МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение высшего образования
«Тольяттинский государственный университет»

Кафедра Прикладная математика и информатика
(наименование)

01.04.02 Прикладная математика и информатика
(код и наименование направления подготовки)

Математическое моделирование
(направленность (профиль))

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА (МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ)

на тему «Модели и алгоритмы прогнозирования фондового рынка на основе интеллектуального анализа текста»

Обучающийся А.Д. Шаброва
(Инициалы Фамилия) (личная подпись)

Научный руководитель д.т.н., доцент, С.В.Мкртычев
(ученая степень (при наличии), ученое звание (при наличии), Инициалы Фамилия)

Тольятти 2024

Оглавление

Введение

В настоящее время фондовые рынки существуют практически во всех развитых и большинстве развивающихся стран.

Фондовый рынок является динамическим центром для торговых ценных бумаг, включая акции, облигации и валюты. Это облегчает передачу активов, выдачу новых ценных бумаг и налогообложения. Рынок также играет решающую роль в создании справедливых цен.

Российский фондовый рынок, хотя и относительно молодой, представляет значительный потенциал роста. По мере того, как инвесторы все больше предпочитают ценные бумаги как инвестиционный автомобиль, понимание и прогнозирование цен на безопасность становится все более важным. Это особенно актуально, учитывая динамическую природу современных фондовых рынков. Моделирование цен на акции является центральным как для управления портфелем, так и для оценки финансовых инструментов [7].

Актуальность данной работы заключается в увеличении вероятности благоприятного прогнозирования фондового рынка с использованием анализа текста, поскольку использование машинного обучения и фундаментального анализа является менее эффективным.

Объектом настоящего исследования является прогнозирование фондового рынка на основе интеллектуального анализа текста.

Предметом исследования являются модели и алгоритмы прогнозирования фондового рынка на основе интеллектуального анализа текста.

Цель магистерской диссертации состоит в исследовании и разработке математических моделей и алгоритмов прогнозирования фондового рынка на основе интеллектуального анализа текста.

Для достижения цели необходимо решить основные задачи:

– проанализировать современное состояние проблемы

прогнозирования фондового рынка;

- проанализировать источники по теме исследования;
- выполнить прогнозирование производительности с использованием классических подходов;
- создать новые критерии для выбора функций, которые могут улучшить обобщаемость моделей результатов и быть настроены для различных уровней приемлемого риска.

Гипотеза исследования заключается в том, что применение предлагаемых в работе математических моделей и алгоритмов позволит повысить точность и качество прогнозирования фондового рынка в условиях кризиса.

Методы исследования. В процессе исследования использованы следующие подходы и методы: основы прогнозирования фондового рынка, интеллектуального анализа текста и автокодеров, структурные модели равновесия, методы машинного обучения.

Новизна исследования заключается в разработке новых моделей и алгоритмов прогнозирования фондового рынка.

Практическая значимость исследования заключается в создании кодировки для прогнозирования цен и возможности практического применения реализованной программы на основе предлагаемых моделей.

Теоретической основой диссертационного исследования являются научные труды российских и зарубежных ученых, занимающихся проблемами прогнозирования фондового рынка.

На защиту выносятся:

- модели и алгоритмы прогнозирования фондового рынка на основе интеллектуального анализа текста;
- результаты проверки адекватности предлагаемых моделей и алгоритмов.

По теме исследования опубликована 1 статья:

Шаброва А.Д. Математическое моделирование технических объектов //

Сборник материалов IX Международной научно-практической конференции (школы-семинара) молодых ученых. Тольятти, 2023. С. 198-201.

Магистерская диссертация состоит из введения, четырех глав, заключения и списка литературы.

Во введении обоснована актуальность темы исследования, представлены объект, предмет, цели и задачи, выносимые на защиту диссертации.

В первой главе представлено современное состояние проблемы прогнозирования фондового рынка.

Во второй главе проведен анализ существующих работ по теме исследования.

Третья глава посвящена прогнозированию производительности с использованием классических подходов.

В четвертой главе представлено моделирование математической модели и объединение текстовых идей с традиционными моделями.

В заключении приводятся результаты исследования.

Работа изложена на 73 страницах и включает 14 рисунков, 6 таблиц и 53 источника.

Глава 1 Современное состояние проблемы прогнозирования фондового рынка

«Прогнозирование фондового рынка – это попытка определить будущую стоимость акций компании или другого финансового инструмента, торгуемого на бирже. Успешное предсказание будущей цены акции может принести значительную прибыль. Гипотеза эффективного рынка предполагает, что цены акций отражают всю имеющуюся в настоящее время информацию, и любые изменения цен, которые не основаны на вновь выявленной информации, таким образом, по своей сути непредсказуемы» [4], [23].

К основным методам прогнозирования фондового рынка относятся:

- структурные модели равновесия;
- неструктурные модели временных рядов;
- неструктурные модели, построенные на основе фьючерсных цен;
- модели искусственного интеллекта;
- качественные модели прогнозирования;
- статистические методы, включающие в себя корреляционные и регрессионные методы;
- машинное обучение;
- анализ текста.

Прогнозирование фондового рынка, постоянное стремление к инвесторам, остается сложной и постоянно развивающейся проблемой. В то время как поле стало свидетелем достижения в области методологии и технологий, стремление к совершенному прогнозу остается неуловимым [40].

Вот подробный взгляд на текущее состояние проблемы.

Задача волатильности и сложности:

- непредсказуемость. Фондовый рынок по своей сути нестабилен, под влиянием множества факторов, включая экономические показатели, геополитические события, настроения инвесторов и корпоративные

новости. Эта непредсказуемость делает точное долгосрочное прогнозирование чрезвычайно трудным;

– нелинейность. Движения рынка часто демонстрируют нелинейные закономерности, что означает традиционные статистические модели, которые полагаются на линейные отношения, могут не запечатлеть полную картину. Это создает значительное препятствие для точного прогноза;
– информационная перегрузка. С огромным количеством доступных данных становится сложным выявлять соответствующие сигналы и отфильтровать шум, что еще больше усложняет усилия по прогнозированию.

Ограничения традиционных методов:

– технический анализ. Этот метод фокусируется на выявлении шаблонов и тенденций в исторических данных о ценах для прогнозирования будущих движений. Хотя он может предложить понимание, он часто страдает от предвзятости подтверждения и ограниченной прогнозирующей силы;
– фундаментальный анализ. Этот подход анализирует финансовые показатели компании и экономические данные, чтобы оценить внутреннюю стоимость акций. Несмотря на то, что он ценен для долгосрочных инвестиционных стратегий, может быть медленным реагировать на внезапные рыночные сдвиги;
– количественные модели. Эти модели полагаются на математические формулы и алгоритмы для анализа данных и генерации прогнозов. Хотя они могут быть сложными, они часто изо всех сил пытаются объяснить непредсказуемые события и могут быть склонны к переоснащению.

Новые подходы и инновации:

– машинное обучение. Алгоритмические методы, такие как глубокое обучение и нейронные сети, демонстрируют перспективу при

обработке больших наборов данных и определении сложных отношений. Тем не менее, они требуют обширных данных и могут быть подвержены предвзятости и переоснащению;

– обработка естественного языка (NLP). Эта технология позволяет анализировать текстовые данные, такие как новостные статьи, настроения в социальных сетях и корпоративные документы, предлагая представление о рыночных настроениях и потенциальных тенденциях;
– альтернативные источники данных. Источники данных, помимо традиционных финансовых показателей, набирают обороты, включая спутниковые образы, данные потребительских расходов и деятельность в социальных сетях. Это предлагает более целостный взгляд на динамику рынка.

