчетов, основанных на моральных, гедонистических или экономических мотивах. Вероятно, большинство наших решений позитивного характера, последствия которых скажутся в полной мере лишь по прошествии многих дней, принимается под влиянием одной лишь жизнерадостности - этой спонтанно возникающей решимости действовать, а не сидеть сложа руки, но отнюдь не в результате определения арифметической средней из тех или иных количественно измеренных выгод, взвешенных по вероятности каждой из них» [35].

Хотя по существующему стандарту теории Кейнса не относятся к школе поведенческой экономике, очевидно, что он опровергал предположения о рациональности экономических агентов, общедоступности и возможности обработки всей информации, эффективности рынка.

Школа поведенческой экономики заложена Робертом Шиллером, Ричардом Таллером, множество практических исследований проведено Дэниэлом Канеманом и Амосом Тверски, подтверждающих ограниченную рациональностей людей.

В ответ на гипотезу эффективного рынка, поведенческая экономика предполагает, что рынки не эффективны и случайное блуждание может быть объяснено человеческим поведением, поскольку оно объясняет принятие решений экономическими агентами, а человеку свойственно принятие нерациональных решений и совершение систематических ошибок. Именно ошибки влияют на неэффективность рынка.

Теории поведенческой экономики подтверждаются исследованиями в психологии, социологии и экономике. Были зафиксированы наблюдения, в которых одна и та же информация имела различную интерпретацию у участников рынка [30].

Гипотеза адаптивного рынка

В качестве общей теории, объединяющей теорию эффективного рынка и поведенческий подход, была предложена теория адаптивного рынка (adaptive market hypothesis) [54]. Она предполагает, что степень эффективности рынка связана с факторами среды, такими как число конкурентов на рынке, количество возможностей для получения выгоды, адаптивность участников рынка.

Финансовые рынки являются экологическими системами, в которых различные группы конкурируют друг за другом за ограниченные ресурсы. То, что поведенческая экономика считает нерациональным поведением, - неприятие рынка, чрезмерная уверенность и пр., гипотеза адаптивного рынка описывает эволюционной моделью поведения участников рынка, адаптирующихся к изменениям среды.

Заявляется, что теория эффективного рынка может выполняться только в том случае, если на рынке капитала отсутствуют транзакционные издержки, налоги, институциональные жесткие правила, ограниченность когнитивных способностей участников рынка.

Согласно теории Эндрю Ло, рынки развиваются циклически, периоды, когда конкуренция истощает существующие торговые возможности, сменяются периодами появления новых возможностей. Следовательно, на финансовых рынках существует возможность получения прибыли.

1.1.3 Технический и фундаментальный анализ

Инструментарий для анализа рынков при предположении о том, что предсказания поведения цен на рынке возможны, разделен на группы: технический и фундаментальный анализ.

Технический анализ

Технический анализа – это изучение предшествующих изменений цены с целью предсказания будущих изменений. Технический анализ финансовых рынков рассматривает исключительно рыночные данные, но не другую информацию. Основной тезис технического анализа состоит в том, что цена акции зависит только от предложения и спроса на рынке, единственно важная информация для оценки акций – объем и цена ценных бумаг [18]. Предполагается, что невозможно охватить и полностью понять информацию о компании, чтобы принять на ее основе решение о покупке или продаже.

В рамках технического анализа рассматриваются тренды и изменения в них (переломы и разрывы) на протяжении некоторого периода времени. Основными инструментами являются анализ графиков ценных бумаг на рынке, оценка скользящих средних, линий тренда и расчет статистических метрик. Большое внимание уделяется устойчивым паттернам на дневных данных,

например, «голова и плечи», «double top» и др. для выявления линий поддержки или сопротивления [19].

Основы технического анализа заложил Уильям Питер Хамильтон в книге «The Stock Market Barometer» [29], описав теорию Доу – первую теорию анализа графиков и развитую впоследствии Робертом Реа в книге «Теория Доу» [56]. Теория Доу предполагает, что вся необходимая информация о прошлом, настоящем и даже будущем. Также выделяется три типа трендов: первичный, вторичный и третичный. Первичные тренды намного более зависимы от случайных новостей и потому их сложно определить.

Технический анализ широко распространен среди брокеров и других участников рынка для принятия инвестиционных решений. Тем не менее технический анализ обусловлен человеческой интерпретацией паттернов на рынке. Субъективная оценка зачастую приводит к тому, что аналитики приходят к противоположным решениям на основании одного сигнала.

Ранние исследования Alexander (1961) [47], Fama и Blume (1966) [23], Levy (2007) [55], Jensen (1967) [52], Jensen и Bennington (1970) [51] показывали, что технический анализ бесполезен. Однако многие более поздние исследователи находили возможность для получения прибыли на рынке, используя методы технического анализа (например, Bessembinder and Chan (1995) [70], Lai et al. (2003) [15] показали возможность выгодного инвестирования на развивающихся рынках). Эндрю Ло, впоследствии предложивший гипотезу адаптивного рынка, в 2005 году участвовал в исследовании, подтвердившим возможность принятия выгодных решений на основе некоторых технических результатов [54].

Фундаментальный анализ

В рамках фундаментального анализа изучают всю релевантную информацию, касающуюся цены акции, чтобы определить внутреннюю (реальную) ценность актива.

Фундаментальный анализ стал использоваться для принятия торговых решений в 1928 году, первая посвященная ему книга была опубликована в 1934 году [28]. Фундаментальный анализ рассматривает финансовые индикаторы (в зарубежной литературе их называют “fundamentals”). Если говорить о рынке акций, то это показатели, описывающие финансовое состояние компании: объем активов и пассивов, доходы, а также рынок компании, конкурентов, руководство, новости об изменениях в технологии. При изучении рынка фьючерсов или ForEx рассматривают макроэкономические индикаторы: процентные ставки, налоги, занятость, ВВП, оптовые и розничные цены, а также новости о политике и даже погоде [7].

Возможность выгодного инвестирования, следуя фундаментальному анализу, появляется для неправильно оцененных финансовых инструментов. Например, выгодно приобрести акции компании, когда она недооценена относительно своих фундаментальных индикаторов, и продать, когда рынок сгладит эту разницу или компания будет переоценена. Фундаментальный анализ часто используется для долгосрочных инвестиционных стратегий, так как требуется время для того, чтобы изменились фундаментальные индикаторы

компаний или государств. Другая стратегия инвестирования – «buy and hold», когда на основе фундаментальных показателей выбирается компания для инвестирования с низким риском и стабильной доходностью от акций.

Основная разница между техническим и фундаментальным анализом в том, что технический анализ игнорирует макроэкономические индикаторы и информацию о состоянии компании, поскольку эта информация уже должна быть учтена в текущих ценах.

Сейчас широко используются программы для фундаментального анализа, постоянно отслеживающие рынки и события. Однако анализ во многом выполняется человеком и потому может быть субъективным, как и предполагает поведенческий подход [76]. Поэтому разработка полностью автоматизированных систем анализа представляется очень многообещающей.

Технологические методы

Развитие компьютерных технологий отразилось на методах исследования и прогнозирования рынка. Использование Интернет-технологий открывает новые возможности для анализа. Одним из примеров может быть статья 2013 года, в которой торговые стратегии были основаны на анализе поисковых запросов 98 финансовых терминов (с помощью Google Trends) и позволили получить 326 % аккумулированной доходности за восемь лет ретроспективной симуляции [57]. В том же году была выведена взаимосвязь между числом просмотров некоторых финансовых статей в Wikipedia и движениями фондового рынка. Таким образом изменение поведения пользователей Google или Wikipedia, такое как увеличение запросов по некоторым статьям или терминам, может использоваться как ранний сигнал, предупреждающий о движении рынка [58].

Объем информации и количество инвестиционных стратегий, которые трейдер может быстро проанализировать, весьма ограниченно. Кроме того, как упоминалось выше, когнитивные искажения могут привести к ошибке [30]. Увеличение вычислительной мощности компьютеров стало толчком к усложнению методов моделирования и привело к быстрому развитию методов машинного и глубокого обучения, позволяющих автоматически принимать решения в условиях ограниченного времени. Возможность обработки неструктурированных данных, таких как текст, позволяют учитывать идеи поведенческого подхода при прогнозировании финансового рынка.

1.2 Практикоориентированные исследования

1.2.1 Исследования об эффективности рынка

Во-первых, нет единого мнения относительно гипотез случайного блуждания и эффективного рынка. В середине 70-х гг. гипотеза эффективного рынка и случайного блуждания были центральными предположениями в теории финансов. На данный момент «нет настоящего ответа на вопрос, следуют ли цены случайному блужданию, но возрастает число доказательств, что нет» [79].

В статье Самюэля Дюпернекса приводится обзор различных исследований, посвященных вопросу [79]. Автор приходит к выводу, что результаты слишком противоречивы, чтобы прийти к единому выводу. Подтверждение гипотез обычно предполагает следующие условия:

- Инвесторы рациональны и рационально торгуют активами;
- Некоторые инвесторы нерациональны, но их торговля случайна и взаимоисключаема;
- Некоторые инвесторы нерациональны, но арбитраж обнуляет их влияние на цены [59].

Доказательство несостоятельности гипотез обычно основывается на кратко- и долгосрочной серийной корреляции и явном тренде, в том числе сезонном, в данных, а также влиянии новостей.

Общий вывод, к которому пришёл Самюэль Дюпернекс, – существуют доказательства того, что рынки в определенной степени могут быть предсказаны. Это не создает возможности для арбитража, поскольку они слишком быстро нивелируются [79].

Хашем Песаран говорит, что статистические доказательства против гипотезы случайного блуждания были отклонены как незначимые с экономической точки зрения (поскольку на их основе нельзя предложить доходные торговые правила с учетом транзакционных издержек) и подозрительные с точки зрения статистики (из-за отсутствия единого подхода к подготовке и исследованию данных) [45].

Во-вторых, эффективность рынка требует, чтобы цены отражали всю возможную информацию [32]. Понятно, что современный рынок включает в себя публичную информацию, но, как правило, не имеет своевременного доступа к частной. В 1978 году Pratt и DeVere показали, что инвесторы, покупающие акции на основе публичного доступного сигнала “buy” могут получать сверхприбыль даже после двухмесячного лага [46]. Ferreira (1995) утверждал, что агенты корпораций используют инсайдерскую информацию для корректировки цены акций собственной фирмы относительно будущих движений рынка [32]. К похожему выводу приходит Jaffe (1974), проанализировав доходность транзакций, проведенных представителями 200 крупных фирм в 1962 – 1968 гг. [61]. Он обнаружил, что инсайдеры старались покупать акции перед резким ростом их цены и продавать перед резким падением. Seyhun (1986) показал, что только инсайдеры могут предсказать эти резкие скачки цен на основе своей, частной информации [31]. Рынки эффективны, пишет он, и аутсайдерам недостаточно использовать общедоступной информацию о сделках инсайдеров чтобы получить сверхприбыль [33]. Таким образом, рынок ценных бумаг можно считать средне эффективными, в терминах Фама.

В то же время есть и исследования, ставящие под вопрос теорию эффективного рынка по итогам анализа долгосрочного периода. Alvarez-Ramirez et al. оценивали степень эффективности финансовых рынков в течение времени [34]. Относительная эффективность рынка ценных бумаг США варьировалась в течение 1929 – 2012, со спадом в конце 2000-х гг., вызванным

экономической рецессией. Наиболее эффективным периодом был 1973 – 2003 гг.

Другое исследование за авторством Abounoori E., Shahrazi M., Rsekhi S. посвящено изменению эффективности рынка IRR/USD в течение 2005 – 2010 гг. в связи с отрицательной долгосрочной зависимостью курса [4].

В статье Kim J.H., Lim K.P. (2011) достигнут подобный результат: рынок индекса Доу Джонса был изучен за период 1900 – 2009 гг. и во время экономических и политических кризисов был более предсказуем [65]. Из перечисленных исследований можно сделать вывод, что сколь значимым бы ни был информационный фон для некоторого рынка, котировкам также свойственно зависеть от доминирующих трендов и предыдущих данных [40].

Также выдвигалась теория о том, что форма эффективности рынка варьируется между странами: рынок США более эффективен, чем рынки развивающихся стран [10]. В другой работе этот тезис дополняется выводом о том, что финансовый кризис значительно увеличил эффективность развивающихся рынков, таких как Китай, Гонконг, Корея и африканский рынок.

Для валютного рынка средняя форма эффективности была разделена на две категории: эффективности единственного рынка и мультирыночная эффективность [60].

Эффективность единственного рынка означает, что вся публичная информация об одном обменном курсе является частью доступной информации. Мультирыночная эффективность информация обо всех обменных курсах включена в доступную информацию.

Эффективность пары EUR/USD изучены в работе Boboc I.A., Dinica M.C., где генетическими методами симулируется поведение 100 участников рынка, использующих методы технического анализа для принятия решения о покупке/продаже валюты [9]. Результаты показали, что гипотеза о слабой форме неэффективности рынка для данной валютной пары не может быть отклонена, так как ни один из используемых наборов стратегий не позволил симулированным участникам обыграть рынок.

Для валютного рынка основной инсайдер – это центральный банк. Bailliem, McMahon (1989) утверждают, что сильная форма эффективности маловероятна на валютном рынке из-за неслучайных интервенций центральных банков [68]. Sweeney (2000) продолжает это утверждение добавляя, что если центральный банк получает нулевой доход от интервенций, то рынок характеризуется сильной степенью эффективности [16].

1.2.3 Исследования о влиянии информации

Судя по всему, работа Wuthrich et al. была первой, в которой был построен прототип использования text mining техник для предсказания цен на финансовых рынках с помощью анализа финансовых новостей [80].

Chan et al. подтвердили влияние новостей на рынок [69]. В их исследовании показано, что экономические новости всегда оказывают

позитивный или негативный эффект на рассматриваемые ценные бумаги. Они использовали значимые политические и экономические новости в качестве публичной информации. Оба типа новостей влияли на такие показатели рыночной активности, как волатильность доходности, цены, количество торгуемых акций и частота сделок. Интересным результатом стало то, что влияние новостей на волатильность цены начиналось за день до публикации, достигало максимума в день публикации новости и затем снижалось. Влияние политических новостей начиналось за день до публикации также, но достигало пика спустя день после публикации новости.

Использование машинного обучение в анализе текста с тех пор набрало популярность. Обычно в машинном обучении используется предварительно размеченный набор документов для обучения модели, затем построенная модель используется для классификации на отложенном тестовом сете.

В качестве методов классификации используются наивный байесовский классификатор, NewsCAT с методом опорных векторов, AZFinText system, включающая регрессию с сегментацией временных рядов, деревья решений, регрессионные модели [67].

В работе Vakeel J., Dey S. рассматриваются новости за четырехмесячный период накануне и после выборов в Индии для анализа влияния на индекс S&P BSE SENSEX на финансовом рынке [77]. Для классификации новостей был выбран следующий подход: если в течение часа после выхода новости индекс вырос, сообщению присваивалась метка «rise», если снизился – «fall». Препроцессинг данных был выполнен с помощью Weka (Java приложение с открытым исходным кодом), в том числе для расчета TF-IDF индекса (term frequency – inverse document frequency). Вес термина (term weight) – это статистическая мера важности термина в документе. В качестве терминов были рассмотрены юниграмы и биграмы (unigrams and bigrams), т.е. отдельные слова или пары слов. Слова, использованные более чем трижды в одном документе, считались атрибутами. Оценка information gain (IG) была использована для сокращения набора предикторов (выбранных атрибутов). Для моделирования используется SVM, обученная на классифицированных новостях и оцененная для двух временных интервалов – до выборов и после. Точность обеих моделей составила около 63 % при практически сбалансированных классах (вероятность случайного угадывания составила 50 – 60 %). Точность модели недостаточно высокая, однако можно утверждать, что текстовые сообщения оказывают влияние на фондовый рынок и могут использоваться для предсказания движения индексов наряду с техническим анализом.

Отличие работы Ding et al. в отказе от простых подходов к text mining (bag-of-words, noun phrases, named entities) и переходу к вложениям событий (event embeddings) [14]. Структурированные события извлекаются из неструктурированного текста с помощью технологии Open IE и разбора зависимостей (dependency parsing). Вложения обучаются с помощью NTN (novel tensor network) таким образом, чтобы похожие события (вида субъект – действие – объект) представлялись схожими векторами, даже если слова в них не пересекаются.

Модель прогнозирования была построена для долгосрочного (за предшествующий месяц), среднесрочного (за последнюю неделю) и краткосрочного (за день) наборов новостей. Для оценки влияния новостей на индекс S&P используются глубокая CNN (convolutional neural network) и стандартная нейронная сеть с одним скрытым слоем и одним output слоем: вложения событий за день усредняются в одно наблюдение, классификация осуществляется для направления фондового индекса (рост или снижение). Помимо стандартных оценок качества прогноза, использовалась симуляция реальной торговли на фондовом рынке по стратегии Lavrenko et al (2000) [36]. Полученная оценка показала, что предложенная модель работает эффективнее предложенных ранее и представление событий в виде вложений гарантирует лучший результат.