Этические и практические проблемы:

– алгоритмическое смещение. Прогнозирующие модели могут непреднамеренно отражать и увековечивать существующие смещения в данных, на которые они обучены. Это может привести к несправедливым результатам для некоторых участников рынка;
– стаивание технологий. Хотя технология играет важную роль, слепая вера в алгоритмы без человеческого суждения может быть вредной;
– Манипуляция на рынке. Потенциал для неправильного использования прогнозных моделей для манипуляции с рынком остается проблемой.

Важность контекста и человеческого суждения:

– несмотря на достижения, прогнозирование остается неопределенным усилием. Крайне важно распознать ограничения любого прогнозирующего метода и рассмотреть разнообразные перспективы.

Будущее прогнозирования:

– гибридные подходы. Сочетание традиционных и современных методов, интеграция человеческих суждений с алгоритмическими

прогнозами и использование альтернативных источников данных, вероятно, станет все более распространенным;

– фокус на управлении рисками. Вместо того, чтобы искать абсолютную уверенность, фокус может сместиться в направлении управления рисками и генерирования вероятностных прогнозов, которые объясняют неопределенность;
– этические соображения. Разработка и применение инструментов прогнозирования должны будут решать этические проблемы, и обеспечить справедливость и прозрачность.

Стремление к точному прогнозированию фондового рынка является непрерывным путешествием без определенного решения. В то время как технологии предлагают мощные инструменты, важно подходить к прогнозированию с критическим умом, признавая ограничения и рассматривая этические последствия. Принимая многогранный подход, интегрируя человеческие суждения с алгоритмическим пониманием и оставаясь в курсе новых технологий, инвесторы могут ориентироваться в сложной и нестабильной ландшафте фондового рынка с большей осведомленностью и информированными решениями [36], [52].

В современном мире в условиях кризиса каждая крупная компания хочет не только оставаться на плаву, но и быть лидером продаж и услуг. В этом может помочь прогнозирование с помощью анализа текста, которое сделает прогнозирование более качественным и поможет компании допустить как можно меньше ошибок [33] .

Поэтому прогнозирование фондового рынка представляется весьма перспективным направлением развития, а особенно перспективным вариантом является прогнозирование с помощью анализа текста, которое мало используется в крупных компаниях.

Выводы по главе 1

В результате анализа современного состояния проблемы прогнозирования фондового рынка на основе интеллектуального анализа текста можно сделать вывод, что данная тема является актуальной и требует дальнейших исследований. Существующие модели и алгоритмы прогнозирования фондового рынка на основе интеллектуального анализа текста имеют определенные ограничения и недостатки, которые не позволяют достичь высокой точности прогнозирования. Принимая силу интеллектуального анализа, мы можем выйти за рамки ограничений традиционных подходов к прогнозированию и раскрыть весь потенциал текстовых данных. Этот инновационный подход имеет ключ к созданию более точных, проницательных и, в конечном счете, более прибыльных моделях фондового рынка, прокладывая путь к будущему, когда инвестиционные решения руководствуются более глубоким пониманием человеческих настроений и психологии рынка.

Глава 2 Анализ существующих работ по теме исследования

Для того чтобы продемонстрировать актуальность решаемой научной задачи в практическом плане, необходимо рассмотреть существующие исследовательские работы по схожей тематике с целью проведения их критического анализа и определения проблемной области, с которой предстоит работать в диссертации, а именно:

- недостаток данных;
- сложность алгоритмов;
- этические вопросы;
- непредсказуемость рынка.

В работах [28], [32], [47] подробно рассматриваются методы прогнозирования фондового рынка и акцентируется внимание на анализе с использованием текста, рассматриваются три варианта: использование только заголовка статьи, всего содержания статьи и краткого изложения статьи. Но автор отдает предпочтение использованию только заголовков статьи, что может привести к неправильному прогнозированию фондового рынка.

В статьях [22], [26], [43] применяется прогнозирование фондового рынка с использованием анализа текста, а также с добавлением вектора с моделью unigram. Автор приводит в таблице свои результаты, где показано, что шанс благоприятного прогнозирования увеличился на 1%, но также есть большой минус – краткосрочное прогнозирование. Несмотря на увеличение прогнозирования в краткосрочный период, с увеличением периода шанс благоприятного прогнозирования будет только уменьшаться.

Авторы в работах [14], [45] предложили совместить интеллектуальный анализ текста, регрессионный анализ и разработанный визуальный алгоритм. Прогнозирование не показало должного результата из-за того, что использовались только название и аннотация патентных документов.

В статьях [9], [17], [31] было рассмотрено долгосрочное прогнозирование с использованием анализа текста. В этих работах авторы для

автоматизации процесса использовали анализ текста с применением программного обеспечения. Благодаря ПО можно анализировать объект более систематически. Но также можно сказать, что долгосрочное прогнозирование не всегда бывает выгодно для фондового рынка, т.к. нельзя учесть все факторы.

В статьях [5], [10] предложили провести анализ продаж и спрогнозировать дальнейший объем продаж на кратковременный период, но с применением инструментов программной среды Statistica.

С использованием программы было увеличено прогнозирование, но авторы не смогли добиться нужного результата, так как данное отклонение находится в пределах доверительного коридора ARIMA, поэтому прогнозирование можно считать неудачным.

В статье [24] рассматриваются плюсы и минусы прогнозирования фондового рынка с использованием анализа текста.

Из важного можно выделить, что в лучшую сторону влияет напоминание и точность, но также есть факторы, которые не дают прогнозированию раскрыться полностью – это отрицания, ирония и сарказм.

В работах [3], [15], [37] рассказывается о методах анализа текста.

Можно выделить, что отправным моментом, который объединяет все методики анализа текста, является то, что в их основе лежат представления о единице анализа.

Понятие единиц анализа является крайне важным аспектом, поскольку выступает своего рода аналогом исследуемых.

В статьях [13], [44] производиться анализ тональности текста.

Целью анализа тональности является нахождение мнений в тексте и определение их свойств.

Главными атрибутами являются: автор, тема и тональность.

И можно сделать вывод, что при использовании анализа текста важно тестировать разные параметры, чтобы подобрать те, которые работают лучше на тестовых данных.

Также рассмотрим работы [2], [30], [51], которые рассказывают про тональность текста.

Здесь делается упор на анализ мнений, но, несмотря на многочисленные исследования в области прогнозирования цен с помощью анализа мнений, все еще присутствует множество ограничений, которые еще не изучены.

Выводы по главе 2

Исследований в области прогнозирования довольно много, но большая часть использует методы машинного обучения и фундаментального анализа.

После анализа источников, заметно, что метод анализа текста используют не многие и в основном для долгосрочного или краткосрочного периода, где не учитываются факторы, например, кризис.

Можно сделать вывод, что проблема исследования прогнозирования фондового рынка с помощью анализа текста актуальна.

Глава 3 Прогнозирование производительности с использованием классических подходов

3.1 Технический анализ

В области технического анализа существует множество устоявшихся методов. Существуют также три основные предпосылки, лежащие в основе технического анализа:

- рыночное действие обесценивает все. Предположение, лежащее в основе технического анализа, заключается в том, что все факторы, влияющие на цену: фундаментальные, политические, психологические – включены в данные о прошлых ценах. Это устраняет необходимость прямого знания фундаментальных принципов, поскольку, как полагают, исторические цены инкапсулируют их влияние;
- цены движутся в соответствии с тенденциями. В тексте утверждается, что успех коротких или длинных продаж зависит от предположения, что тенденции с большей вероятностью будут продолжаться, чем обратный. Это убеждение позволяет трейдерам предсказать будущее движение рынка и потенциально прибыль;
- технический анализ предполагает последовательное поведение человека, что позволяет аналитикам применять постоянные модели для выявления рынков. Эта зависимость от неизменной психологии лежит в основе непрерывности этих моделей.

Технический анализ развивался: «техник» и «чартист», первоначально представляющие ту же поле, сфокусированное на использовании диаграмм для прогнозирования рыночных тенденций. Тем не менее, в 1990-х годах произошел сдвиг, когда специалисты включали статистические методы наряду с традиционным анализом диаграммы [27], [48].

Чартисты полагаются на такие методы, как уровни поддержки и сопротивления, модели цен, скользящие средние и линии тренда для выявления тенденций. Их основной подход состоит в том, чтобы наблюдать исторические данные о ценах и определить закономерности, которые предполагают продолжение или изменение существующих тенденций. Эта зависимость на визуальном анализе, однако, остается субъективной, в то время как статистические методы обеспечивают более объективную перспективу [16]. Существует пять наиболее часто используемых графиков разворота (четыре из них изображены на рисунке 1) являются:

– голова и плечи;
– тройные верха и низа;
– двойные верха и низа;
– вершины и низы шипов;
– узор блюдец.

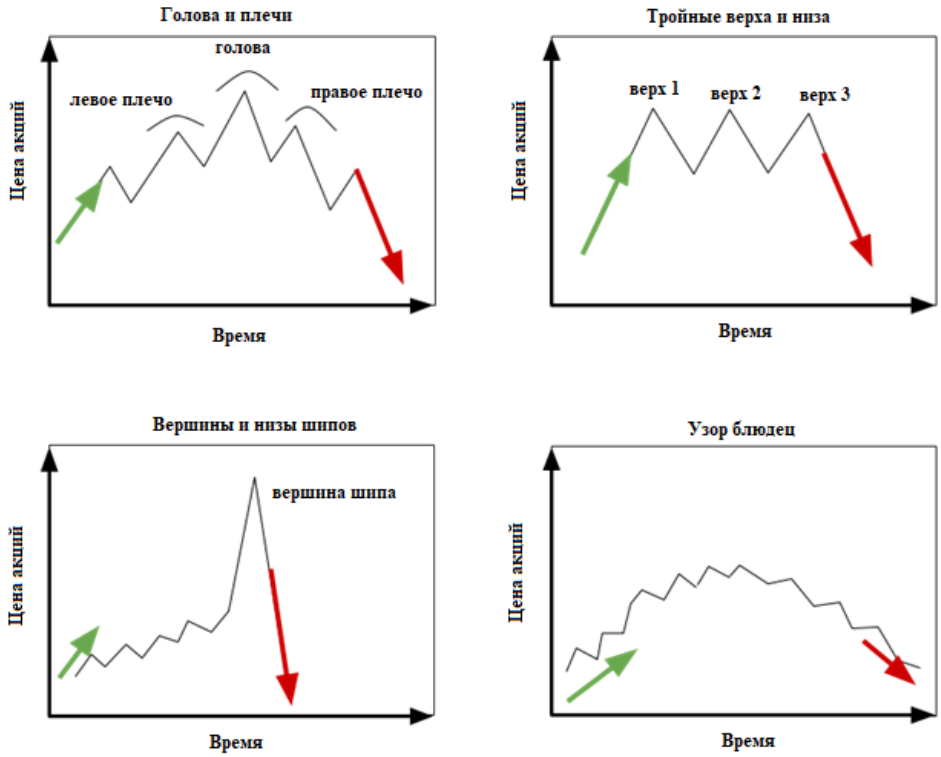

Рисунок 1 – Графики разворота

Есть несколько моментов, которые следует учитывать при попытке использовать шаблоны разворота, такие как требуемое наличие текущего тренда и то, что чем больше шаблон, тем больше последующее движение.

Например, когда применяется "голова и плечи", аналитик проверяет, соответствует ли недавняя кривая цен прототипу [20], [35]. Первая восходящая линия указывает на текущий тренд. Если кривая недавнего движения обладает характерными плечами и головой, то ожидается, что существующий тренд претерпевает разворот.

3.2 Фундаментальный анализ

Фундаментальный анализ углубляется в базовое здоровье и стоимость компании для прогнозирования будущих движений цен на акции.

В отличие от технического анализа, который фокусируется исключительно на исторических данных о ценах, фундаментальные аналитики изучают такие факторы, как финансовые показатели компании, управление, промышленность и конкурентная среда [18], [49]. Схема фундаментального анализа изображена на рисунке 2.

Чтобы сравнить компании разных размеров, финансовые коэффициенты играют решающую роль.

Эти соотношения, которые измеряют различные аспекты эффективности компании, позволяют сравнивать компании. Например, сравнение соотношения цены к прибыли (P/E) двух компаний в одном и том же секторе помогает оценить, какая компания дороже по сравнению с ее доходом.

Рисунок 2 – Схема фундаментального анализа

Абсолютные изменения цен являются менее значимыми, чем процентные изменения при анализе инвестиций в акции. Большое повышение цен в массовой компании может быть менее эффективным, чем меньшее увеличение меньшей компании, как с точки зрения процентного роста, так и вашей потенциальной прибыли [42]. Используя финансовые коэффициенты, фундаментальные аналитики могут лучше идентифицировать компании с сильными основными принципами и потенциалом для будущего повышения цен:

- «коэффициенты прибыльности, которые измеряют доходность фирмы;
- коэффициенты ликвидности, которые измеряют способность фирмы выплачивать свои непосредственные обязательства;
- коэффициенты задолженности, которые измеряют способность фирм выплачивать долговые обязательства с течением времени;
- коэффициенты использования активов, которые измеряют

способность фирмы эффективно использовать свои активы;

- коэффициенты рыночной стоимости представляют собой дополнительную группу коэффициентов, которые отражают рыночную стоимость акций и фирмы.

Некоторые коэффициенты рентабельности определяются следующим образом:

- маржа валовой прибыли = валовая прибыль / прибыль (чистый объем продаж);
- маржа операционной прибыли = операционная прибыль /прибыль;
- маржа прибыли до налогообложения = прибыли до налогообложения / прибыль;
- маржа чистой прибыли = чистый доход / прибыль» [49].

«Маржа – это положительная разница, которая осталась в результате какого-либо действия» [41].

Операционная прибыль является ключевым показателем для оценки прибыльности компании, отражая эффективность, с которой она управляет своими основными бизнес операциями. Это соотношение, полученное путем деления операционной прибыли на чистые продажи, подчеркивает влияние управленческих решений на контроль затрат. Его стабильность и надежность делают его ценным инструментом для сравнения компаний и прогнозирования будущей производительности.

Тем не менее, очень важно помнить, что только высокая маржа операционной прибыли не гарантирует успех инвестиций. Внешние факторы, такие как рыночные тенденции и эффективность отрасли, могут значительно повлиять на общую прибыльность компании. Кроме того, комплексная инвестиционная стратегия требует оценки не только текущих показателей, но и будущего потенциала роста и способности компании адаптироваться к изменению рыночных условий [38], [53].

3.3 Прогнозирование производительности с использованием машинного обучения

Теперь мы рассмотрим работы, в которых использовалось машинное обучение (рисунок 3) для прогнозирования производительности акций, сосредоточив внимание на конкретной методологии, которая использует рецидивирующие нейронные сети (RNNS) для прогнозирования относительных изменений цен. В отличие от подходов, основанных на текстовых данных, эта работа фокусируется на исторических данных о цене и прибыльности [12].