В диссертации Lopes Antunes P.A. (2015) для подготовки текстовых данных наряду с обычными техниками (исключение стоп-слов, возврат к корню слова) создается словарь на основе доступных словарей Opinion Lexicon, OpinionFinder, SentiWordNet, AFINN, NRC для анализа окраски текста (Sentiment analysis) на уровне документа, предложения и группы слов [5]. Метки новостей были выставлены вручную и сравнивались с предсказаниями, полученными с помощью Sentiment classifier.

Таким образом, задача оценки влияния новостей на валютный рынок сводится к преобразованию текстовых сообщений, сопоставлению их с временными метками и изменениями новостей в положительную или отрицательную сторону с целью классификации. Альтернативным способом классификации новостных сообщений является анализ тональности (sentiment analysis). Желательным результатом анализа является выделение блоков слов и/или словосочетаний, которые имеют наиболее значимое влияние на рынок.

1.3 Подходы к моделированию

1.3.1 Анализ тональности

Проблема анализа тональности может быть сформулирована как две отдельные проблемы классификации или как проблема трехклассовой классификации.

Если рассматривать задачу как две проблемы классификации, то первая из них состоит в определении того, является текст субъективным или объективным, то есть выражает ли он мнения. Эта проблема получила название классификация субъективности (subjectivity classification). Тогда вторая проблема состоит в том, чтобы классифицировать предложения в субъективном документе как позитивные или негативные. Задача бинарной классификации документа как выражающего в целом позитивное или негативное мнение названа (sentiment) polarity classification.

Если задачу классификации тональности рассматривать как трехклассовую классификацию, то текст классифицируется как позитивный, негативный или нейтральный [42]. В литературе метка «нейтральный» иногда

используется для объективного класса (как текст, не выражающий никакого мнения) или как степень тональности, промежуточная между позитивной и негативной.

Классификация по тональности может начинаться на уровне слова в документе (новостном сообщении, в конкретном случае). На основе существующих лексиконов каждое слова документа классифицируют как позитивное (например, «рост», «доход»), негативное («падение», «убыток») или нейтральное (это слова, не найденные ни в положительном, ни в отрицательном лексиконах). Если число позитивных слов в документе превышает число негативных, то документ классифицируется как позитивный. Если в документе негативные слова преобладают над позитивными, то он классифицируется как негативным. Если ни одно из условий не удовлетворено, то документ считается нейтральным.

В некоторых исследованиях авторы отступали от трехклассовой задачи к классификации к более сложным, например, присваивая текстам метки от единицы до пяти [43].

Такой метод классификации может быть рассмотрен как мультиклассовая проблема категоризации текста (или ordinal classification) [27].

Помимо использования существующих лексиконов, есть опция составления нового, вручную присваивая метку каждому слову из словаря всех текстов. Например, Taboada et al. (2011) создал свой лексикон вручную, поскольку автоматически сгенерированные лексиконы не обеспечивали стабильность классификации [37].

Другой подход основывается на использовании словарей [53]. Начальный набор слов размечается вручную, а затем расширяется с помощью словарей антонимов и синонимов. Вновь найденные слова добавляются к первоначальному списку и процесс повторяется итеративно, пока станет невозможно добавить новые слова к составленному списку. После завершения процесса составленный лексикон проверяется вручную.

Несомненным плюсом этого метода является простота подбора большого числа тональных слов. Однако метод не учитывает изменения смысла в зависимости от контекста. Например, «рост» в фразе о доходе имеет положительный оттенок, но в контексте долга – негативный.

Corpus-based подход основывается на синтаксических паттернах или паттернах словосочетаний, пытаясь решить проблему контекста при отнесении слов к позитивным или негативным [53].

Анализ тональности широко используется в изучении отзывов на продукты и услуги, сообщений в социальных сетях и новостей. Тональность сообщений может быть полностью оценена экспертно для обучения модели, однако это трудоемкий и время затратный процесс, выбор автоматического или полуавтоматического ранжирования текстов более предпочтителен.

1.3.2 Градиентный бустинг

Бустинг — это подход к построению композиций, в рамках которого:

- базовые алгоритмы строятся последовательно, один за другим;
- каждый следующий алгоритм строится таким образом, чтобы исправлять ошибки уже построенной композиции.

Благодаря тому, что построение композиций в бустинге является направленным, достаточно использовать простые базовые алгоритмы, например, неглубокие деревья.

Градиентный бустинг является одним из лучших способов направленного построения композиции на сегодняшний день. В градиентном бустинге строящаяся композиция

является суммой, а не усреднением N базовых элементов . Это связано с тем, что алгоритмы обучаются последовательно и каждый следующий корректирует ошибки предыдущих.

В задаче классификации функция потерь $L(y,z)$, где y – истинный ответ, z – прогноз алгоритма на некотором объекте, обычно имеет вид:

В начале построения композиции по методу градиентного бустинга нужно ее инициализировать, то есть построить первый базовый алгоритм . В задаче классификации это может быть алгоритм, который всегда возвращает метку самого распространенного класса в обучающей выборке:

Обучение базовых алгоритмов происходит последовательно. Пусть к некоторому моменту обучены $N+1$ алгоритмов , то есть композиция имеет вид:

Затем к текущей композиции добавляется еще один алгоритм , который обучается так, чтобы как можно сильнее уменьшить ошибку композиции на обучающей выборке:

Сначала решается более простая задача: определить, какие значения должен принимать алгоритм на объектах обучающей выборки, чтобы ошибка на обучающей выборке была минимальной:

где – вектор сдвигов. Направление наискорейшего убывания функции задается направлением антиградиента, его можно принять в качестве вектора *s*:

Компоненты вектора сдвигов *s*, фактически, являются теми значениями, которые на объектах обучающей выборки должен принять новый алгоритм , чтобы минимизировать ошибку строящейся композиции. Обучение , таким образом, представляет собой задачу обучения на размеченных данных, в которой – обучающая функция.

1.3.3 Сверточные нейронные сети

Сверточные нейронные сети – мощный инструмент машинного обучения, нацеленный на эффективное распознавание и классификацию, в первую очередь, изображений и предложенный Яном Лекуном. Однако эффективность в сокращении размерности позволяет использовать их также для классификации текстов.

Вычисление сверточных нейронных сетей (convolutional neural network; далее – CNN) основано на работе с тензорами. Под тензором понимается многомерный массив чисел (не элемент тезорного пространства). Оперирование тензорами, а не матрицами позволяет не терять информацию за счет перехода к меньшей размерности. В случае работы с изображениями CHC принимает в качестве аргументов трехмерный тензор – изображение с *H* строк, *W* колонок и тремя каналами (R, G, B – красный, зеленый и синий цвета). Для задач работы с текстом размерность тензора может быть большей.

Каждый шаг работы с данными обычно называют слоем сети. Архитектура сверточной нейронной сети обычно представляет собой чередование сверточных слоев (convolutional layers), слоев пулинга (pooling layers), нормализации (normalization layers) и полносвязных слоев (fully-connected layers) на выходе. Результат операций на каждом слое подается на вход следующему. Особым слоем можно считать операцию обратного распространения ошибки (error backpropagation), который позволяет переоценить веса в сети.

Сверточный слой

Пусть – сверточный слой. На вход слой принимает карт признаков (feature maps) из предыдущего слоя, каждая размером . На первом слое ( ) на вход подаются сырые данные. На выходе слой возвращает карт признаков размерностью . Карта признаков рассчитывается следующим образом:

где – матрица смещения,

– фильтр размерности , связывающий карту признаков слоя с картой признаков слоя , размерности фильтра получены из равенств:

Часто фильтры, используемые для расчета фиксированной карты признаков одинаковы, т.е. для .

Уравнение (8) может быть переписано как многослойный перцептрон. Каждая карта признаков слоя *l* состоит из юнитов. Каждый юнит в позиции рассчитывается как

Фильтры и матрицы смещений содержат веса, значения которых подбираются во время моделирования.

Нелинейный слой

Особенность нейронных сетей в том, что они способны моделировать сложные нелинейные зависимости. Поэтому обязательным также является нелинейный слой.

Пусть – нелинейный слой, входные данные задаются картами признаков, а на выходе получаем карт признаков, каждая размерностью и , тогда

где – функция активации для слоя .

Ректификация

Этот слой может быть рассмотрен как часть нелинейного слоя или выделен как самостоятельная операция. Пусть – слой ректификации. На входе он принимает карт признаков размерностью и рассчитывает абсолютное значение для каждой компоненты карт признаков:

На выходе снова получаем карт признаков той же размерности. Слой ректификации обычно обозначают .

Субдискретизация

Следующий слой нужен для сокращения размерности матриц, с одной стороны, и повышения устойчивости карт признаков к шуму и искажениям.

Пусть – субкредитизирующий слой (pooling layer). На выходе он дает карт признаков уменьшенного размера. В общем случае суть операции пулинга состоит в рассматривании данных через окно (окно сдвигается с шагом (stride), чтобы не допустить перекрытия одного и того же фрагмента карты признаков). Из всего окна выбирается только одно значение, что позволяет в конечном счете значительно сократить размерности карт признаков. Наиболее распространена субдискретизация max pooling – в этом случае берется максимальное значение из каждого окна. Слой называют .

Сокращение размерности карт признаков позволяет намного быстрее достичь сходимости на стадии обучения.

Полносвязный слой

Пусть – полносвязный слой. Если предыдущий слой также полносвязный, можно применить уравнение

где

– выход слоя *l*,

– весовой коэффициент, связывающий *j*-й юнит слоя *(l-1)* с юнитом *i* слоя ,

– число юнитов слоя ,

получают с помощью алгоритма обратного распространения ошибки (error backpropagation),

– функция активации. Обычно имеет вид пограничной (threshold function), частично-линейной или сигмоидальной функции.

В противном случае, слой получает карт признаков размерности , -й юнит рассчитывает:

где

– весовой коэффициент, связывающий юнит в позиции *j*-й карты признаков слоя и *i*-й юнит слоя .

На практике для целей классификации часто используется более одного полносвязных слоев. Этот слой также включает нелинейность (функция преобразования не выделяется в отдельный слой).

В зависимости от специфики задачи могут быть подобраны различные архитектуры, в том числе включающие несколько слоев свертки и/или пулинга.

Предпоследний слой сети для задач классификации должен преобразовать выход предыдущих слоев в метки класса или, что еще лучше, в вероятности принадлежности наблюдения к некоторому классу. Для этого на предпоследнем слое используется softmax преобразование.

Функция потерь

Последний слой – слой функции потерь (cost or loss function), сопоставляющей результат работы алгоритма с истинной меткой класса. Как правило, для задач классификации в нейронных сетях используют функцию кросс-энтропии:

где – k-й элемент целевой переменной ,

– предсказание нейронной сети.

Для обучения нейронной сети может использоваться стохастический метод (входное значение выбирается случайно, веса обновляются на основе ошибки) или метод мини-батчей (случайные входные значения объединены в выборки, веса обновляются на основе общей ошибки выборки .

Алгоритм обратного распространения ошибки

Для того, чтобы подобрать оптимальные веса , минимизирующие функцию потерь , используются различные модификации алгоритма стохастического градиентного спуска. Алгоритм обратного распространения ошибки нужен для оценки градиента функции потерь на каждой итерации.

Описание алгоритма:

1. Прямой прогон модели: входные данные используются моделью для расчетов в каждом нейроне и получения финального предсказания. В качестве весов используются значения, полученные на предыдущем слое, или (для первого слоя) инициализированные случайным образом.
2. Для каждого выходного нейрона рассчитывается ошибка :

3. Для всех скрытых слоев l рассчитывается ошибка :

4. Оценка производных:

Алгоритм обратного распространения ошибки позволяет после каждой итерации обновлять веса между нейронами, улучшая качество модели.

Искусственные нейронные сети с использованием алгоритма свертки позволили значительно улучшить качество прогноза в задачах классификации изображений и текстов.

Сложность моделирования с помощью нейронной сети, с одной стороны, состоит в большом числе параметров (количество скрытых слоев, слоев свертки и субдискретизации, размер шага и окна, количество нейронов на скрытых слоях, выбор функций активации и др.). Помимо вопроса о правильной спецификации параметров модели возникают также такие проблемы как

- переобучение модели на тренинговых данных – решается с помощью изменения параметров скрытых слоев, использования более крупных батчей данных для обучения, уменьшения числа прогонов модели, добавление дропаута [84];

- несходимость алгоритмов градиентного спуска или достижение только локального минимума – решается с помощью различных алгоритмов инициализации весов, выбора альтернативных методов оптимизации на основе градиентного спуска;

- процесс обучения время- и ресурсозатратен.

ГЛАВА 2

СБОР, ОБРАБОТКА И АНАЛИЗ НОВОСТЕЙ

2.1 Особенности подготовки текстовых данных

Единицами text mining ветви машинного обучения являются документ и коллекция документов. Документов – это единственное текстовое сообщение, в данном случае новость. Коллекция документов – набор текстов. Коллекции могут быть статичными, что означает, что они не меняются с течением времени, или динамичными, то есть возможно обновление документов в коллекции или добавление новых. Признаком в документе может быть буква, цифра, символ слово, термин, концепция. Даже небольшой документ может быть представлен большим набором признаков, такой набор признаков называют репрезентационной моделью документа (representational model).

При представлении текста как на уровне букв могут использовать представления как набора букв (bag-of-characters approach) или биграм/триграм (bigrams/trigrams), в которых учитывается позиция букв. Это наиболее полное, но не оптимальное представление текста.

Представление текста на уровне слов не имеет значительного преимущества над набором букв.

Следующим уровнем представления является термин. Термины (terms) это слова или фразы из нескольких слов выбранные из корпуса текста методами term-extraction. Как правило, такие методы конвертируют сырой текст документа в набор нормализированных терминов (проведена токенизация, лемматизация и слова разобраны по частям речи). Иногда используется сторонний лексикон для нормализации терминов.

Наконец, концепции (concepts) – это признаки документа, созданные при помощи ручной, статистической, основанной на правиле или гибридной категоризации. Чаще всего они генерируются для документа комплексными метода обработки, которые идентифицируют отдельные слова, словосочетания и фразы. Это наилучший метод представления текста, так как решает такие проблемы как синонимы, многозначные слова и др. Однако это сложный и трудоемкий метод.

Текст представляет собой неструктурированные данные. Иногда выделяют «слабо» структурированные данные – это документы, в которых есть некая общая схема и разметка. Например, научные публикации, законодательные акты или новости. «Полуструктурированными» данными называют документы, содержащие метаданные, метки полей. Например, email-сообщения, веб-страницы HTML, PDF файлы.

Помимо отсутствия структуры в данных, при работе с ними возникает ряд сложностей:

- значение слова может зависеть от контекста;
- необходимость различать синонимы и многозначные слова;

- таблица частот слов обычна имеет большую размерность и сильно разрежена;

- в текстах могут встречаться опечатки, ошибки, различные варианты написания, аббревиатуры.

Miner, Elder, etc. выделяют следующие типы работы над текстом:

- поиск и получение информации: хранение и использование текстовых документов, поиск внутри документа и поиск по ключевым словам;

- кластеризация документа – группирование и категоризация терминов, параграфов или документов используя data mining clustering методы;

- классификация документов – группирование и категоризация отрывков, глав или документов с помощью data mining classification методов, обученных на разменных данных;

- web mining – преобразование данных и текстов в Интернете;

- natural language processing (NLP) – низкоуровневое преобразование языка (например, определение частей речи, токенизация, лемматизация);

- concept extraction – группирование слов и фраз в семантические близкие группы.

Вне зависимости от выбранного метода работы с текстом, техники text mining предполагают, что текст прошел предварительную подготовку. Зачастую качество работы text mining техник определяется тем, как были подготовлены данные. Подготовка текста (или нормализация) включает следующие процедуры:

- исключение новостных сообщений с нерелевантной информацией. Пустые записи или записи, не относящиеся к предмету анализа, добавляют шум в задачу классификации;

- преобразование текста к нижнем регистру. Эта процедура не меняет значения слова и позволяет программам избежать ошибки распознавания одного слова как два по причине заглавной буквы;

- коррекция ошибок и опечаток. Для подбора наиболее близких слов могут использоваться словари;

- исключение стоп-слов. Стоп-слова – это часто встречающиеся слова, которые не добавляют дополнительной информацией и специфичны для конкретного языка. К стоп-словам относятся предлоги, союзы, артикли. Некоторые списки включают около 400 – 500 стоп-слов английского языка;

- исключение лишних пробелов, любых пунктуационных символов и цифр;

- лемматизация (stemming). Это процесс исключения суффиксов и префиксов из слова на основе гипотезы, что общая лемма или корень слова в большинстве случаев описывают одну и ту же сущность в тексте.

Porter (1980) заключил, что качество работы систем обработки информации улучшается, если группа терминов преобразована в единственный термин [3]. Например, “connect”, “connected”, “connection”, “connections”, “connecting” имеют общую лемму “connect” и значение связи, соединения.

Предварительная обработка данных позволяет значительно повысить качество алгоритмов анализа данных благодаря сокращению размерности данных и кластеризации одинаковых и схожих по значению слов и фраз.