Исследование начинается с учета ограничений традиционных систем ранжирования, основанных на стоимости, особенно их плохих результатов в краткосрочном прогнозировании. Чтобы решить эту проблему, они использовали набор данных из 1452 акций из Valueline Universe, охватывая период с 1992 по 2001 год. Для регрессионной модели был выбран небольшой RNN с одним скрытым слоем [34].

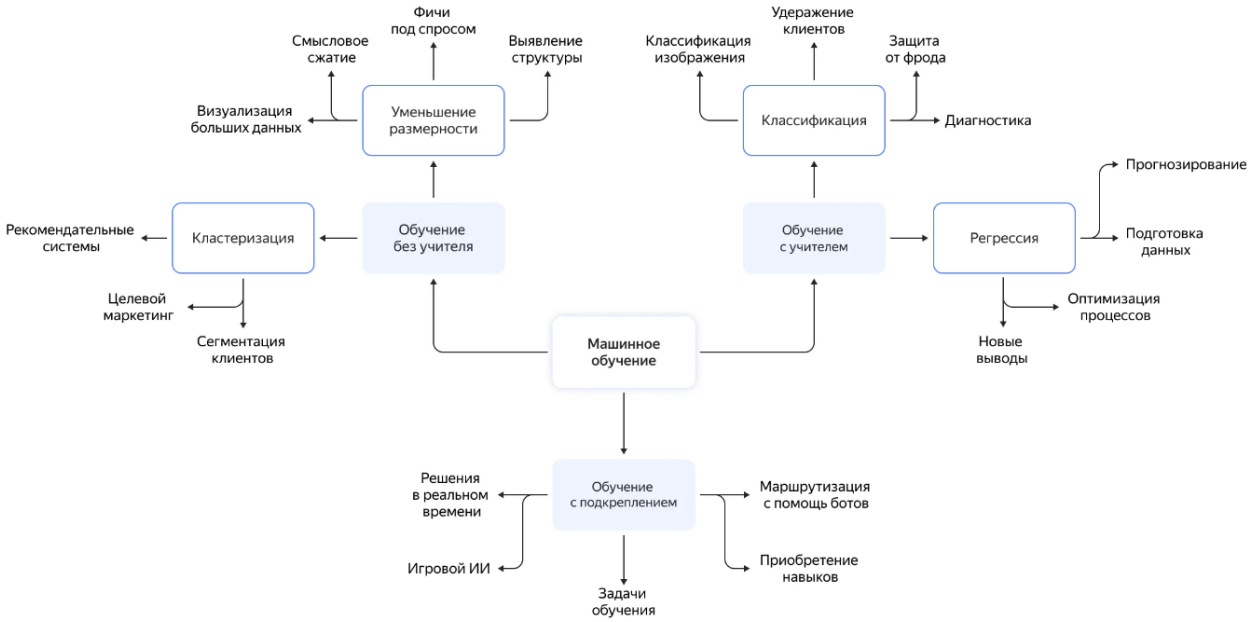

Рисунок 3 – Методы машинного обучения

В качестве вклада, модель получила предыдущие десяти квартальных процентных изменений, как в цене, так и для прибыли для каждой акции. Эти входные данные были преобразованы в относительные ранги и масштабировались вокруг нуля, обеспечивая последовательное представление данных. Выход нейронной сети был прогнозом относительного изменения цен в следующем квартале.

Чтобы усовершенствовать рейтинг дальше, прогнозируемые изменения цен были затем отсортированы в системе ранжирования. Этот метод, который объединил исторический относительный доход и запатентованный алгоритм, ранжировал все акции во вселенной (около 1700) от лучших (1) до худшего (5). Модель использовала множественные инициализации для нейронной сети и усредняет результаты, чтобы смягчить эффекты случайных начальных весов [21].

В отличие от традиционного рейтинга на основе стоимости, исследование вышло за рамки прогнозирования индивидуальных результатов акций. Он использовал прогнозируемые рейтинги для создания ряда портфелей, каждый из которых увеличивал свои ассигнования на высочайшие акции, одновременно продавая равное количество акций с самым низким рейтингом.

В течение 29-четвертейного периода с 1994 по 2001 год модель последовательно превосходила традиционную систему ранжирования на основе стоимости с точки зрения общей прибыльности. Дальнейший анализ с использованием коэффициента Sharpe, мера, которая учитывает риск, показал, что наиболее эффективный портфель достиг коэффициента Sharpe 0,55, превышая традиционный рейтинг на основе стоимости. Тем не менее, это было значительно ниже, чем типичные соотношения Sharpe от 1 до 2, наблюдаемых в хедж-фондах.

Это исследование подчеркивает потенциал методов машинного обучения, таких как RNNS в прогнозировании фондового рынка, особенно в повышении краткосрочной точности прогнозирования. Хотя предлагаемый

метод продемонстрировал превосходную производительность по сравнению с традиционными рейтингами на основе стоимости, необходимы дальнейшие исследования и уточнения для достижения уровней, сопоставимых с уровнями, наблюдаемыми в хедж-фондах.

Это исследование посвящено разработке прогнозирующей модели эффективности фондового рынка, в частности, нацеленном на выявление наиболее перспективных акций для инвестиций и коротких продаж. Вместо того, чтобы искать универсальную модель, которая предсказывает поведение всех акций, исследователи концентрируются на компаниях с высоким капитализацией, признавая, что модели, адаптированные к конкретным группам, могут привести к более точным прогнозам.

Модель использует уникальный набор характеристик для прогнозирования будущей производительности запасов. Они включают в себя возврат акций в течение предыдущей недели и двух недель, а также коэффициент объема, рассчитанные путем сравнения объема торгов в течение прошлой недели с объемом две недели ранее. Затем эти данные анализируются с использованием нового алгоритма, называемого «ранжированием прототипов», который использует модифицированную методологию конкурентного обучения для прогнозирования ранжирования акций [46].

Чтобы оценить эффективность модели, исследователи провели реальные торговые симуляции с использованием исторических данных обмена с 1963 по 2004 год. Они использовали стратегию выбора лучших акций с самым высоким прогнозируемым рейтингом для закупок и нижних «N» с акциями с самым низким рейтингом для коротких продаж в течение каждой моделируемой недели. Производительность модели была оценена на основе средней доходности портфеля, достигнутой в течение периода моделирования.

Результаты демонстрируют превосходство алгоритма ранжирования прототипа по сравнению с традиционными методами, такими как метод Coopers. Модель последовательно генерировала портфели со значительно

более высокой средней прибыльностью и поправкой на риск, измеренные по отношению к Шарпу. Коэффициент Шарпа, мера отдачи от поправки к риску, постоянно превышало 0,5 за период с 1978 по 1993 год и оставался выше 0,2 за период с 1994 по 2004 год, что указывает на надежные показатели в различных рыночных условиях.

Для дальнейшей проверки модели исследователи применили ее к набору данных трех крупных компаний с Иорданского фондового рынка, охватывая период с апреля 2005 года по май 2007 года. В этом тематическом исследовании предоставили дополнительные доказательства способности модели прогнозировать эффективность акций и определять потенциал инвестиционные возможности.

Это исследование предлагает ценный вклад в область финансового моделирования, предоставляя новый подход к прогнозированию эффективности фондового рынка. Способность модели генерировать высокую доходность с последовательной поправкой на риск, подчеркивает свой потенциал для руководства инвестиционными решениями. Кроме того, использование легкодоступных данных, таких как еженедельные доходные отношения и объемные отношения, делает его практическим инструментом, как для отдельных инвесторов, так и для финансовых учреждений, стремящихся оптимизировать свои стратегии управления портфелем.

«Входные данные для их модели состоят из пяти функций, связанных с ценой:

- цена закрытия предыдущего дня;
- цена открытия текущего дня;
- минимальная цена на текущий день;
- максимальная цена на текущий день;
- цена закрытия текущего дня.