Нормализация текста

Текстовые данные для построения подобной модели могут иметь различные источники и типы, например, социальные медиа, блоги и форумы, онлайн-новости. Большинство источников используют крупные новостные порталы, такие как The Wall Street Journal, Financial Times, Reuters, Dow Jones, Bloomberg, Forbes, Yahoo! Finance и так далее. Из всех типов новостей обычно выбирают общие или финансовые новости. Большинство моделей используют финансовые новости, поскольку в них содержится меньше шума, излишней информации, чем в общих новостях. С сайта изымают текст или заголовки новостей.

В данной работе использовались новостные данные на английском языке. Ежедневно в полночь они сохранялись с сайта Reuters.com с раздела Markets News – новости рынка [41]. На странице публикуются заголовки новостей и краткое содержание. Заголовки, как правило, мало информативны, могут содержать преувеличения или игру слов для привлечения внимания читателей. Однако машинное обучение плохо справляется с такого рода задачами, поэтому в данной работе анализируются непосредственно краткие описания новостей (см. пример ниже).

Первый шаг в нормализации – это очистка текста от нерелевантной информации.

На примере случайного набора записей, из особенностей текста сразу можно выделить следующие моменты:

- текст скопирован из HTML и содержит такие символы как «\n» для определения разрыва строки. Такие сочетания должны быть заменены пробелом;

- часть сообщений прерываются (пример, сообщение 12). Это не должно вызвать проблемы для машинных алгоритмов;

- часть сообщений написана прописными буквами. В любом случае, во время обработки данных весь текст будет преобразован к нижнему регистру;

- часть сообщений начинается со знака «*». Судя по всему, так выделяют новости, относящиеся в большей степени к отдельной компании, нежели ко всем рынкам. Неясно, стоит ли включать данные новости в анализ: крупные изменения в жизни монополистической компании могут оказать влияние на расположение сил на валютном рынке в том числе. Однако информации в этих сообщениях может оказаться недостаточно: после удаления цифр и дат сообщение может оказаться слишком неинформативным даже для человека;

- в некоторые сообщения включены ссылки (см. девятое сообщение в таблице). В данном случае сообщение содержит ссылку на новостную подписку и не имеет ценности для анализа. Также похоже, что вместе с информацией о рассылке в одной поле вышли несколько новостных сообщений. Подобные тексты должны быть удалены;

- в скобках, как правило, указаны суммы или ссылки на источник информации. Эти данные также малоинформативны для анализа и могут быть исключены.

Таблица – Пример «сырого» текста новостей

Текст новостей
[1] "Mexican miner and\ninfrastructure company Grupo Mexico on Thursday priced an\ninitial public offering of its rail unit at 31.50 pesos per\nshare, two sources said, raising around 19 billion pesos ($998\nmillion)."
[2] "Mexican tycoon Carlos Slim's\nbank Inbursa said on Thursday that a unit expects to see profit\nof 4.65 billion pesos ($245 million) from the sale of GMexico\nTransportes shares."
[3] "Workers at a Chinese-owned auto glass\nplant in southwestern Ohio voted heavily against union\nrepresentation, the United Auto Workers (UAW) said on Thursday,\nin a major blow for the union's strategy for organizing\nforeign-owned auto factories."
[4] "* BNK Petroleum Inc -average production was 1,097 barrels\nof oil\nequivalent per day for Q3 2017, an increase of 7% compared to Q3\n2016 production"
[5] "* Alterra Power announces results for the quarter ended\nseptember\n30, 2017"
[6] "* It will export $2.2 billion in vehicles and parts from the\nUnited States to its China joint venture SAIC Motor\nover next three years\nFurther company coverage:\n(Reporting by Norihiko Shirouzu)"
[7] "* Senate Republicans propose 1-year delay in corporate tax\ncu
[8] "* Concerns rise as Senate tax bill differs from House\nversion"
[9] "To access the newsletter, click on the link:\n \n http://share.thomsonreuters.com/assets/newsletters/Indiamorning/MNC_IN_11102017.pdf\n \n \nIf you would like to receive this newsletter via email, please register at: https://forms.thomsonreuters.com/india-morning/\n \n \n FACTORS TO WATCH\n 08:30 am: Aurobindo Pharma post-earnings analyst conference call in Mumbai.\n 09:15 am: Mahindra Logistics lists on stock exchanges in Mumbai.\n 10:00 am: Prime Minister Economic Advisory Council"
[10] "Japan's Nikkei share average fell\nsharply on Friday morning hit by drops in tech shares after\ntheir U.S. counterparts languished overnight, while Toshiba Corp\nstumbled on dilution fears after media reported it will issue\nnew shares to raise funds."
[11] "South Korea's Hyundai Motor Co\nis considering building its Tucson and Kona sport\nutility vehicles and its pickup truck at its sole U.S. factory\nto reverse a sales slump, Seoul Economic Daily reported on\nFriday."
[12] "London Metal Exchange copper\nheld above one-month lows on Friday as a weaker dollar broadly\nlifted base metals, although it remained on track for a weekly\nloss.\n \"Copper prices will continue to trade sideways over the\ncoming weeks as a stronger U.S. dollar outlook and weakening\nChinese imports of refined copper weigh on prices,\" said BMI\nResearch in a report.\n \"High-frequency indicators such as stable stock levels, a\npersistently ne"
[13] "* Reuters Tankan closely correlates with BOJ tankan\n(Adds analyst's quote, detail)"

Примечание – Источник: собственная разработка

Таким образом, для очистки текста от излишней информации выполнены следующие шаги:

1) Выполнена проверка на непустые записи;

2) Удалены 260 сообщений со словами «This Diary is filed daily» и еще 289 со словами «Political and General News». Эти сообщение представляли

собой вставки между сообщениями, печатающиеся, как правило, в фиксированное время, и были не информативны. Также вставки со словами «The following are the top stories»;

3) Еще 113 записей были удалены из-за содержащейся информации о новостной рассылке;

4) Все ссылки на Интернет-ресурсы были заменены пробелом;

5) С помощью регулярных выражений удалены непечатные знаки, например, «<U+200D>», заменены пробелами разрывы строки «/n» и информация, заключенная в круглые скобки;

6) Выполнена проверка на сообщения, дублирующиеся по тексту и времени публикации. 1919 сообщений удалены.

Следующий этап – приведение текста к нижнему регистру.

Коррекция ошибок и опечаток выполняться не будет, поскольку текст взят с крупного новостного портала и вероятность неисправленных редакторами ошибок маловероятна, а с другой стороны, тексты содержат множество собственных имен.

Четвертый шаг – исключение стоп-слов. Использован список из 174 слов, включающий местоимения, вспомогательные глаголы, частицы, предлоги и союзы. Далее удаляются все пунктуационные знаки и цифры.

Наконец, заключающий шаг подготовки данных к анализу - лемматизация.

2.2 Разметка новостей относительно временного ряда

2.2.1 Подходы к разметке новостей

В работе A.K. Nassirtoussi et al. представлена общая схема проведения подобного анализа [66, с. 98]. На основе ее разработана схема, представленная ниже (рис. 1).

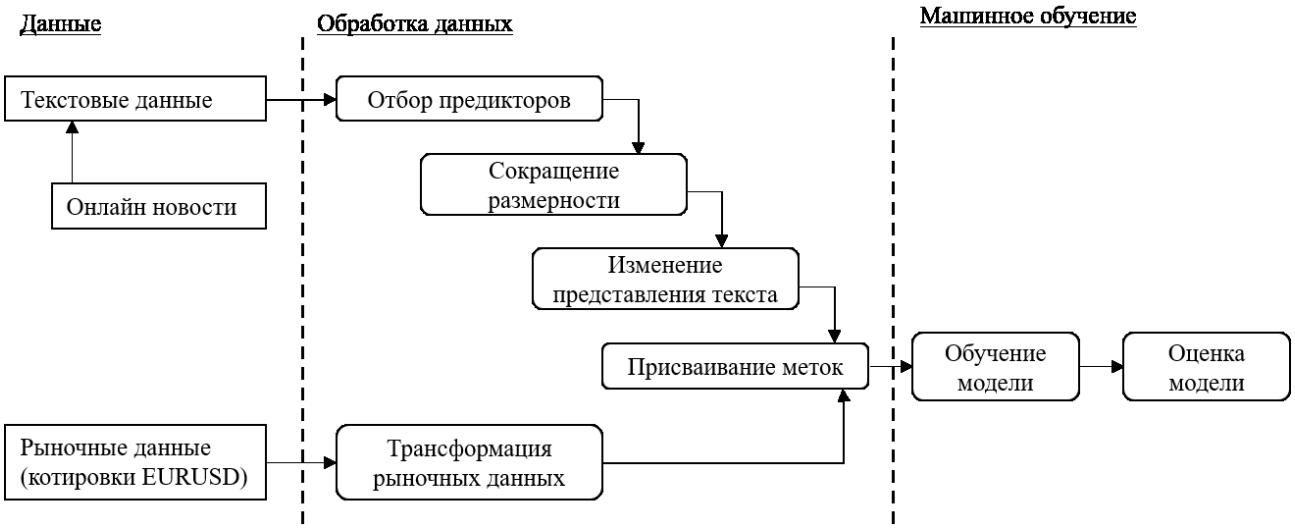

Рисунок 1 – Схема проведения анализа

Примечание – Источник: собственная разработка на основе [66]

Выше был рассмотрен процесс сбора и очистки данных, далее рассмотрим подходы к определению соответствия новости временному ряду [20]:

1) Наблюдаемый временной лаг. Предполагается, что воздействие новости отразится на финансовом рынке спустя некоторое время после публикации. Причем используется достаточно длительный период времени: в Lavrenko et al (2000) [36] лаг равен двум часам. Этот подход также использовался в работах Permunetilleke and Wong (2002) [44], Thomas and Sycara (2000) [72].

2) Эффективный рынок. Финансовый рынок отражает информацию сразу же после публикации новости без временного лага. Этот подход предполагает, что рынок эффективен и обычно нет возможности для арбитража.

3) Репортаж (reporting). Предполагается, что новостное сообщение публикуется уже после изменений на финансовом рынке. То есть движение рынка не определяется новостными сообщениями, и информация может быть использована только для отражения текущей ситуации, но не предсказания будущего.

Нет единого предпочтения одного из подходов, выбор может зависеть от природы самого рынка.

Далее представлен график изменения котировки евро/доллар в течение 30 апреля 2018 года (рис. 2).

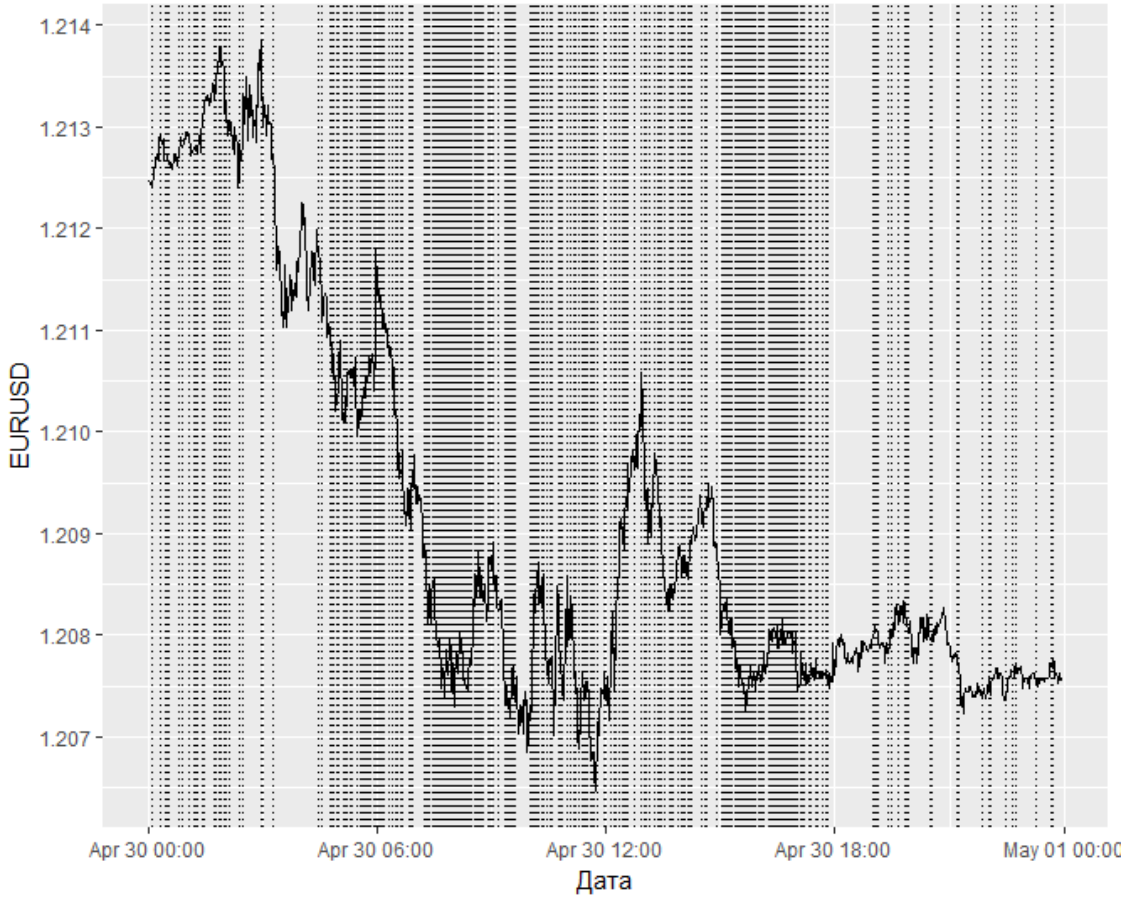

Рисунок 2 – Изменение котировки евро/доллар 30 апреля 2018 г. на фоне публикации новостей

Примечание – Источник: собственная разработка

Вертикальными линиями на графике отмечено время публикации новости. Видно, что в течение дня частотность новостей очень различна. 30 апреля большая часть новостей поступила в промежуток времени с 6:00 до 18:00.

Чтобы понять влияние новостей на динамику рынка рассмотрим ненасыщенный событиями интервал времени с 01:00 до 02:00 (рис. 3).

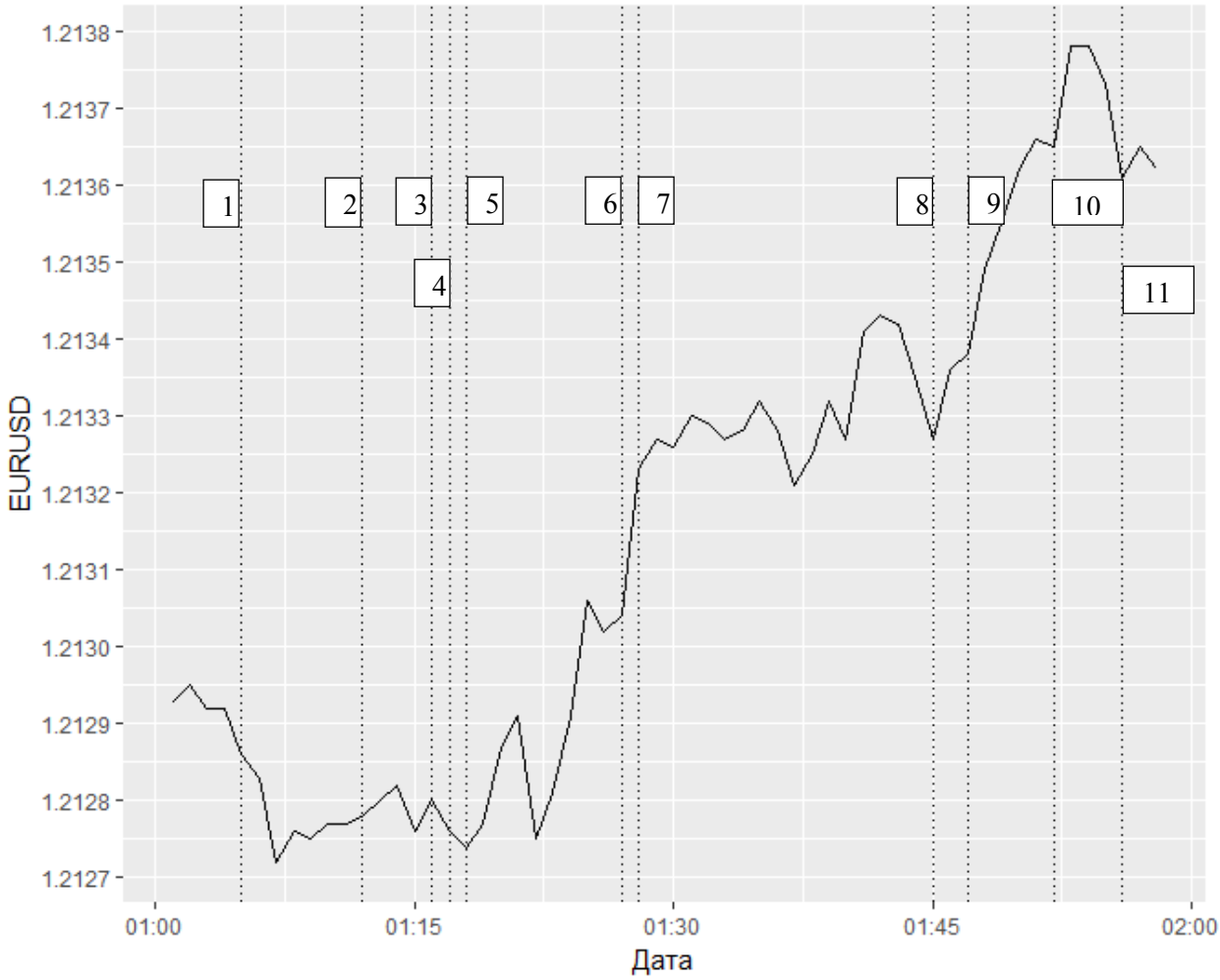

Рисунок 3 – Изменение котировки евро/доллар с 01:00 до 02:00 30 апреля 2018 г. на фоне публикации новостей

Примечание – Источник: собственная разработка

Во-первых, очевидно, что новости не включают в себя всю информацию, объясняющие колебания рынка. Какая-то информация может быть вовсе не представлена в новостях, также есть место и влиянию спекуляций.

Во-вторых, влияние некоторых новостей краткосрочно, в то время как после других сообщений котировки меняются на протяжении более длительного периода.

В-третьих, некоторым из новостей предшествовали изменения котировки, что может подтверждать репортажный подход к определению времени влияния новости на рынок.

Рассмотрим детально новости, появившиеся в данный период времени (возьмем текст до очистки, что облегчит анализ; табл.1):

Таблица 1 – Новости, опубликованные 30 апреля 2018 г. с 01:00 до 02:00

№	Время	Новость	Оценка
1	1:05	The world's biggest advertising group WPP reported better than expected first-quarter net sales and reiterated its full-year guidance in the first set of results to be published without founder Martin Sorrell.	0

2	1:12	Sainsbury's and Asda, the UK arm of Walmart, confirmed on Monday they are to merge to create Britain's biggest supermarket group by market share, surpassing current leader Tesco.	-1
3	1:16	* FDA TO CONDUCT PRIORITY REVIEW OF CEMIPLIMAB AS A POTENTIAL TREATMENT FOR ADVANCED CUTANEOUS SQUAMOUS CELL CARCINOMA	0
4	1:17	German monthly retail sales unexpectedly fell in March, data showed on Monday, marking a fourth consecutive drop and dampening cheer around a consumer-led upswing in Europe's biggest economy.	-2
5	1:18	* FY UNDERLYING PRETAX PROFIT ROSE 1.4 PERCENT TO 589 MILLION STG	0
6	1:27	The leader of the Northern Irish political party that props up British Prime Minister Theresa May said the European Union's chief negotiator was not an honest broker, the BBC reported on Monday.	-1
7	1:28	* Jitters over whether strong corporate earnings can continue	0
8	1:45	* PROVIDE OPERATIONAL UPDATE CONCERNING FORTHCOMING BACK-TO-BACK WELLS TO BE DRILLED AT PERMIAN BASIN ASSETS IN MITCHELL COUNTY	0
9	1:47	* NOW HOLDS FOLLOWING ECONOMIC INTERESTS IN US STRATEGIC INVESTMENTS	-1
10	1:52	Sainsbury's chief executive Mike Coupe said that it was possible that competition authorities would force some stores to be sold to allow the British supermarket chain's proposed merger with Asda to go ahead.	0
11	1:56	Britain's FTSE 100 index is seen opening 17 points higher on Monday, according to financial bookmakers, with futures up 0.2 percent ahead of the cash market open.	+1

Примечание – Источник: собственная разработка

Из 11 новостей, опубликованных в выбранный период времени, 7 касались конкретных компаний, нежели экономик стран в целом. Ожидается, что их влияние будет незначительным. Условно экспертную оценку влияния можно представить по шкале -2, -1, 0, 1, 2, где 0 соответствует отсутствию любого влияния новости на рынок, -1 – незначительному влиянию на снижение котировки, +1 – незначительному влиянию в пользу роста котировки, 2 и -2 – значительному влиянию на рост или снижение котировки соответственно. Оценка влияния основывалась на следующих предпосылках:

- новости отдельный компаний или городов не оказывают значительное влияние на рынок (оценка 0 для новостей под номером 1, 3, 5, 6, 8 и 10);

- новости, затрагивающие отдельный рынок одной страны, маркируются единицей. Например, новость под номером 2 касается всей отрасли ритейла в Великобритании и может повлиять на фондовые индексы (в то же время,

новость 10 лишь уточняет условия для совершения сделки и может иметь меньший эффект на рынок);

- наконец, новости, отражающие состояние экономики, были отмечены двойкой.

Оценка влияния новостей априори не может быть объективной. Во-первых, информация в новости может влиять разнонаправленно. Например, в 1:05 опубликована новость о поглощении британским офисом Wallmart другой компании. Вероятно, это позитивно повлияет на акции Wallmart, крупнейшего ритейлера США, что может в целом улучшить позицию США на фондовом рынке (при прочих равных) и отрицательно повлиять на котировку евро к доллару. В то же время, существует вероятность и повышения инвестиционной привлекательности Великобритании и улучшения обменного курса. Существует и третий вариант: если информация просочилась на рынок ранее, то текущие котировки уже отреагировали на новость и либо не изменятся вовсе, либо несколько откатятся по сравнению с предыдущим изменением.

Наиболее показательна в выбранном наборе данных четвертая новость о **внезапном** падении объема розничных продаж в Германии. Во-первых, эта информация наверняка должна отрицательно отразится на рынке. Во-вторых, слово «внезапно» дает некоторую гарантию того, что информация все же не была учтена в текущих ценах. Падение котировки произошло не сразу после публикации новости, а в момент с 1:21 до 1:22. За минуту котировка значительно снизилась с 1,21291 до 1,21275 евро за доллар США. Но если за начальный уровень изменения принимать момент публикации новости, то курс практически не изменился.

Рассмотрим, как решался вопрос о выборе величины лага другими исследователями.

1. Оценка влияния рассчитывалась для отдельных слов [1]. Для каждого слова собиралась статистика по его встречаемости (A) и по количеству дней, когда цена на рынке росла после публикации новости с данным словом (B). Влияние слово рассчитывалось как отношение B/A.

2. NewCATS категоризировал новости как «хорошие», «плохие», «нейтральные» [39]. «Хорошая» новость – это та, после которой рыночная цена выросла в течение 60 минут с пиком как минимум +3% от цены на момент публикации.

3. AZFinTex – это регрессионная система, разработанная для прогнозирования цен, а не классификации [67]. Т.е. каждая новость маркируется не как «хорошая» или «плохая», а привязывается к цене на рынке спустя 20 минут после публикации.

4. Pegah Falinouss предлагал два варианта привязки цены к группам новостей [20]. Первый вариант основан на гипотезе эффективного рынка, т.е. основывается на предположении, что лаг отсутствует вовсе. Вторая теория допускает существование положительного лага.

5. В Aase, et al. скомбинированы два метода присваивания меток новостям в зависимости от изменения дневных цен [2]. Первый схож с NewCATS системой и считает изменение средней, максимальной и

минимальной цены относительно момента публикации. Новость считается позитивной, если в среднем новость выросла больше, чем на некий параметр *a*, и максимальное ее изменение больше некоего параметра *b* (рис. 4). Второй метод использует метод наименьших квадратов, чтобы провести линию через цену момента публикации новости и цену спустя некоторое время. Если наклон этой линии больше некоторого параметра *k*, то новость считается «позитивной» или «негативной» в зависимости от направления наклона. В противном случае новость считается «нейтральной». Из общего массива новостей используются только те, которые были одинаково помечены двумя способами.

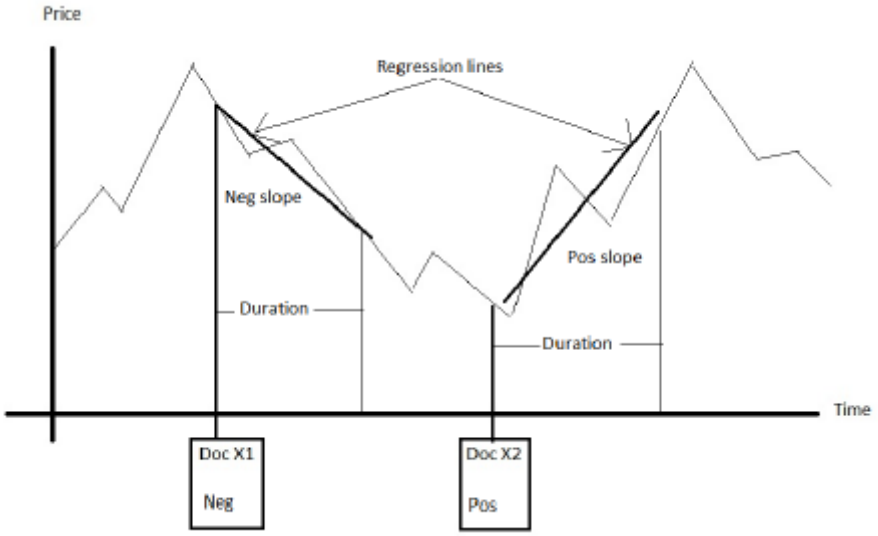

Рисунок 4 – Схема разметки новостей в Aase et al. [2]
Примечание – Источник: Aase et al. [2]

Два альтернативных подхода также используются при определении первой точки отсчета изменения цены. Согласно первому подходу, используется цена накануне публикации новости (цена открытия дня для дневных новостей и цена закрытия для ночных сообщений). Согласно второму, используют цену после публикации новости (цена закрытия дня для дневных новостей и цена открытия для предшествующих ночных сообщений).

Таким образом, новости за один день будут иметь одинаковые метки, даже если они совершенно противоположны по реальному влиянию на тренд цен. Поэтому дополнительно новости кластеризуют, чтобы выделить группы с негативным, позитивным или нейтральным влиянием.

6. Другой подход представлен в работе Beckmann [7]. Каждой цене на рынке присваивается категория: «Surge», «Plunge», «Not recommended» для высоких, низких и всех прочих значений цены за день. Далее новостям присваивается та категория, к которой они принадлежат согласно времени публикации.

7. Gidofalvi and Elkan (2003) использовали интервал 20 минут до и после публикации новости для определения волатильности цены в этот период [26]. Новости, соответствующие низкой волатильности, относились к категории неизменных, остальные делились на позитивные и негативные. Новости, опубликованные после закрытия, во время выходных и праздников не

рассматривались. Наконец, использовался наивный байесовский классификатор (Naïve Bayes text classifier), вычисляющий вероятность для каждой новой новости принадлежать к одному из классов.

8. Исследование Fung et al. (2002) следует гипотезе эффективного рынка, не допуская возможность лага между новостью и ее влиянием на рынок [24].

Далее будут рассмотрены три альтернативных варианта (рис. 5):

1. Следуя гипотезе эффективного рынка, предположим, что влияние новости на рынок происходит практически мгновенно. Тогда интервал изменения цены составит плюс-минус одну минуту с момента публикации новости;

2. Второй вариант – допустить существование лага между временем публикации новости и ее отражением на финансовом рынке;

3. Наконец, третий подход к разметке слов – репортаж, предполагающий, что вначале происходит изменение на рынке и лишь затем, с запозданием публикуется новость.

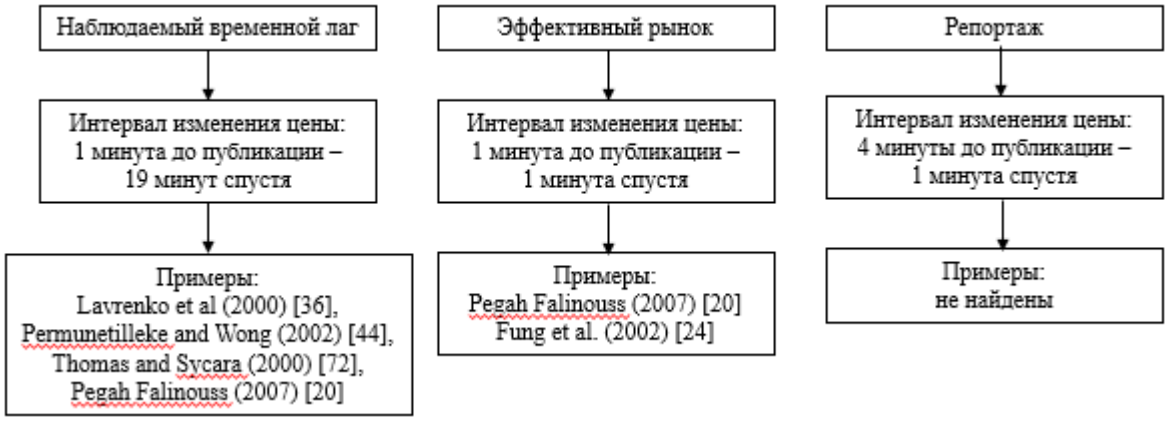

Рисунок 5 – Подходы к определению соответствия новости временному ряду, использованные в работе

Примечание – Источник: собственная разработка

2.2.2 Отсутствие значимого лага после публикации новости

Рассмотрим интервал изменения цены плюс-минус 1 минута. Как видно из рисунка 6, в таком случае почти все новости сосредоточены около нуля – отсутствие значимого изменения в курсе, связанного с публикацией новости. Если быть более точным, то 44,2 % новостей не отразились на рынке (изменение котировки равно 0,00 %), 55,8 % повлияли на курс евро к доллару (из них равное количество повлияло в положительную и отрицательную сторону).

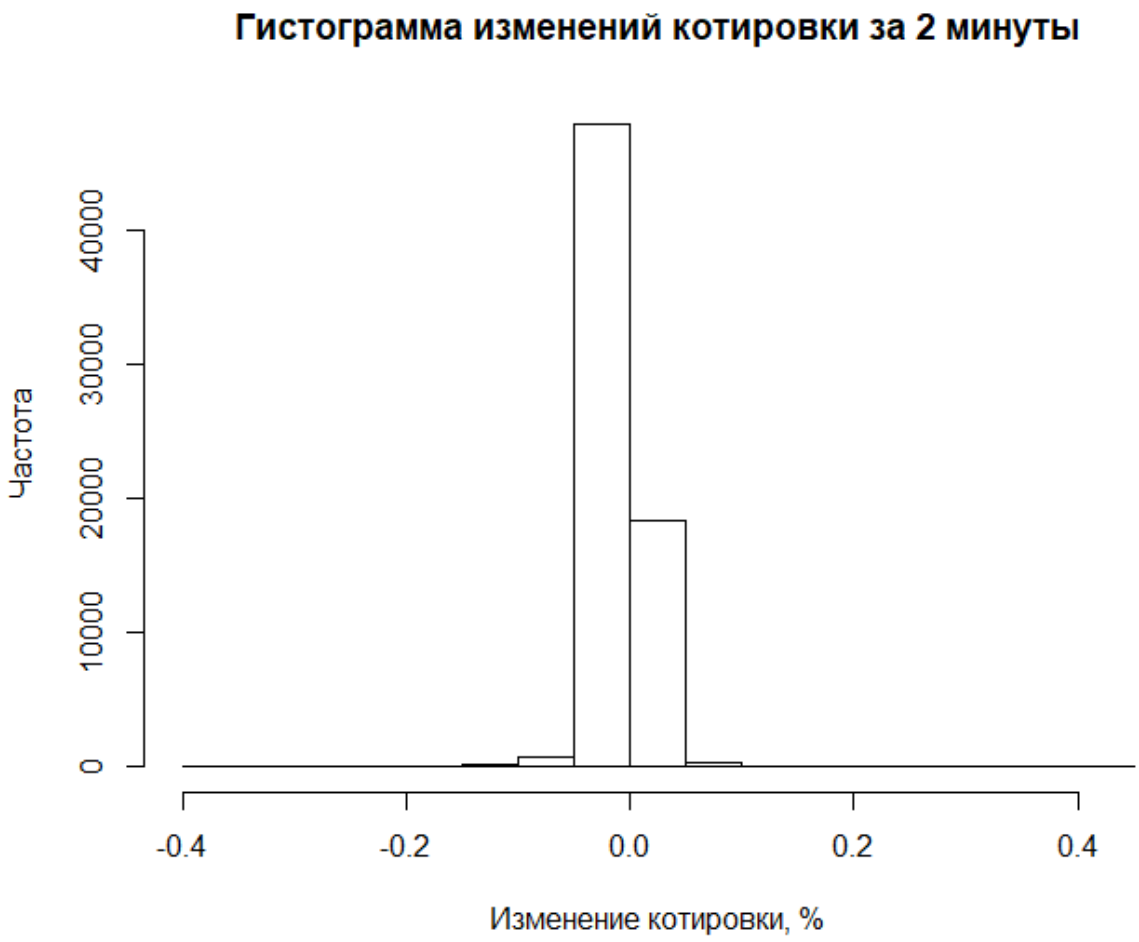

Рисунок 6 – Гистограмма изменений котировки за двухминутный интервал
Примечание – Источник: собственная разработка

Попробуем сравнить распределение слов по группам позитивных, нейтральных и негативных новостей (табл. 2).