В этом исследовании рассматривается уникальный подход к построению портфеля, используя классификационную модель для прогнозирования решений о покупке или продаже. Исследователи преобразовали непрерывные

значения атрибута в три класса: положительный, равный или отрицательный, сравнивая значения между последовательными днями. Хотя этот упрощенный формат классификации не идеален для ранжирования задач, его все еще можно использовать для генерации торговых сигналов» [24].

Для классификации они использовали алгоритмы дерева решений, такие как ID3 и C4.5, выбирая их из-за их удобства пользователя, интерпретации и вычислительной эффективности. Эти алгоритмы не требуют расширенного опыта машинного обучения, что делает их доступными для более широкой аудитории.

Производительность модели была оценена с использованием как 10-кратной перекрестной проверки, так и 66/33 тренировочных/тестов. Тем не менее, точность классификации варьировалась от 45% до 55%, что исследователи считают относительно низким.

Эта производительность значительно ниже по сравнению с другими подходами, такими как нейронная сеть, которая достигла точности в размере приблизительно 74% по ценам на акции от Гонконга и Шанхайских банковских корпораций в период с января 2004 года по декабрь 2005 года [11].

Исследование подчеркивает потенциал использования моделей классификации для построения портфеля, даже с упрощенным представлением данных.

Тем не менее, низкая точность классификации предполагает дальнейшие улучшения, потенциально путем изучения более сложных алгоритмов или включения дополнительных функций данных.

Это исследование также подчеркивает необходимость тщательной оценки различных методологий и их пригодности для конкретных финансовых задач.

3.4 Прогнозирование производительности с помощью анализа текста

В этом разделе мы, наконец, рассмотрим предыдущую работу, в которой машинное обучение сочеталось с анализом текста с целью прогнозирования производительности акций.

Исследование, проведенное в течение 5-недельного периода с 26 октября по 28 ноября 2005 года, сосредоточено на прогнозировании цен на акции S&P 500 через двадцать минут после выхода новых статей. Для этого был использован набор данных, состоящий из 9211 новостных статей и котировок акций. В качестве методов технического анализа было рассмотрено сочетание точных текстовых представлений и исторической информации о ценах [6], [29], [50].

«Анализ тональности относится к использованию обработки естественного языка, текстового анализа и вычислительной лингвистики для идентификации и извлечения субъективной информации в исходных материалах. Вообще говоря, анализ тональности направлен на определение отношения говорящего или писателя к какой-то теме или на определение общей контекстуальной направленности документа. Отношением может являться суждение или оценка, эмоциональное состояние или предполагаемый эмоциональный посыл. Основная задача при анализе тональности – классифицировать полярность данного текста в документе, предложении, любом другом уровне текста на позитивную, негативную или нейтральную. Продвинутая, «сверхполярная» классификация тональности выражается в таких состояниях как «сердитый», «грустный» и «счастливый»» [17].

«С ростом интернета за последние десять лет мнения стали доступны почти везде: на блогах, социальных сетях типа Facebook и Вконтакте, новостных порталах, сайтах электронной коммерции и других платформах. Хотя эти мнения должны быть информативными, количество и доступность информации становятся ошеломляющими для анализа. Последние годы

привели к увеличенному интересу к задачам суммирования мнений в области обработки естественного языка (NLP) и текстового анализа. "Мнения" часто представлены текстовой информацией из отзывов или блогов вместе с числовыми данными, такими как рейтинги. Несмотря на различия в толковании конечного результата между разными группами людей, целью является объективное обобщение нескольких точек зрения для получения всестороннего представления с определенной тональностью. Задачи предсказания тональности или самоклассификации активно изучались уже давно» [21].

«Новое поколение обзоров содержит структурированные сводки с хорошо организованным распределением по темам и временной визуализацией. Различные форматы подведения итогов дополняют друг друга, помогая аналитикам лучше понять информацию на новом уровне осознания данных. Например, прогнозирование настроений отзывов о продукте может указать пользователю его уровень удобства использования продукта» [36].

Разнообразные методы подведения обобщений имеют широкий спектр применений: от текстовой кластеризации до интеллектуальной обработки текста и анализа естественного языка. Некоторые из них базируются на простых правилах поведения ("эвристика"), тогда как другие используют более сложные статистические модели.

«Метод опорных векторов (SVM) эффективен для обработки многомерных массивов данных. Исследователи применяют SVM с линейным ядром в качестве модели прогнозирования. Для оценки анализа настроений на форумах было разработано шесть наборов функций. Первый использовал только исторические цены, в то время как другие методы интегрировали информацию о настроениях в модель прогнозирования» [45]. Для извлечения объектов из текстовых представлений использовался набор слов, словосочетаний существительных и именованных объектов. Для прогнозирования цен была использована производная SVM, специально разработанная для дискретного числового прогнозирования, а также модели,

содержащие различные переменные, зависящие от запасов. Простая модель использовала только регрессию, применяемую к историческим ценам на акции.

В другой модели учитываются как термины статьи, так и цены акций для прогнозирования. Эта модель показала наилучшие результаты при измерении близости к истинной будущей цене акций, точности при прогнозировании направления движения цены (57,1%) и доходности при моделировании торговли (2,06%).

Также было обнаружено, что схема имен собственных работает лучше, чем стандартный метод BoW.

Предполагается, что причина может заключаться в том, что BoW полагается на слишком шумные данные.

Для оценки показателей доходности модели был разработан торговый движок с правилами, направленными на максимизацию краткосрочной торговой прибыли.

В целом можно сделать вывод о более высокой эффективности подходов с использованием имен собственных по сравнению с набором слов [8], [19].

Чтобы получить метки для каждой акции (L(x)) в каждый день, значения ежедневной доходности (R(x)) преобразуются в положительный, нейтральный или отрицательный класс, как изображено на рисунке 4.