Таблица 2 – Топ-10 часто встречающихся слов по группам новостей

Негативные новости	**Нейтральные новости**	**Позитивные новости**
compani (3963)	compani (7041)	compani (4022)
said (3029)	inc (4590)	said (3118)
us (2564)	coverag (4495)	us (2533)
inc (2382)	quarter (4422)	inc (2476)
million (2347)	million (4419)	million (2327)
coverag (2252)	announc (4233)	coverag (2258)
year (2182)	report (3596)	announc (1971)
announc (1930)	year (3409)	quarter (1917)
report (1846)	share (3089)	report (1818)
quarter (1883)	financi (2832)	share (1607)

Примечание – Источник: собственная разработка

Наиболее часто встречающиеся слова во всех группах: «компания», «сказать», «квартал», «миллион», «год», «США» и пр.

Сравним также биграммы и триграммы по трем группам (табл. 3).

Таблица 3 – Топ-10 часто встречающихся биграмм и триграмм по группам новостей

Негативные новости	**Нейтральные новости**	**Позитивные новости**
compani coverage (2470)	compani coverag (4334)	compani coverage (2499)
sec file (811)	sec file (1672)	sec file (819)
c us (786)	fourth quarter (1549)	c us (778)
fourth quarter (655)	per share (1488)	fourth quarter (676)
per share (619)	financi result (1368)	per share (595)
financi result (483)	first quarter (1179)	financi result (500)
first quarter (478)	full year (1002)	said tuesday (488)
said thursday (440)	earn per (953)	first quarter (476)
said tuesday (422)	earn per share (929)	said thursday (466)
full year (404)	file compani (862)	file compani (406)

Примечание – Источник: собственная разработка

Очевидно, что сколько-нибудь значимых отличий между тремя предложенными группами новостей нет, сообщения в них достаточно однородны.

Альтернативный подход заключается в том, чтобы предположить, что большинство слов равномерно распределены по группам новостей. В таком случае интересно выбрать слова и словосочетания, которые преимущественно сосредоточены лишь в одной из групп.

Для проверки будем сравнивать по отдельности частоты слов в группах позитивных и нейтральных новостей, а также в группах негативных и нейтральных новостей. Для каждого слова будет выдвигаться нулевая гипотеза, что частота встречаемости слова в группе позитивных (или негативных) новостей та же, что и в группе нейтральных новостей. В таком случае будем считать, что слово одинаково распространено в различных группах новостей и не вносит вклад в оценку характера новости. Альтернативная гипотеза заключается в том, что исследуемое слово встречается чаще в группе позитивных (негативных) новостей, чем в группе нейтральных новостей. Такое слово будет влиять на характер новости.

Для проверки гипотезы используем t-статистику Стьюдента с односторонней областью принятия решения для случая независимых выборок. Критерий позволяет найти вероятность того, что оба средних значения в выборке относятся к одной и той же совокупности. Статистика критерия для случая независимых выборок равна:

$$\frac{\quad}{\quad}$$

где

– среднее арифметическое в экспериментальной группе (позитивные или негативные новости),

– среднее арифметическое в контрольной группе (нейтральные новости),

– стандартная ошибка разности средних арифметических, которая находится из следующей формулы:

где и соответственно – величины первой и второй выборки,

.

Будем рассматривать только те слова, которые в общем числе новостей на тренинговом сете встречались более 100 раз.

После расчета статистики найдены 1007 слов, которые встречаются в позитивных новостях чаще, чем в нейтральных (на 0,1 %-м уровне значимости). Для негативных новостей статистика схожая – 956 слов чаще встречаются в них, чем в нейтральных новостях.

Ниже в таблице 4 представлены списки слов с наиболее высокой статистикой Стьюдента (значимой на 0,1 %-м уровне значимости) для двух сравнений: а) позитивных и нейтральных новостей, б) негативных и нейтральных новостей. В скобках указано количество новостей с приведенным словом.

Таблица 4 – Топ-10 слов, преобладающих в одной из групп по сравнению с группой нейтральных новостей

Слова, преобладающие в позитивных новостях	**Слова, преобладающие в негативных новостях**
bank (4695)	trade (3973)
trade (3973)	bank (4695)
european (2083)	european (2083)
thursday (4778)	thursday (4778)
add (2208)	friday (3301)
tuesday (4588)	central (1868)
friday (3301)	euro (1493)
minist (1529)	minist (1529)
market (4219)	add (2208)
yield (1595)	govern (2347)

Примечание – Источник: собственная разработка

Списки слов, преобладающих в негативных или позитивных новостях по сравнению с нейтральными сообщениями, пересекаются. Их можно рассматривать как слова, обладающие большим весом, большей важностью. Ведь их наличие в новости чаще приводит к смещению сообщения из зоны нейтральности.

В продолжение анализа сравним группы позитивных и негативных новостей. Таблица 5 ниже демонстрирует два списка слов. В первой колонке представлены слова, чаще встречающиеся в позитивных, чем в негативных или нейтральных новостях. В скобках указан уровень значимости: *** - значимо на 1 %-м уровне, ** - на 5 %-м уровне, * - на 10 %-м уровне. В правом столбце

представлены наиболее значимые слова, преобладающие в негативных новостях, по сравнению с позитивными и нейтральными.

Таблица 5 – Топ-10 слов, преобладающих в группе позитивных или негативных новостей по сравнению с другой группой

Слова, преобладающие в позитивных новостях	**Слова, преобладающие в негативных новостях**
johnson (***)	europ (***)
januari (***)	germani (***)
tuesday (***)	london (***)
power (***)	billion (***)
parliament (***)	buy (***)
domest (***)	monday (***)
uncertainti (***)	ahead (***)
solid (***)	war (***)
decis (***)	day (***)
arabia (***)	china (***)

Примечание – Источник: собственная разработка

Слова, способствующие уменьшению курса евро/доллар, относятся в большей степени к европейским странам (“europ”, “germani”, “london”, “itali”). Слова, увеличивающие курс евро к доллару, относились скорее к теме политики и финансовых рынков: “parliament”, “uncertainti”, “decis”, “sanction”, “power”, “stabil”, “dow”, “johnson”.

В целом, в группе слов, характерных только для позитивных новостей, 351 слово встречается чаще, чем в негативных новостях, на 5 %-м уровне значимости, из них 153 слова – на 1 %-м уровне. Среди негативных новостей таких слов 412 и 171 соответственно.

Проведем аналогичный анализ для биграмм и триграмм (табл. 6).

Таблица 6 – Топ-10 словосочетаний, преобладающих в группе позитивных или негативных новостей по сравнению с другой группой

Словосочетания, преобладающие в позитивных новостях	**Словосочетания, преобладающие в негативных новостях**
feder reserv (***)	european central (***)
three year (***)	european central bank (***)
said tuesday (***)	c london (***)
economy growth (***)	long term (***)
c euro zone (***)	global trade (***)
periphery govt (***)	us presid donald (***)
periphery govt bond (***)	month low (***)
zone periphery govt (***)	trade war (***)
govt bond (***)	treasuri yield (***)
gov bond yield (***)	us presid (***)

Примечание – Источник: собственная разработка

Словосочетания встречаются реже и их частоты более равномерно распределены по группам. В каждой группе новостей было отобрано около 650

биграмм и триграмм, которые встречались в новостях не менее 60 раз, поскольку нет смысла в анализе редко встречающихся словосочетаний. В результате, 146 словосочетаний преобладают в позитивных новостях на 5 %-м уровне и 159 словосочетаний более характерны для негативных новостей на том же уровне значимости.

В целом позитивные новости чаще включают слова, касающиеся финансовых рынков государственных ценных бумаг (“govt bond”, “periphery govt bond”, “feder reserv ”), негативные больше касаются монетарной политики (“european central bank ”, “c us treasuri”, “monetari polici”).

2.2.3 Значительный лаг после публикации новости

Далее допустим существование лага между временем публикации новости и ее отражением на финансовом рынке.

Для разметки новостей будем сравнивать состояние курса евро/доллар за минуту до публикации новости и спустя 19 минут после публикации.

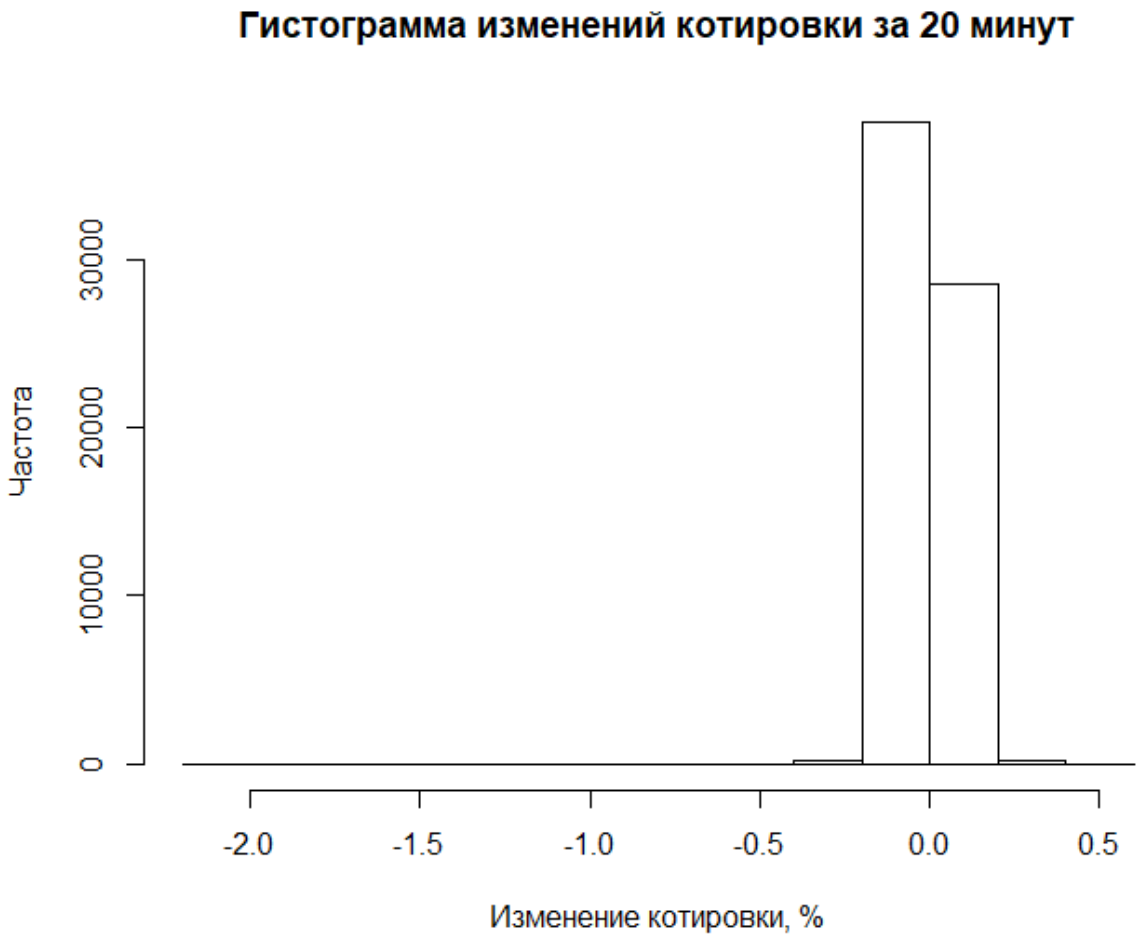

Рисунок 7 – Гистограмма изменений котировки за 20-минутный интервал
Примечание – Источник: собственная разработка

Распределение новостей смещено (рис. 7): только 12,9 % сообщений являются нейтральными, группы позитивных и негативных новостей составляют 42,9 % и 44,2 % соответственно.

Списки наиболее часто встречающихся слов очень похожи на списки из анализа двухминутного интервала: списки пересекаются со списками для двухминутного интервала примерно на 70 %. Рассмотрим группы слов, более свойственных позитивным или негативным новостям (табл. 7).

Таблица 7 – Топ-10 слов, преобладающих в группе позитивных или негативных новостей по сравнению с другой группой

Слова, преобладающие в позитивных новостях	**Слова, преобладающие в негативных новостях**
govern (***)	thursday (***)
fund (***)	money (***)
detail (***)	alleg (***)
africa (***)	european (***)
pct (***)	central (***)
replac (***)	execut (***)
zone (***)	interview (***)
russian (***)	friday (***)
bank (***)	weigh (***)
internet (***)	trade (***)

Примечание – Источник: собственная разработка

Теперь слова в каждом из списков более разнообразны и сложнее объединить их в кластеры. Вместе с тем, значимость частот по группам стала меньше, слова распределены по новостям более равномерно. В группе слов, преобладающих в позитивных новостях, только 23 % слов имеют значимость на 5 %-м уровне и только 7 % (65 слов) значимы на 1 %-м уровне. Схожая статистика наблюдается по группе слов, преобладающих в негативных новостях: 27 % слов имеют значимость на уровне 5 % и 10 % слов – на уровне 1 %.

На уровне биграмм и триграмм значимость также снизилась. 37 словосочетаний значимы в группе негативных новостей и 29 словосочетаний – в группе позитивных новостей на 5 %-м уровне значимости.

Очевидно, данный временной интервал является слишком большим и влияние новости нивелируется.

2.2.4 Репортажный подход

Для выбора интервала был проведен анализ промежутков времени от 3 до 20 минут с шагом 1 минута. Цель анализа была определить процент слов, имеющих значимость для группы позитивных или негативных новостей. Наибольшее число таких слов обнаружилось для 5-минутного интервала. Для остальных же промежутков времени распределение слов более равномерно по группам, значимость слов и словосочетаний много ниже, чем для двухминутного интервала, рассмотренного в самом начале анализа.

Для разметки новостей будем сравнивать состояние котировки валютной пары за 4 минуты до публикации новости и спустя 1 минуту. Распределение новостей по группам симметрично: 36.7 % новостей относятся к группе позитивных, 36.5 % - к группе негативных, оставшиеся 26.7 % формируют группу нейтральных новостей (рис. 8).

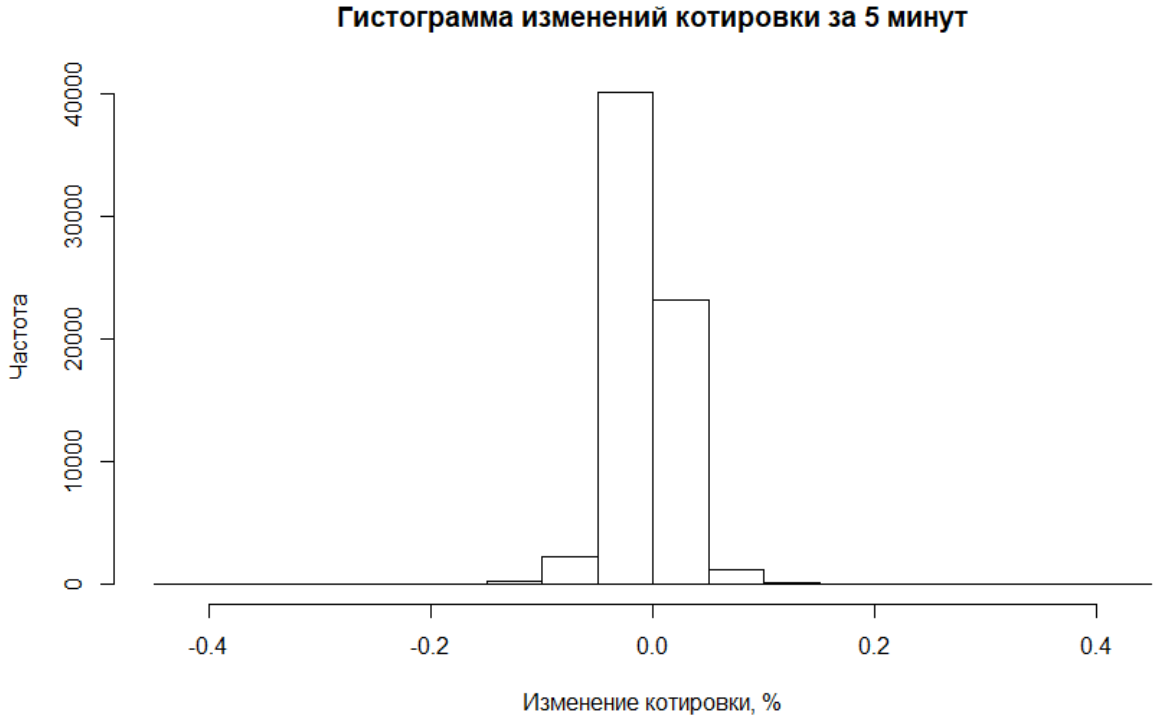

Рисунок 8 – Гистограмма изменений котировки за 5-минутный интервал
Примечание – Источник: собственная разработка

Сравним частоты встречаемости слов по трем группам новостей (табл. 8). Видно, что списки 10 наиболее значимых слов практически идентичны со списками на 2-хминутном интервале. В целом, они пересекаются более чем на 80 % для каждой группы новостей.

Таблица 8 – Топ-10 часто встречающихся слов по группам новостей

Негативные новости	**Нейтральные новости**	**Позитивные новости**
compani (5346)	compani (4274)	compani (5405)
said (3689)	inc (2829)	said (3873)
inc (3258)	quarter (2810)	inc (3360)
million (3239)	coverag (2714)	us (3177)
us (3191)	million (2703)	coverag (3151)
coverag (3140)	announc (2545)	million (3151)
year (2846)	report (2223)	announc (2825)
announc (2764)	year (2202)	quarter (2713)
quarter (2699)	said (2159)	year (2689)
report (2546)	us (2039)	report (2491)

Примечание – Источник: собственная разработка

Сравним также биграммы и триграммы по трем группам (табл. 9). Снова списки 10 часто встречающихся слов между тремя группами новостей очень

похожи и сильно пересекаются с анализом новостей на двухминутном интервале.

Таблица 9 – Топ-10 часто встречающихся биграмм и триграмм по группам новостей

Негативные новости	Нейтральные новости	Позитивные новости
compani coverage (3444)	compani coverag (2959)	compani coverage (3482)
c us (1177)	fourth quarter (1094)	sec file (1120)
sec file (1098)	sec file (999)	c us (1110)
fourth quarter (954)	per share (963)	per share (927)
per share (908)	financi result (929)	fourth quarter (919)
financi result (755)	first quarter (780)	first quarter (755)
first quarter (716)	full year (717)	financi result (707)
said thursday (634)	c us (709)	said tuesday (672)
central bank (625)	earn per (618)	central bank (627)
said tuesday (609)	earn per share (600)	said thursday (621)

Примечание – Источник: собственная разработка

Проведем анализ, аналогичный анализу для двухминутного интервала: рассчитаем t-статистики для выделения слов, не характерных для группы нейтральных новостей, затем сопоставим слова в позитивных и негативных новостях, чтобы выбрать наиболее значимые в каждой из групп. В таблице ниже представлены списки десяти наиболее значимых слов в группе позитивных или негативных новостей (табл. 10).

Таблица 10 – Топ-10 слов, преобладающих в группе позитивных или негативных новостей по сравнению с другой группой

Слова, преобладающие в позитивных новостях	Слова, преобладающие в негативных новостях
januari (***)	export (***)
dow (***)	court (***)
gas (***)	dip (***)
tuesday (***)	lose (***)
pct (***)	turkey (***)
lawyer (***)	talk (***)
close (***)	greec (***)
anoth (***)	surg (***)
policymak (***)	track (***)
govern (***)	target (***)

Примечание – Источник: собственная разработка

Среди характерных для позитивных новостей слов часто встречаются слова, относящиеся к финансовым рынкам («dow”, “pct”, “nasdaq”) и политическим решениям (“govern”, “policymak”, “agenc”). Для негативных новостей характерно упоминание европейский стран (“germani”, “Italian”, “europ”).

Однако значимость слов значительно снизилась по сравнению с двухминутным интервалом. В группе слов, характерных только для позитивных новостей, 198 слов встречаются чаще, чем в негативных новостях, со значимостью 5 % (на уровне 1 % - всего 71 слово). Среди негативных новостей преобладающих слов 147 на уровне 5 % (или 69 слов на уровне 1 %).

Наконец, проведем аналогичный анализ для биграмм и триграмм (табл. 11).

Таблица 11 – Топ-10 словосочетаний, преобладающих в группе позитивных или негативных новостей по сравнению с другой группой

Словосочетания, преобладающие в позитивных новостях	**Словосочетания, преобладающие в негативных новостях**
s p (***)	add detail (***)
p pct nasdaq (***)	c us (***)
p pct (***)	donald trump (***)
s p pct (***)	european central (***)
pct nasdaq (***)	presid donald trump (***)
pct s (***)	european central bank (***)
pct s p (***)	unit state (***)
dow pct (***)	c new (***)
said tuesday (***)	c dollar (***)
nasdaq pct (***)	trade war (***)

Примечание – Источник: собственная разработка

Топовые выражения в позитивных новостях оказались связаны с показателями финансовых рынков, такими как SP, Nasdaq, Dow Jones. Негативные словосочетания в основном касаются президента США или Европейского центрального банка, спектр тем охватывает торговую политику. Однако значимость словосочетаний намного ниже, чем в случае двухминутного анализа. Всего около двух десятков словосочетаний значимы на уровне 1 %.

Таким образом, в данной главе были проведены следующие этапы работы: сбор новостей с сайта Reuters.com, очистка их от шума и нормализация текста, следуя правилам, описанным в литературе по теме обработки текста, а также проведен анализ с целью разметки новостей по классам. В рамках последнего рассмотрены три подхода к влиянию новостей на рынок. Согласно гипотезе эффективного рынка, рассмотрено предположение об отсутствии лага при влиянии новости после публикации, метки новостям были проставлены на основании изменении курса валютной пары в течение двух минут (за минуту до публикации и минуту спустя). Второй рассмотренный вариант предполагал существование значительного лага – 20 минут – во влиянии новости на рынок. Наконец, третий подход соответствовал гипотезе репортажа о том, что новость публикуется уже после того, как воздействие на рынок оказано.

Наибольший интерес для последующего анализа представил собой подход эффективного рынка, так как в случае двухминутного интервала слова оказались распределены по группам наиболее неравномерно. Это позволяет предположить, что использование текста для предсказания направления

влияния новости будет наиболее эффективно как раз в случае двухминутного интервала.

ГЛАВА 3

ИСПОЛЬЗОВАНИЕ НОВОСТЕЙ ДЛЯ ПРОГНОЗИРОВАНИЯ КОТИРОВКИ

3.1 Создание классификатора новостей на основе слов и словосочетаний

В предыдущей главе была выдвинута гипотеза о возможности использования новостей для предсказания направления изменения котировки на двухминутном интервале. Попробуем вручную создать простой классификатор, который будет предсказывать движение на финансовом рынке на основе публикуемых новостей. Если такой классификатор будет работать, он может быть использован для принятия решений о покупке-продаже евро или доллара на краткосрочных интервалах.

Для предварительного анализа, проведенного в предыдущей главе, использовалась часть выборки – обучающая выборка. Она представляла собой данные за период времени с 10 ноября 2017 года по первое августа 2018 года, что составило 67 154 новости. Оставшаяся часть данных до 19 октября 2018 года образует тестовую выборку, на которой будет оцениваться качество классификатора. Размер тестовой выборки – 6 442 новости.

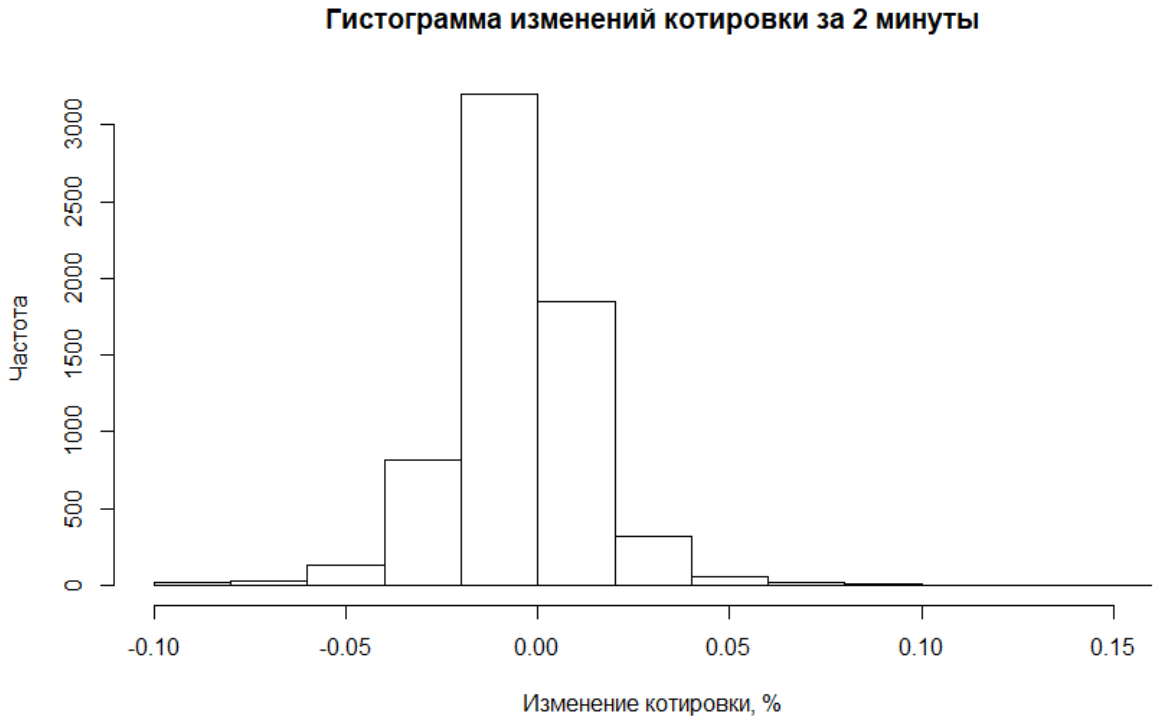

Рисунок 9 – Гистограмма изменений котировки за 20-минутный интервал
Примечание – Источник: собственная разработка

Для присвоения новостям истинных меток о классе используем двухминутный интервал (рис. 9) и правило, принятое на обучающей выборке. Если через минуту после публикации новости котировка выросла по сравнению с моментом времени за минуту до публикации, то новости относится к категории позитивных. Если за это время котировка снизилась, - к категории негативных новостей. Наконец, если значимых изменений на рынке не произошло, то новость считается нейтральной. Следу этому правилу, на тестовой выборке имеем 2220 негативных, 1973 нейтральных и 2248 позитивных новостей. В процентном соотношении это 34,5 % негативных, 30,6 % нейтральных и 34,9 % позитивных новостей.

Всю совокупность слов на тестовой выборке представим как вектор длиной 6 691 элементов (стоп-слова исключены). Второй вектор – вектор биграмм и триграмм, длиной 141 400 элементов. В данном случае не будет проводить очистку редко встречающихся словосочетаний, поскольку их частота может быть высокой на обучающей выборке.

Теперь каждой новости на тестовой выборке можно присвоить векторы слов и словосочетаний. Каждому элементу вектора присваиваем нуль, если это слово или словосочетание не встречается в новости, или единицу в обратном случае. Если объединить векторы слов по всем новостям, получим TF-IDF матрицу, показывающую частоты встречаемости слов по документам (в данном случае, по новостным сообщениям).

Далее единицы в векторах можно заменить, используя информацию из обучающей выборке о том, какие слова более свойственны негативным или позитивным сообщениям. В таком случае для слов из позитивной группы можно оставить единицу, а для слов из негативной группы – заменить ее на минус единицу. Если просуммировать значения вектора слов для каждой конкретной новости, можно получить условный индекс, который можно интерпретировать как преобладание положительных или негативных слов в новости. Его можно использовать как классификатор: если индекс больше нуля, новость классифицируется как позитивная и спустя минуту после ее публикации ожидается увеличение котировки евро-доллар.

В таком простом подходе никак не используется информация о словосочетаниях и о том, что слова имели разную статистическую значимость. Поэтому предложен подход с использованием весовых коэффициентов. Знак весового коэффициента будет зависеть от того, характерно данное слово или словосочетание для позитивных или негативных новостей. Величина веса зависит от уровня значимости статистики Стьюдента, рассчитанной для слова или словосочетания. Предлагается присвоить следующие веса: 0 – если лексическая единица не значима, +/- 0.3 – в случае значимости на 10 %-м уровне, +/- 0.6 – значимость на 5 %-м уровне и, наконец, +/- 1 – для значимости на 1 %-м уровне.

Рассмотрим на примере конкретной новости: "Qualcomm Inc has signed $12 billion worth of deals with three Chinese mobile handset makers on the sidelines of a state visit to Beijing by U.S. President Donald Trump."

Представим данное выражение как нормализированный набор слов (табл. 12). Поскольку одному предложению соответствует множество комбинаций слов, образующих биграммы и триграммы, добавим в таблицу только те из них, которые является значимыми хотя бы на 10 %-м уровне.

Таблица 12 – Расстановка весовых коэффициентов словам и словосочетаниям на примере одного предложения

Слово	Значимость	Вес
qualcomm		0
inc		0
sign	**	0.6
billion	*	0.3
worth	**	-0.6
deal	*	0.3
three		0
chines	**	-0.6
mobil		0
handset		0
maker		0
sidelin		0
state		0
visit		0
beij	***	-1
us	*	-0.3
presid	*	-0.3
donald		0
trump	***	-1
us presid donald	***	-1
us presid	***	
presid donald	***	
donald trump	***	
presid donald trump	***	
chines mobil	*	0.3
…		0

Примечание – Источник: собственная разработка

Из 19 слов в предложении только 9 имели статистическую значимость для анализа. Весовые коэффициенты были проставлены им с учетом уровня статистической значимости.

Несколько более сложная ситуация возникает при анализе биграмм. Из одной комбинации слов «us presid donald trump» было создано пять словосочетаний: «us presid donald», «us presid», «presid donald», «donald trump», «presid donald». Все выражения имеют самую высокую значимость, свойственную новостям из группы негативных. Очевидно, что нет смысла считать каждое из них как самостоятельную единицу, дающую собственный

вклад в оценку новости. В таком случае одно упоминание Дональда Трампа практически однозначно гарантировало бы укрепление доллара США.

Чтобы не допустить такой эффект, добавим следующее правило для анализа биграмм и триграмм. В случае двух и более пересекающихся выражений (т.е. имеющих хотя бы одно общее слово) в одной новости, если эти выражения имеют оценку одного знака, в индексе они будут засчитываться как одно выражение с весов, соответствующем максимальному из весов объединенных словосочетаний. Говоря проще, если в новости встречается несколько словосочетаний с одинаковыми словами и все словосочетания относятся только к положительному или только к отрицательному классу, то они заменяются одним выражением того же класса. В данном случае, все пять словосочетаний о президенте США могут быть заменены одним выражением «us presid donald trump» с отрицательной оценкой и максимальным весом – минус единица.

Итого, общий индекс представленной новости – минус 3.3, новость предсказана как негативная. После ее публикации ожидается снижение курса евро к доллару США.

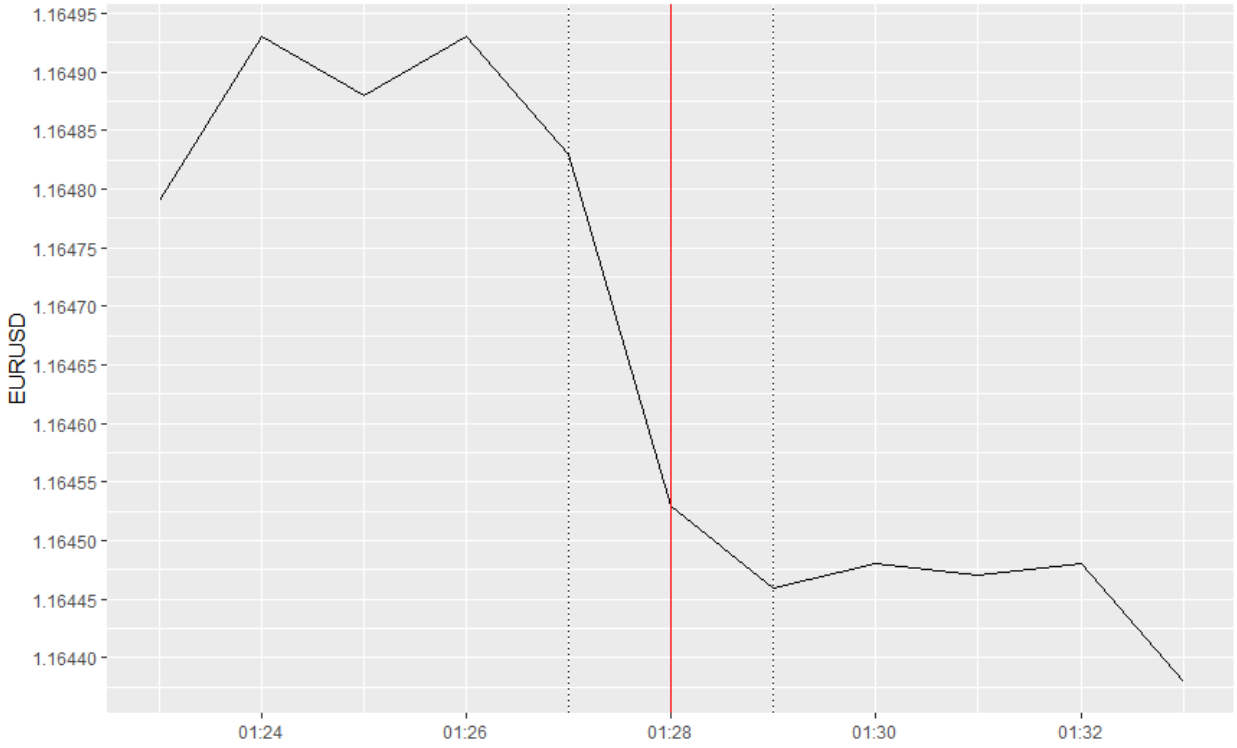

Рисунок 10– Изменение котировки в момент публикации новости
Примечание – Источник: собственная разработка

Рассмотрим, что в действительности происходило с котировкой в окрестности момента публикации новости (рис. 10). За минуту до публикации курс EURUSD был равен 1.16483, в момент публикации он опустился до 1.16453, спустя еще минуту – до 1.16446. За две минуты изменение составило 3.2 %. Таким образом, в выбранном временном интервале данная новость действительно оказала негативное влияние.