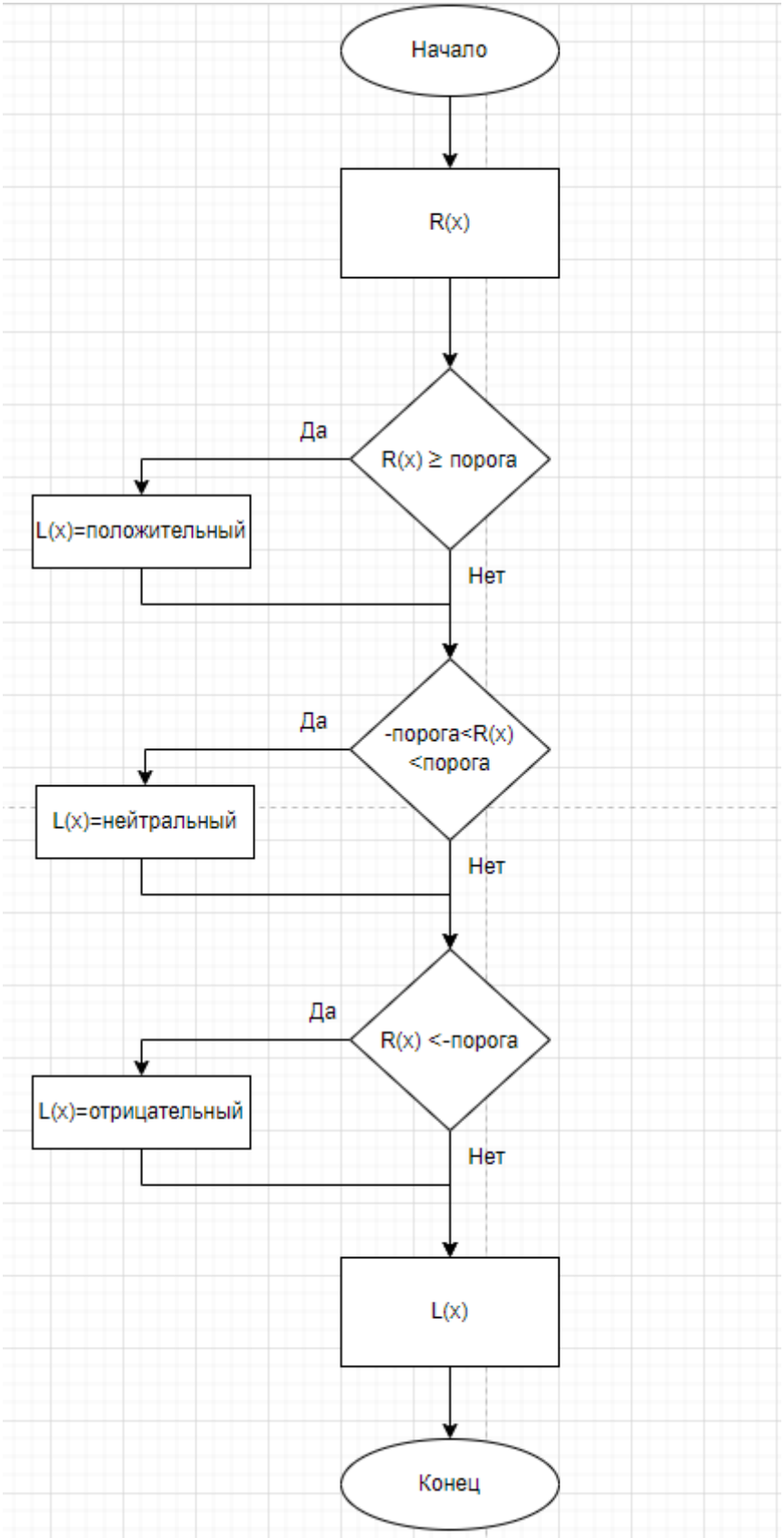

Рисунок 4 – Алгоритм получения метки для каждой акции

Производительность каждого тестируемого метода анализа настроений затем определяется точностью классификации машинной модели опорных векторов, обученной с использованием функций, предоставляемых методом анализа настроений. Они показывают, что на уровнях отдельных акций, секторов и индексов модели, использующие анализ настроений, а не набор слов, лучше работают как с данными проверки, так и с ранее невидимыми данными. Результаты показывают, что простая полярность настроений не может обеспечить прогностическую силу. Словарь настроений Гарварда IV содержит 182 измерения настроений, сгруппированных в 15 категорий для каждого из 10 000 слов, словарь настроений Лоуграна и Билла Макдональда содержит всего около 4000 слов и отображает слова в 6 измерениях настроений [1], [39]. Для дальнейшего исследования будем использовать прогнозирование фондового рынка в условиях кризиса с помощью анализа текста и автокодера, которые в совокупности должны дать благоприятный прогноз для компаний и организаций.

Выводы по главе 3

В данной главе были исследованы различные методы прогнозирования производительности фондового рынка. Проведен технический анализ, который основан на анализе исторических данных ценных бумаг с целью выявления трендов и паттернов. Фундаментальный анализ, в рамках которого изучены финансовые показатели компаний и экономические факторы для прогнозирования будущей производительности рынка.

Рассмотрено применение методов машинного обучения для прогнозирования производительности фондового рынка. а так же представлено использование анализа текста для прогнозирования производительности рынка. Исследовано, как текстовая информация из новостей, социальных медиа и других источников.

Глава 4 Моделирование математической модели и объединение текстовых идей с традиционными моделями

4.1 Математическая модель

«Математическая модель – это упрощенное представление реальной системы или явления с использованием математического языка и символов. Он включает в себя создание уравнений, формул или алгоритмов, которые можно использовать для описания, прогнозирования или анализа поведения системы. Математические модели широко используются в различных областях, таких как физика, биология, экономика, инженерия и социальные науки, поскольку они обеспечивают основу для понимания сложных систем и прогнозирования их будущего поведения» [25].

Математическая модель для анализа текста имеет большое значение, поскольку позволяет извлекать значимые идеи и закономерности из больших объемов текстовых данных. Применяя статистические и вычислительные методы, мы можем определить ключевые темы, настроения и даже предсказать будущие результаты на основе текстовых данных.

Эта модель имеет множество применений в различных областях, включая маркетинг, анализ социальных сетей, анализ отзывов клиентов и академические исследования. Например, в маркетинге компании могут использовать анализ текста, чтобы понять настроения потребителей и их мнения о продуктах и услугах. При анализе социальных сетей это может помочь определить популярные темы, мнения и разговоры. В академических исследованиях исследователи могут использовать текстовый анализ для анализа больших объемов литературы, выявления закономерностей и получения осмысленных выводов.

В целом, математическая модель для анализа текста помогает нам лучше понять значение больших объемов текстовых данных, что в конечном итоге может привести к более обоснованному принятию решений и лучшим

результатам в различных областях.

Для моделирования математической модели будем использовать анализ текста и автокодер для извлечения объектов из текста, который использует тип нейронной сети. Ранее было описано, как мы используем автокодер для извлечения функций для нашей диссертации.

Чтобы предсказать производительность акций, используя текстовые данные, мы используем автокодер для извлечения значимой информации. Это включает в себя представление каждого документа как вектор «пакетов слов», а затем сжатие этих векторов. Это сжатие, если сделано эффективно, кодирует суть каждого документа. Мы считаем, что конкретные параметры кодирования в этом процессе дают подсказку о будущем поведении запасов. Чтобы достичь этого, мы следим за пятиэтапным процессом:

Шаг 1. Сначала мы должны получить заявки 10-K и 10-Q для подгонки автокодера. Мы можем выбрать те файлы, к которым у нас есть доступ, которые не будут находиться внутри обучающих или тестовых данных прогностических моделей.

Шаг 2. Далее нам нужно предварительно обработать текст в этих файлах и преобразовать текст в нормализованные векторы "пакет слов". Более подробно, то, как мы это сделали, выглядит следующим образом:

а) для каждой заявки извлекается соответствующий текст. Выполнение этого состоит в разборе XML-файлов для получения основного текста документа;

б) извлеченный текст будет очищен. Это состоит в том, чтобы сделать весь текст строчным и удалить XML-термины, такие как & и &bull. Все символы, кроме цифр, букв и пробелов, также удаляются;

в) найдены наиболее распространенные 3000 слов среди заявок. Поскольку мы используем автокодеры, появление стоп-слов не должно иметь большого эффекта, поскольку частота их появления будет примерно одинаковой для каждого документа, что не повлияет на кодировку. Однако может также случиться так, что слова, обычно

классифицируемые как стоп-слова, действительно обладают прогностической силой в отношении показателей акций, и в этом случае выгодно сохранить эти слова;

г) векторы набора слов нормализованы:

1) масштабируем каждый вектор так, чтобы сумма была равна 1. Это устраняет эффект длины текста путем преобразования количества слов в плотность слов,
2) нормализация Z-балла выполняется таким образом, что каждый признак преобразуется таким образом, чтобы иметь нулевое среднее значение и единичную дисперсию. Это преобразование гарантирует, что если слово имеет очень высокую или низкую частоту в каждой записи набора данных, то это измерение в векторах не принимает значения только в небольшом диапазоне,
3) значения в каждом векторе обрезаются между -10 и 10, если в некоторых заявках есть слово с необычно высокой или низкой частотой, то после нормализации Z-балла соответствующее значение для этого слова будет очень высоким или низким (большое количество стандартных отклонений от средней частоты). Когда это значение используется в обратном распространении для обучения автокодера, это может привести к очень большим градиентам, которые негативно влияют на обучение. Эта проблема наиболее заметна при обучении рекуррентных нейронных сетей, и отсечение значений является одной из стратегий ее преодоления.