Распространим данный метод на весь массив новостей на тестовой выборке. Итоговая точность предложенного алгоритма оказалась 0.3305. Это очень низкий результат, который равносилен случайному угадыванию меток класса. Если отнести все новости к группе позитивных (преобладающий класс), точность составит 0.3490.

Матрица ошибок показывает, что практически все новости в рамках данного подхода были отнесены к числу нейтральных (табл. 13).

Таблица 13 – Матрица ошибок для алгоритма расстановки весов

		Предсказания модели		
		Негативная	**Нейтральная**	**Позитивная**
Истинные метки	**Негативная**	79	1567	574
	Нейтральная	71	1474	428
	Позитивная	92	1580	576

Примечание – Источник: собственная разработка

Таким образом, предложенный алгоритм не позволил выявить значимые взаимосвязи между словами и словосочетаниями в новости и ее влиянием на рынок. Попробуем применить к данным более сложные методы, относящиеся к машинному и глубокому обучению.

3.2 Использование методов машинного и глубокого обучения для категоризации новостей

3.2.1 Градиентный бустинг

Огромно разнообразие методов, которые исследователи пытались применить к проблеме классификации текста. Среди таких методов: иерархическая классификация, мультивариативные регрессионные модели, классификация методом К ближайших соседей, байесовский классификатор, решающие деревья, нейронные сети, метод опорных векторов [14].

В данной главе попробуем применить наиболее популярные методы машинного и глубокого, которые используются для категоризации текста. К ним относятся ансамбли деревьев (градиентный бустинг) и нейронные сети от простых до сверточных и рекуррентных.

Каждый из рассматриваемых алгоритмов обладает сложной архитектурой и имеет множество параметров. Для подбора оптимального набора параметров на обучающей выборке будет использоваться перекрестная проверка (кросс-валидация). Перекрестная проверка – это метод оценки аналитической модели и её поведения на независимых данных. При оценке модели имеющиеся в наличии данные разбиваются на k частей. Затем на k−1 частях данных производится обучение модели, а оставшаяся часть данных используется для тестирования. Процедура повторяется k раз; в итоге каждая из k частей данных используется для тестирования. В результате получается

оценка эффективности выбранной модели с наиболее равномерным использованием имеющихся данных.

В ряде публикаций авторы использовали один из видов градиентного бустинга для классификации текстов новостей.

Градиентный бустинг будет применен к новости на уровне слов. Подготовка текста проведена практически без изменений: текст очищен от шума, применен стемминг, текст превращен в двумерный тензор слов, стоп-слова и редко встречающиеся слова исключены из анализа.

Для подбора параметров используется кросс-валидация на 10 % данных.

Для применения данного алгоритма использованы следующие параметры (табл. 14):

Таблица 14 – Параметры градиентного бустинга

Параметр	**Значение**
Learning rate	0.1
Loss	“deviance”
Min impurity decrease	0.0
Min samples leaf	1
Min samples split	2
# estimators	100
Min weight fraction leaf	0.0

Примечание – Источник: собственная разработка

Для реализации использовался модуль GradientBoostingClassifier из библиотеки Python sklearn.

Оценим качество модели на тестовом наборе данных. Точность алгоритма (accuracy) равна 0.3493, матрица ошибок (confusion matrix) имеет следующий вид (табл. 15):

Таблица 15 – Матрица ошибок для градиентного бустинга

		Предсказания модели		
		Негативная	**Нейтральная**	**Позитивная**
Истинные метки	**Негативная**	2090	111	19
	Нейтральная	1825	133	15
	Позитивная	2092	129	27

Примечание – Источник: собственная разработка

Точность алгоритма находится на уровне случайного гадания. Матрица ошибок показывает, что алгоритм не научился точно классифицировать новости, несмотря на высокое качество на обучающей выборке.

3.2.2 Нейронная сеть

Далее попробуем применить искусственную нейронную сеть для текста на уровне слов. Искусственная нейронная сеть (artificial neural network; далее – ANN) имеет более простую архитектуру, чем CNN. В ней нет сверточного слоя

и слоя субдискретизации, архитектура состоит только из полносвязных слоев. Среди параметров модели функция инициализации, активации и число нейронов на каждом слое, возможен дропаут для предупреждения переобучения, для обучения всей сети выбираются метод оптимизации, функция потерь, число эпох и размер батча для обучения.

Архитектура предложенной нейронной сети имеет вид (табл. 16):

Таблица 16 – Архитектура искусственной нейронной сети

Номер слоя	**Функция инициализации**	**Функция активации**	**Число нейронов**	**Дропаут**
1	glorot_uniform	ReLU	25	0.2
2	glorot_uniform	ReLU	20	0.2
3	glorot_uniform	ReLU	10	0.2
4	glorot_uniform	ReLU	5	0.2
5	glorot_uniform	Softmax	3	-

Примечание – Источник: собственная разработка

Предложенная архитектура несколько более сложная, чем искусственная нейронная сеть, предложенная в исследовании Yashwant Singh Patel и Supriyo Mandal для классификации новостей, касающихся фондового рынка [63].

Для обучения искусственной нейронной сети используются следующие параметры:

- метод оптимизации: rmsprop;
- функция потерь: категориальная кросс-энтропия;
- число эпох: 100;
- размер батча: 32.

Точность алгоритма на тестовой выборке составила 0.3479, матрица ошибок представлена ниже (табл. 17).

Таблица 17 – Матрица ошибок для искусственной нейронной сети

		Предсказания модели		
		Негативная	**Нейтральная**	**Позитивная**
Истинные метки	**Негативная**	2109	96	15
	Нейтральная	1851	108	14
	Позитивная	2107	117	24

Примечание – Источник: собственная разработка

Таким образом, точность нейронной сети, обученной на словах, сравнима с качеством предсказаний градиентного бустинга. В обоих случаях, модель не выявила взаимосвязи в данных.

Следующие алгоритмы будут основаны на альтернативном представлении данных – непрерывный мешок со словами (continuous bag of words; CBOW). Это один из вариантов word2vec. Работа этой технологии осуществляется следующим образом: word2vec принимает большой текстовый корпус в качестве входных данных и сопоставляет каждому слову вектор, выдавая координаты слов на выходе. Сначала он создает словарь, «обучаясь»

на входных текстовых данных, а затем вычисляет векторное представление слов. Векторное представление основывается на контекстной близости: слова, встречающиеся в тексте рядом с одинаковыми словами, в векторном представлении будут иметь близкие координаты векторов-слов.

3.2.3 Простая CBOW модель

Начнем с примера неглубокой модели, включающей в себя слой с предварительной обработки данных для представления их как непрерывного мешка слов. Пример использования подобной модели можно найти в работе Ding X. et al., где использовались новости для предсказания цены акций [14].

Архитектура этой модели следующая:

1. Слой вложений (embedding layer). На нем происходит трансформация представления входных данных;
2. Слой субдискретизации по среднему;
3. Полносвязный слой, возвращающий 3 выхода из сети.

Для обучения искусственной нейронной сети используются следующие параметры:

- метод оптимизации: rmsprop;
- функция потерь: категориальная кросс-энтропия;
- число эпох: 100;
- размер батча: 256.

Для подбора параметров использовалась кросс-валидация на 10 % выборки. Точность алгоритма на тестовой выборке составила 0,3442 (табл. 18).

Таблица 18 – Матрица ошибок для искусственной нейронной сети с CBOW

		Предсказания модели		
		Негативная	**Нейтральная**	**Позитивная**
Истинные метки	**Негативная**	373	1388	459
	Нейтральная	246	1397	330
	Позитивная	365	1436	447

Примечание – Источник: собственная разработка

Очевидно, что переход к альтернативному представлению слов на примере простой модели не увеличил качество алгоритма. Результаты модели свидетельствуют об отсутствии значимой взаимосвязи между меткой новости и набором слов, в нее входящих. Попробуем оценить результаты на более сложной модели, используя предобученный массив вложений.

3.2.4 Сверточная нейронная сеть с использованием GloVe

GloVe расшифровывается как Global vectors for word representation. Это метод обучения без учителя, применяемый с целью получения векторного представления слов. Обучение было произведено на агрегированной статистике смежности слов. Для обучения алгоритмы авторы использовали огромные массивы текста из Википедии, Твиттера и Common crawl.

Результативность сверточных нейронных сетей, а также использования предобученных векторных представлений исследована для классификации различных текстов в работе Yoon Kim. В частности, там доказано, что использование готовых векторных представлений улучшает качество сверточных моделей.

Обученную модель GloVe можно применить к данным из новостей, чтобы получить более корректное векторное представление слов.

Ниже представлена архитектура сверточной сети, в которой для преобразования пространства входных данных будет использоваться предобученная модель GloVe (табл. 19).

Таблица 19 – Архитектура сверточной нейронной сети

Номер слоя	**Тип слоя**	**Функция активации**	**Параметры слоя**	**Дропаут**
1	Слой вложений (GloVe)	ReLU	300	-
2	Сверточный слой	ReLU	128, 3	-
	Субдескритизация по максимуму	-	3	0.2
3	Сверточный слой	ReLU	64, 3	-
	Субдескритизация по максимуму	-	3	0.2
4	Сверточный слой	ReLU	64, 3	0.2
	Flatten	-	-	-
5	Полносвязный слой	Softmax	3	-

Примечание – Источник: собственная разработка

Для обучения сверточной нейронной сети используются следующие параметры:

- метод оптимизации: Adam;
- функция потерь: категориальная кросс-энтропия;
- число эпох: 100;
- размер батча: 32.

Точность алгоритма на тестовой выборке составила 0.3535, матрица ошибок представлена ниже (табл. 20).

Таблица 20 – Матрица ошибок для сверточной нейронной сети

		Предсказания модели		
		Негативная	**Нейтральная**	**Позитивная**
Истинные метки	**Негативная**	912	1002	306
	Нейтральная	715	979	279
	Позитивная	851	1011	386

Примечание – Источник: собственная разработка

Использование предобученной модели для обработки входных данных и более сложного алгоритма не улучшило качество модели значимым образом. Точность предсказаний лишь немного выше точности наивной модели, которая предсказывала бы все новости одним классом.

Таким образом, было использовано пять методов, включая предложенный автором метод присваивания весовых коэффициентов словам и словосочетаниям и более сложные методы машинного и глубокого обучения. Качество работы всех методов было оценено на тестовом наборе данных, который не участвовал в обучении модели. Наилучший результат был достигнут при помощи сверточной нейронной сети, использующей предобученное векторное представление слов. Ни один из методов не смог подтвердить наличие значимой взаимосвязи между словами в тексте новости и ее влиянием на рынке. Следовательно, гипотеза о том, что на основании анализа только текста новости можно предсказать поведение котировки евро/доллар не была подтверждена и должна быть отвергнута.

ЗАКЛЮЧЕНИЕ

С развитием технологий значительно возросла скорость сделок на рынках всех видов. Финансовый рынок – один из самых волатильных: цены активов, котировки валют изменяются каждую секунду под влиянием огромного количества факторов. Один из таких факторов – информация в виде новостей. На курсы валют влияет информация разного характера, поступающая из самых разных источников: о политических и экономических решениях, военных конфликтах, изменении цен на сырьевые ресурсы, высказываниях представителей власти. Существует даже понятие вербальной интервенции: когда уполномоченный представитель крупной государственной или надгосударственной структуры высказывает мнение с целью влияния на валютный рынок.

Существует несколько теорий, концентрирующихся на возможности влияния информации на финансовый рынок. Первая – гипотеза эффективного рынка, предложенная Юджином Фама в 1946 году. В рамках гипотезы предполагается, что цена актива отражает всю доступную информацию и все участники рынка имеют доступ к информации. Теория Фама имеет три формы: слабую, среднюю и сильную. При низкой эффективности рынка цены учитывают только историческую информацию. Средняя форма предполагает, что цены также включают общедоступную информацию. Только те участники рынка, которые обладают частной информацией, могут получить выигрыш. Наконец, при высокой эффективности рынка цены учитывают и частную информацию, что делает невозможным получение прибыли участниками рынка.

Другой взгляд, похожий на теорию Фама при средней эффективности рынка, представляет гипотеза случайного блуждания. Согласно ей, цены на рынке случайны и прогнозирование неэффективно.

Из этих двух теорий сформировались два подхода к торговле на финансовом рынке – технический и фундаментальный анализ. В рамках первого используются только исторические данные о ценах для предсказания, как для рынка низкой эффективности. В рамках второго технический анализ дополняется исследование все доступной информации, в том числе новостей и макроэкономических данных.

Теория поведенческой экономики предполагает, что рынки не эффективны и случайное блуждание может быть объяснено человеческим поведением. Чтобы примирить поведенческую экономику с гипотезой эффективного рынка была разработана гипотеза адаптивного рынка. Суть ее сводится к тому, что периоды эффективности и неэффективности на рынке сменяют друг друга.

Несмотря на множество попыток исследовать отдельные рынки, до сих пор не сложилось однозначное мнение о том, выполняется ли гипотеза эффективного рынка. «Для каждой публикации, приводящей эмпирические доказательства рыночной эффективности, найдется противоречащая работа,

эмпирически подтверждающая рыночную неэффективность» [74]. Несмотря на противоречивость мнений относительно фондового рынка, можно заключить, что в большинстве случаев исследователи нашли подтверждение влияния информации на рынок и возможность для использования ее с целью прогнозирования. Намного менее изучен валютный рынок. В рассмотренных работах авторы приходили к выводу о малой возможности для прогнозирования такого рынка. Малая изученность рынка, в особенности такого важного показателя как котировка евро к доллару США, обусловила интерес к проведению исследования на данную тему.

Анализ публикаций показал, что не существует единого подхода к подготовке данных к анализу и моделированию. В большинстве публикаций авторы предпочитали прогнозировать направление движения цены построению регрессионных моделей. Спектр используемых алгоритмов включает классификацию на основе опорных векторов, регрессионные алгоритмы, наивный байесовский классификатор, деревья и комбинаторные алгоритмы [66].

Подготовка текстовых данных выполнялась по образцу большинства исследований с учетом особенностей собранного текста. Она включала очистку текста, исключение стоп-слов, пунктуационных знаков и чисел, лемматизацию.

Важным этапом было принятие решение об алгоритме присваивания меток новостям. В проанализированной литературе использовались самые различные временные интервалы между публикацией новости и изменением цены актива. В общем случае все подходы могут быть разделены на три группы. Первая предполагает существование лага после публикации новости и прежде, чем ее влияние отразится на рынке. Второй подход допускает, что рынок эффективен и в момент публикации новости немедленно происходит изменение на рынке. Наконец, третий вариант – репортаж – основан на допущении, что новость публикуется постфактум, уже после ее влияния на рынок.

В связи с неопределенностью по поводу используемого временного интервала были исследованы следующие варианты: существование лага в 20 минут после публикации новости, отсутствие значимого лага, лаг в 5 минут до публикации новости. Для всех вариантов был проведен статистический анализ с целью выбора того временного интервала, при котором больше слов имели бы значимость для групп позитивных, негативных или нейтральных новостей. Наибольшее число значимых слов соответствовало подходу, допускающим лаг в плюс-минус одну минуту, что практически эквивалентно немедленному влиянию новости на рынок.

Для попытки спрогнозировать поведение рынка на основе новостей предложен анализ на основе выставления весовых коэффициентов словам в зависимости от их статистической значимости для групп новостей. Также использованы такие алгоритмы как градиентный бустинг, искусственная нейронная сеть, CBOW модель (на основе искусственной нейронной сети) и сверточная нейронная сеть, использующая предобученный массив вложений GloVe. Именно последняя модель дала наилучший результат.

В результате исследования не были найдены доказательства того, что возможно использование публикуемых новостей для прогнозирования котировки евро/доллар, гипотеза о средней эффективности рынка не может быть отклонена. Тем не менее, интересным результатом работы стало выявление ключевых слов, оказывающих значимое влияние на курс.