Шаг 3. «Далее нам нужно обучить автокодер кодировать и декодировать векторы записи таким образом, чтобы свести к минимуму потери при восстановлении. Автокодеры установлены таким образом, чтобы минимизировать потерю среднеквадратичной ошибки, как в формуле 1, где $\hat{y}$ и y – восстановленный и истинно нормализованный векторы соответственно. В этом тезисе также было рассмотрено использование второй потери при

подборе автокодеров. Вместо того чтобы настраивать автокодер только для кодирования и декодирования гистограмм, они также были обучены таким образом, чтобы кодировки можно было использовать для прогнозирования относительной доходности, связанной с каждой регистрацией. Для этих автокодеров с основными и вспомогательными потерями функция конечных потерь показана в формуле 2, где $\alpha \in [0, 1]$ определяет соотношение между основными и вспомогательными потерями. Мотивация включения этой второй потери заключалась в том, чтобы посмотреть, сможет ли автокодер извлечь из текста характеристики, связанные с производительностью акций» [50].

$$\text{потеря}_{\text{одиночная}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2, \quad (1)$$

$$\text{потеря}_{\text{двойная}} = \frac{1}{n}\left(\alpha\sum_{i=1}^{n}\left(\hat{y}_{main_i} - y_{main_i}\right)^2 + (1-\alpha)\sum_{i=1}^{n}\left(\hat{y}_{aux_i} - y_{aux_i}\right)^2\right), \quad (2)$$

где main – основная потеря, aux – вспомогательная потеря.

Шаг 4. Далее мы можем предварительно обработать данные, которые содержатся в данных обучения и тестирования, для наших прогностических моделей. Предварительная обработка может быть выполнена аналогично тому, как на шаге 2. Важно, чтобы в заявках, используемых для подгонки автокодера и для создания кодировки для прогнозных моделей запасов, не было совпадений, чтобы предотвратить чрезмерную подгонку.

Шаг 5. Наконец, эти новые векторы регистрации могут быть введены в настроенный автокодер, а кодировки сохранены в файл в хэш-таблице для последующего использования.

Прогнозирование фондового рынка является сложной задачей, так как

цены на акции и другие финансовые инструменты зависят от множества факторов, включая экономические, политические, социальные и технологические события. Традиционные методы прогнозирования, основанные на математических моделях и статистических данных, не всегда способны учесть все эти факторы и предсказать будущие изменения цен.

Однако с появлением большого объема текстовых данных, доступных в Интернете, стало возможным использовать анализ текста для прогнозирования фондового рынка. Анализ текста позволяет компьютерам извлекать информацию из текстовых источников, таких как новости, статьи, сообщения в социальных сетях и финансовые отчеты, и использовать эту информацию для предсказания будущих изменений цен на рынке.

В последние годы интеллектуальный анализ текста стал одной из самых активно развивающихся областей в искусственном интеллекте. Он позволяет компьютерам понимать и обрабатывать естественный язык, что открывает новые горизонты для анализа текстовых данных и извлечения полезной информации.

Это исследование представляет компьютерную модель, которая предсказывает движения фондового рынка, анализируя текстовые данные. Он сочетает в себе методы машинного обучения, анализа текста и финансового прогнозирования для выявления скрытых моделей, настроений и событий в рамках новостей, социальных сетей и финансовых отчетов. Признавая эти влияния, модель стремится предсказать будущие колебания цен на акции.

Основные компоненты такой модели включают:

- сбор данных. Модель собирает соответствующие текстовые данные из различных источников, таких как статьи о финансовых новостях, сообщения в социальных сетях, отчеты компаний и мнения аналитиков. Эти данные служат исходными данными для анализа;
- анализ настроений. Модель применяет методы анализа настроений для определения настроений или мнений, выраженных в собранном тексте. Он классифицирует настроения как положительные,

отрицательные или нейтральные, что дает представление о настроениях рынка и потенциальных движениях цен на акции;

- извлечение признаков. Одним из ключевых шагов в анализе текста является извлечение признаков из текстовых данных. Это позволяет преобразовать текстовую информацию в числовые признаки, которые могут быть использованы для обучения модели. Существует несколько методов извлечения признаков. Эти методы позволяют представить текстовые данные в виде векторов, которые могут быть использованы для обучения модели. Соответствующие признаки извлекаются из текстовых данных с использованием таких методов, как набор слов или встраивание слов. Эти функции собирают важную информацию, связанную с тенденциями и настроениями фондового рынка;
- моделирование и прогнозирование. После извлечения признаков можно приступить к построению модели. Можно использовать различные алгоритмы машинного обучения, такие как регрессия, классификация или рекуррентные нейронные сети, для обучения модели на исторических данных и прогнозирования будущих изменений цен на фондовом рынке;
- оценка модели. Эффективность модели оценивается с использованием статистических показателей, таких как точность и полнота. Оценка помогает определить эффективность модели в прогнозировании движений фондового рынка на основе анализа текста.

Сочетая передовые методы анализа текста с финансовым анализом, эта компьютерная модель призвана улучшить прогнозирование фондового рынка за счет использования текстовых данных и предоставления информации о настроениях инвесторов, рыночных тенденциях и потенциальных движениях цен.

Важно отметить, что прогнозирование фондового рынка является

сложной задачей, и результаты модели могут быть подвержены различным факторам, включая изменчивость рынка, неожиданные события и другие внешние влияния. Поэтому модель на основе анализа текста должна использоваться как один из инструментов принятия решений, а не как источник единственной истинной прогнозной информации.

4.2 Интеллектуальный анализ текста

«Хотя существуют различные методы для анализа текста, наиболее эффективный подход часто включает в себя целевой поиск конкретной информации, основанной на потребностях пользователей. В поиске информации обычно используются четыре ключевых метода:

- метод, основанный на терминах (Term Based Method (TBM));
- метод, основанный на фразах (Phrase Based Method (PBM));
- метод, основанный на концепциях или понятиях (Concept Based Method (CBM));
- метод шаблонной систематики (Pattern Taxonomy Method (PTM))» [52].

Рассмотрим подробно каждый метод.

«Term Based Method – это подход, основанный на терминах, включающий в себя тщательное изучение каждого термина и имеет преимущество эффективных вычислительных возможностей, а также из-за того, что он опирается на хорошо зарекомендовавшие себя теории для присвоения весовых значений терминам. Эти методы были разработаны экспертами по машинному обучению в области поиска информации за последние несколько десятилетий» [27]. К сожалению, основанные на терминах подходы подвержены проблемам синонимии и двусмысленности. Синонимия относится к нескольким словам, которые имеют одно и то же значение, в то время как полисемия обозначает слова с несколькими значениями. Семантические значения многих выученных терминов остаются

неясными, что затрудняет определение намерений пользователя.

Phrase Based Method – метод, основанный на фразах. Документ подвергается синтаксическому анализу на уровне фраз, поскольку фразы, как правило, менее двусмысленны и более уникальны, чем единичные термины. Следующие факторы были определены как основные факторы, влияющие на превосходную эффективность метода:

- фразы демонстрируют подчиненные статистические признаки по отношению к терминам;
- фразы встречаются в тексте нечасто;
- среди фраз есть многочисленные повторяющиеся и не относящиеся к делу фразы.