Сложности, которые возникли в процессе проведения исследования и могли повлиять на результаты:

- новостные данные ежедневно скачивались с портала Reuters.com один раз в сутки за несколько минут до полуночи. Во-первых, время новости, публикуемое на сайте, округлено в точностью до минуты, что могло немного исказить интервал влияния. Во-вторых, новость могла быть отредактирована в течение дня. В такой ситуации на сайте сохранялась только последний вариант новости и время последнего редактирования. Это может вызвать присваивание неправильной метки новости из-за неточности времени, а также повлиять на прогноз из-за изменения текста сообщения;

- хотя новости скачивались только из раздела Market news, их характер был очень разных: от макроэкономических и политических до новостей конкретный фирм. С одной стороны, это особенность выбранного актива: на евро и доллар США влияет огромное количества факторов и ситуация в разных странах. Но с другой стороны, часть информации могла оказаться бесполезной. В случае с фондовым рынков и анализом конкретных котировок разнообразие новостей намного меньше и лишние из них легче отфильтровать;

- дополнительные сложности вызвал выбор порога разбивки новостей на три группы. При выбранном пороге новости образовали примерно равные группы, однако, возможно, стоило уменьшить чувствительность и новости, оказавшие минимальное воздействие на рынок, отнести к нейтральным. Это могло отчасти помочь с решением предыдущей проблемы;

- новости также имеют различную значимость, а следовательно, одни могут оказать длительное воздействие на рынок и изменить весь тренд котировки, а другие могут вызвать лишь слабое отклонение от общего тренда. Очень показательный пример произошел на фоне разговор и референдума о Brexit. Задолго до голосования среди инвесторов сложились определенные негативные ожидания, которые были сильны настолько, что новости о результатах референдума практически не повлияли на состояние рынка. Тренд котировки давно изменился и новости о результатах вызвали незначительные колебания. Различная значимость новостей никак не учитывается в данном подходе.

Существует множество направлений для дальнейшей работы с целью улучшения модели. Например, в публикации Robert P. Schumaker и Hsinchun Chen [67] доказывается, что использование альтернативных представлений текста, таких как named entities, улучшают качества моделей по сравнению с мешком слов. Другой вариант улучшения модели – не ограничиваться текстовыми данными, но учитывать еще и временные ряды, как делается в рамках технического анализа. Как минимум, стоит попробовать учесть сам ряд

котировки евро/доллар США. Можно также создать фильтр новостей, чтобы отсеять шум, возможно, используя экспертное мнение.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. A stock prediction system based on news and Twitter / K.Kim et al. // International journal of software engineering and its applications. – 2016. – Col. 10, № 6. – P. 69 – 80.
2. Aase, K.-G. Text mining of news articles for stock price predictions: thesis of Master of Science: June 2011 / K.-G. Aase. – NTNU. – 84 p.
3. An algorithm for suffix stripping / M.F. Porter // Program: electronic library and information systems. – 1980. – Vol. 14, issue 3. – P. 130 – 137.
4. An investigation of Forex market efficiency based on detrended fluctuation analysis: A case study for Iran / Abounoori E, Shahrazi M, Rasekhi S. // Physica A. – 2012. – Vol. 391. – P. 3170 – 3179.
5. Antunes, P. A. L. Sentiment in financial news: master thesis in data analytics / P.A.L. Antunes. – 2015. – 99 p.
6. Automated news reading: stock price prediction based on financial news using context-capturing features / M. Hangenau, M. Liebmann, D. Neumann // Decision support systems. – 2013. - №55. – P. 685-697.
7. Beckmann, M. Stock price change prediction using news text mining: phd thesis / M. Beckman. – Rio de Janeiro, 2017. – P. 152.
8. Behavioral economics and the effects of psychology on the stock market: thesis / J.L. Nagy // Applied economic theses. – 2017. – Vol. 24.
9. Boboc, I.-A. An algorithm for testing the efficient market hypothesis / I.-A. Boboc, M.-C. Dinica // PLoS ONE [Electronic resource]. – 2013. – Mode of access: https://doi.org/10.1371/journal.pone.0078177. – Date of access: 30.01.2018
10. Bodie, Z. Investments / Z. Bodie, A. Kane, A.J. Marcus. – San Diego: University of California, 2014. – P. 1054.
11. Cootner, P.H. The random character of stock market prices / P.H. Cootner. – M.I.T. Press: 1964. – 510 p.
12. Currency exchange rate forecasting from news headlines / D. Peramunetilleke, R. Wong // Australian Computer Science Communications. - № 24. – P. 34-43.
13. Currency jumps, cojumps and the role of macro news / A. Chatrath [et al.] // Journal of International Money and Finance. – 2014. - № 40. – P. 42-62.
14. Ding, X. Deep learning for event-driven stock prediction / X.Ding [et al.] // Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence. – 2015. – P. 2327 – 2333.
15. Do Malaysian investors overreact / M.M. Lai, B.K. Guru, L.L. Chong // Journal of Behavioral Finance. – 2003. – Vol. 2, № 2. – P. 602 – 609.
16. Does the Fed beat the foreign-exchange market? / R.J. Sweeney // Journal of banking and finance. – 2000. - № 24. – P. 665 – 694.
17. Dollar US to Euro historical data [Electronic resource] / Free Forex Historical data. – Mode of access: http://www.histdata.com/download-free-forex-historical-data/?/metatrader/1-minute-bar-quotes/eurusd/2017. – Date of access: 02.11.2018.

18. Dreman, D.N. Psychology and the stock market: investment strategy beyond random walk / D.N. Dreman. – American management association, 1977. – 306 p.

19. Elder, A. Trading for a living: psychology, trading tactics, money management / A. Elder. – John Wiley & Sons, 1993. – 304 p.

20. Falinouss, P. Stock trend prediction using news articles: master thesis in marketing and e-commerce / P. Falinouss. – 2007. – 165 p.

21. Fama, Y.F. Efficient capital markets: a review of theory and empirical work / E.F. Fama // proceedings of the twenty-eighth annual meeting of the American Fincance Association, New York, December, 28 – 30, 1969. – P. 383 – 417.

22. Fang, J. Forex-foreteller: Currency trend modeling using news articles / Jin Fang [et al.] // the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2013. – P. 1470 – 1473.

23. Filter rules and stock market trading / E. Fama, M. Blume // Journal of Business. – 1966. – Vol. 39. – P. 226 – 241.

24. Fung, G.P.C. News sensitive stock trend prediction / G.P.C. Fung, J.X. Yu et al. // Pacific-Asia Conference on knowledge discovery and data mining (PAKDD), Taipei, Raiwan. – 2002.

25. Gholampour, V., Wincoop, E. What can we learn from euro-dollar tweets? / V. Gholmapour, E. van Wincoop. – NBERT working paper. – 2017. – 61 p.

26. Gidofalvi, G. Using news articles to predict stock price movements / G. Gidofalvi, C. Elkan // Department of Computer Science and Engineering, University of California. – 2003. – 67 p.

27. Goldberg, A.B. Seeing stars when theare aren't many stars: graph-based semi-supervised learning for sentiment categorization // A.B. Goldberg, X. Zhu // Proceedings of TextGraphs 2006. – 2006. – P. 45 – 52.

28. Graham, B. Security Analysis / B. Graham, D.L. Dodd. – The McGraw-Hill Co., 2009. – P. 819.

29. Hamilton, W.P. The stock market barometer / W.P. Hamilton. – LLC, 2014. – 370 p.

30. Haselton, M.G. The evolution of cognitive bias / M.G. Haselton, A. Nettle, P.W. Andrews // The Handbook of Evolutionary Psychology / Ed. D.M. Buss. – John Wiley & Sons, 2005. – P. 724 – 746.

31. Insiders' profits, costs of trading, and market efficiency / H.N. Seyhun // Journal of Financial Economics. – 1986. – Vol. 16. – P. 57 – 71.

32. Insider trading activity, different market regimens, and abnormal returns / E.J. Ferreira // Financial and Quantitative Analysis. – 1995. – Vol. 30. – P. 35 – 47.

33. Insider trading: a review of theory and empirical work / A. Doffou // Journal of Accounting and Finance Research. – 2003. – Vol. 11, № 1. – P. 1 – 18.

34. Is the US stock market becoming weakly efficient over time? Evidence from 80-year-long data / Alvarez-Ramirez J, Rodriguez E, Espinosa-Paredes G // Physica A. - 2012. Vol. 391. – P. 5643 – 5647.

35. Keynes, J.M. The general theory of employment, interest and money / J.M. Keynes. – Lefingtom: Stellar Classics, 2016. – P. 66.

36. Lavrenko, V. Mining of concurrent text and time series / V. Lavrenko [et al.] // Workshop on Text Mining. – 2000. – P. 37 – 44.

37. Lexicon-Based Methods for sentiment analysis / M. Taboada et al. // Computational linguistics. – 2011. – Vol. 37, № 2. – P. 267 – 307.

38. Mahajan, A. Mining financial news for major events and their impacts on the market / A. Mahajan, L. Dey, S.M. Haque // IEEE / WIC / ACM international conference on web intelligence and intelligent agent technology. – 2008. – Col. 1. – P. 423 – 426.

39. Mittermayer, M. Forecasting intraday srock price trends with text mining techniques / M. Mittermayer// Proceedings of the 37th Annual Hawaii international conference on system sciences. – 2004.

40. Multi-scale approximate entropy analysis of foreign exchange markets efficiency / G.-J. Wang, C. Xie, F. Han // Systems Engineering Procedia. – 2012. – Col. 3. – P. 201 – 208.

41. News portal Reuters.com [Electronic resource]. – 2018. – Mode of access: https://uk.reuters.com/. – Date of access: 01.11.2018.

42. Opinion mining and sentiment analysis / B. Pang, L.Lee // Foundations and trends in information retrieval. – 2008. – Vol. 2, № 1 – 2. – P. 1 – 135.

43. Pang, B. Seeing starts: exploiting class relationships for sentiment categorization with respect to rating scales / B.Pang, L.Lee // Proceedings of ACL, 2005. – P. 115 – 124.

44. Permunetilleke, D. Currency exchange rate forecasting from news headlines. - D. Permunetilleke, R.K. Wong // Proceedings of the 13th Australian database conference (ADC'02) – Melbourne, 2002. – 62 p.

45. Pesaran, M.H. Market efficiency today / M.H. Pesaran. – IERP working paper 05.41. – 2005. – 15 p.

46. Pratt, S.P. Relationships between insider training and rates of return for NYSE common stocks, 160 – 66 / S.P. Pratt, S.W. Vere // Modern developments in investment management / J. Lorie, R. Brealy. – 1978. – P. 259 – 270.

47. Price movements in speculative markets: trends of random walks / S.S. Alexander // Industrial Management Review. – 1961. – Vol. 2. – P. 7 – 26.

48. Proof that properly anticipated prices fluctuate randomly / A.P. Samuelson // Industrial Management Revies. – 1965. – Vol. 6, № 2. – P. 41 – 49.

49. Purchasing power parity and the real exchange rate / L. Sarno, M.P. Taylor // IMF Staff papers. – 2002. – Vol. 46, № 1. – P. 1 – 5.

50. Rachlin, G. ADMIRAL: a data mining based financial trading system / G.Rachlin, D. Alberg, A. Kandel // IEEE symposium on computational intelligence and data mining, 2007. – P. 720-725.

51. Random walks and technical theories: some additional evidence / M. Jensen, G.A. Bennington // Journal of Finance. – 1970. – Vol. 25, № 2. – P. 469 – 482.

52. Random walks: reality of myth / M.C. Jensen // Financial Analysts Journal. – 1967. – Vol. 10. – P. 1 – 9.

53. Rational herding in microloan markets / J. Zhang, P. Liu // Management science. – May, 2012. – Vol. 58, № 5. – 10 p.

54. Reconciling efficient markets with behavioral finance: the adaptive markets hypothesis / A.W. Lo // Journal of Investment Consulting. – 2005. – Vol. 7, № 2. – P. 21 – 44.

55. Relative strength as a criterion for investment selections / R.A. Levy // Journal of Finance. – 1967. – Vol. 22, T 4. – P. 595 – 610.

56. Rhea, R. The Dow theory / R. Rhea. – Fraser publishing Co., 1994. – 252 p.

57. Quantifying trading behavior in financial markets using Google Trends / T. Preis, H.S. Moat, H.E. Stanley // Scientific reports. – 2013. – Vol. 3. – P. 1684.

58. Quantifying Wikipedia usage patterns before stock market moves / T. Piet, H. Moat, H.Stanley // Scientific reports. – 2013. – Vol. 3. – P. 1801.

59. Shleifer, A. Inefficient markets: an introduction to behavioral finance / A. Shleifer. – Oxford University Press, 2000. – 224 p.

60. Some joint tests of the efficiency of markets for forward foreign exchange / J. Geweke, E.L. Feige // The Review of Economics and Statistics. – MIT press, 1979. – Vol. 61, № 3. – P. 334 – 341.

61. Special information and insider trading / J. Jaffe // Journal of Business. – 1947. – Vol. 47. – P. 132 – 139.

62. Stock-index invest model using news big data opinion mining / Y. Kim [et al.] // Journal of Intelligence and Information systems. – 2012. – Vol. 18, № 2. – P. 143-156.

63. Stock Market Prediction using Daily News Articles / Y.S. Patel, S. Mandal // [Electronic resource]. – Mode of access: http://www.iitp.ac.in/~arijit/dokuwiki/lib/exe/fetch.php?media=courses:2017:cs551:09_report.pdf. – Date of access: 08.01.2018.

64. Stock Market Reaction to Good and Bad Political News / S. Tahir // Asian Journal of Finance & Accounting. – 2012. – Vol. 4, № 10. – P. 229 – 312.

65. Stock return predictability and the adaptive market hypothesis: Evidence from century-long U.S. data / J.H. Kim, A. Shamsuddin, K.P. Lim // Journal of Empirical Finance. – 2011. – Vol. 18. – P. 868 – 879.

66. Text mining for market prediction: a systematic review / A. K. Nassirtoussi [et al.] // Expert Systems with Applications. – 2014. – Vol. 41.

67. Textual analysis of stock market prediction using breaking financial news: the AZFin text system / R.P. Schumaker, H.Chen // ACM transactions of information systems. – 2009. – № 27. – P. 1-19.

68. The foreign exchange market: theory and econometrics evidence / R.T. Bailie, P.C. McMahon // International journal of forecasting. – 1989. - № 3. – P. 385 - 387.

69. The impact of salient political and economic news on the trading activity / Y. Chan et al. // Pacific-basin finance journal. – 2001. - № 9(3). – P. 195 – 217.

70. The profitability of technical trading rules in the Asian stock markets / H. Bessembinder, K. Chan // Pacific-Basin Finance Journal. – 1995. – Vol. 3, № 2-3. – P. 257-284.

71. The use of knowledge in Society / F.A. Hayek // The American Economic Review. – 1945. – Vol. 35, № 4. – P. 519 – 530.

72. Thomas, J.D. Integrating genetic algorithms and text learning for financial prediction / J.D. Thomas, K. Sycara // Proceedings of the Genetic and evolutionary computing 2000 conference workshop on data mining and evolutionary algorithms. – 2000. – P. 101 – 162.

73. Thomsom, G.S. A taxonomy of selected finance theories // SSRN [Электронный ресурс]. – 2007. - Режим доступа: https://ssrn.com/abstract=1087583. – Дата доступа: 01.10.2017.

74. Towards an efficient stock market: empirical evidence from the Indian market / D. Majumder // Journal of policy modeling. – Elsevier, 2013. – Vol. 35 (4). – P. 572 – 587.

75. Twitter mood as a stock market predictor / J. Bollen, M. Huina //Computer. – 2011. - № 44. – P. 91 – 94.

76. Uncertainty about fundamentals and herding behavior in the Forex market / P.R. Kaltwasser // Physica A.: A Statistical Mexhanic and irs Applications. – 2010. – Vol. 389. – P. 1215 – 1222.

77. Vakeel, K. Impact of news articles on stock prices // K. Vakeel, S. Dey // The 6th IBM Collaborative Academia Research Exchange Conference. – 2014. – P. 35 – 37.

78. Vu, T.T. An experiment in integrating sentiment features for tech stock prediction in Twitter / T.T. Vu [et al.] // Proceedings of the workshop on information extraction and entity analytics on social media data. – Mumbai, 2012. – P. 22 – 38.

79. Why might share prices follow a random walk? / S. Dupernex // Student Economic Review. – 2007. – Col. 21. – P. 167 – 179.

80. Wuthrich, B. Daily stock market forecast from textual web data / Wuthrich B. et al. // Conference Systems, Man, and Cybernetics, 1998. – 1998. – Vol. 3. – P. 271 – 279.

81. Yahoo! for Amazon: sentiment extraction from small talk on the web / S.R. Das, M.Y. Chen // Management science. – 2007. - № 53. – P. 75-88.

82. Zhai, Y. Combining news and technical indicators in daily stock price trends prediction / Y. Zhen, A. Hsu, S.K. Halgamuge // Proceedings of the 4th international symposium on neural networks: advances in neural networks. – Nanjing, 2007. – Part 3. – P. 1087-1096.

83. Бернстайн, П.Л. Фундаментальные идеи финансового мира. Эволюция / П.Л. Бернстайн. – Москва: Альпина Паблишер, 2018. – 535 с.

84. Николенко, С. Глубокое обучение / С. Николенко, А. Кадурин, Е. Архангельская. – СПб: Питер, 2018. – 480 с.