«Concept Based Method – метод концептуального анализа, включающий в себя разбивку терминов в соответствии с их релевантностью как на уровне предложения, так и на уровне документа. Хотя методы анализа текста основаны на статистическом анализе отдельных слов или фраз, важно отметить, что частота появления термина в документе не обязательно указывает на его важность для понимания документа. С введением нового подхода к интеллектуальному анализу, основанного на понятиях, акцент должен быть сделан на выявлении терминов, определяющих семантику или смысл текста» [36]. Для этого требуется модель с тремя компонентами: анализ семантической структуры предложений, создание концептуального онтологического графа (conceptual ontological graph (COG)) для описания этих структур и извлечение верхних понятий, которые можно использовать для построения векторов признаков. Используя понятийно-ориентированную модель, можно различать несущественные термины и те, которые вносят свой вклад в значение предложения. Этот подход обычно основан на технологиях обработки естественного языка и включает в себя выбор соответствующих понятий для оптимизации представления и устранения беспорядка или двусмысленности.

«Pattern Taxonomy Method – метод, основанный на анализе документов

проводимый путем ссылки на уже существующий шаблон или образец. Непоследовательность в распознавании шаблонов приводит к неоптимальному синтаксическому анализу текста, тем самым снижая производительность. Для решения этой проблемы метод на основе шаблонов включает два ключевых процесса: активацию шаблона и разработку шаблона» [46]. Благодаря этому подходу можно уточнить обнаруженные закономерности в текстовых документах. В экспериментальных испытаниях модель на основе шаблонов превзошла не только другие модели интеллектуального анализа данных и концепции, но и модель на основе терминов.

На таблице 1 представлен сравнительный анализ основных методов анализа текста и используемого в диссертации метода анализа текста и автокодера.

Таблица 1 – Сравнительный анализ методов

Метод	Преимущества	Недостатки
Метод, основанный на терминах	Эффективность, простота реализации, гибкость, объективность, универсальность	Чрезмерное упрощение, отсутствие контекста, неоднозначность, ограниченный охват, зависимость от качества данных
Метод, основанный на фразах	Эффективность, простота реализации, прозрачность, гибкость	Ограниченное понимание контекста, ограниченная точность, неспособность работать со сложными языковыми структурами
Метод, основанный на концепциях или понятиях	Точный анализ, универсальность, эффективность, согласованность	Отсутствие контекста, чрезмерное использование технологий, субъективность, ограниченный охват слов

Продолжение таблицы 1

Метод	Преимущества	Недостатки
Метод шаблонной систематики	Согласованность, эффективность, стандартизация, объективность	Ограниченный объем, отсутствие гибкости, неточность, стоимость, зависимость от качества данных
Метод анализа текста и автокодера	Эффективность, извлечение признаков, сжатие данных	Сложность, переобучение, ограниченная применимость

После сравнительного анализа опишем подробно преимущества и недостатки по каждому методу.

К преимуществам метода терминологического анализа текста относятся:

- эффективность. Метод на основе терминов быстр и эффективен при анализе больших объемов текста;
- простота реализации: этот метод прост в реализации и не требует передовых технических навыков или дорогостоящих программных инструментов;
- гибкость: метод, основанный на терминах, является гибким и может использоваться для анализа различных типов текстов, включая сообщения в социальных сетях, новостные статьи и научные статьи;
- объективность. Этот метод является объективным, поскольку он основан на статистическом анализе для определения частотности и шаблонов терминов;
- универсальность: метод на основе терминов можно комбинировать с другими методами анализа текста для получения более глубокой информации.

К недостаткам метода анализа текста на основе терминов можно отнести:

- чрезмерное упрощение: метод, основанный на терминах, может

чрезмерно упростить сложные тексты, сведя их к ряду терминов;

– отсутствие контекста: этот метод может не учитывать контекст, в котором используются термины, что приводит к неточному толкованию текста;
– неоднозначность. Некоторые термины могут иметь несколько значений, что может привести к путанице и неправильному толкованию;
– ограниченный охват: метод, основанный на терминах, может не улавливать нюансы и тонкости языка, такие как ирония, сарказм и метафора;
– зависимость от качества данных. Точность анализа зависит от качества используемых данных. Если данные необъективны или неполны, это может повлиять на результаты анализа.

К преимуществам метода, основанного на фразах можно отнести:

– гибкость. Метод можно настроить для работы с различными текстами, языками и областями;
– простота реализации: метод не требует сложных алгоритмов или больших наборов данных;
– прозрачность: метод прост для понимания и интерпретации, что делает его доступным как для экспертов, так и для неспециалистов;
– эффективность: метод быстрый и может быстро обрабатывать большие объемы текста.

Недостатки:

– ограниченное понимание контекста: метод не учитывает контекст, в котором используются слова. Это может привести к неточному или неполному анализу;
– ограниченная точность: метод может давать неточные результаты, если словарь фраз составлен неаккуратно;
– неспособность работать со сложными языковыми структурами: метод

может испытывать затруднения при идентификации и анализе сложных языковых структур, таких как идиомы или метафоры.

К преимуществам метода, основанного на концепциях или понятиях можно отнести:

– более точный анализ. Метод концептуального анализа текста позволяет проводить более точный и тщательный анализ текста, поскольку он фокусируется на конкретных понятиях, идеях или темах, а не на более широких категориях;
– универсальность: метод можно применять к разным типам текстов, включая письменные документы, речи и сообщения в социальных сетях, что делает его универсальным инструментом для исследователей;
– эффективность. Использование программного обеспечения и алгоритмов позволяет быстро и эффективно обрабатывать большие объемы текста, что делает его идеальным для анализа больших данных;
– согласованность: метод позволяет проводить последовательный и воспроизводимый анализ, снижая возможность систематической ошибки и ошибок.

К недостаткам можно отнести:

– отсутствие контекста: метод может упускать важные нюансы и контекстуальную информацию, поскольку он фокусируется на конкретных концепциях или темах и может не учитывать более широкий контекст, в котором был создан текст;
– чрезмерное использование технологий. Использование компьютерных программ и алгоритмов может привести к отсутствию критического мышления и упущению важной информации, которую технология не может получить;
– субъективность. Процесс выбора концепций и определения их

релевантности может быть субъективным, что может привести к систематической ошибке в анализе;

– ограниченный охват слов: метод может не охватывать всю сложность и богатство языка и может упускать важные аспекты текста, которые не имеют прямого отношения к выбранным понятиям или темам.

Преимуществами метода шаблонной систематики являются:

– согласованность. Анализ текста на основе систематики шаблонов обеспечивает согласованность процесса анализа, поскольку ко всем текстам применяется один и тот же набор правил и шаблонов;
– эффективность: это эффективный метод анализа, поскольку он экономит время, предоставляя предварительно определенный набор шаблонов и правил для анализа;
– стандартизация. Использование шаблонов помогает стандартизировать процесс анализа, снижая риск ошибок и повышая точность результатов;
– объективность. Этот метод сводит к минимуму влияние субъективной интерпретации, поскольку анализ основан на заранее определенных шаблонах и правилах.

Из недостатков можно выделить:

– ограниченный объем: шаблоны основаны на заранее определенных критериях, которые могут ограничивать объем анализа и не охватывать все аспекты текста;
– отсутствие гибкости: использование шаблонов может быть ограничивающим и может не допускать творчества или инноваций в процессе анализа;
– неточность: использование шаблонов может привести к ложным срабатываниям или отрицательным результатам, если шаблоны не разработаны должным образом;
– стоимость. Разработка и внедрение системы на основе шаблонов

может быть дорогостоящей и занимать много времени;

– зависимость от качества данных. Точность анализа зависит от качества используемых данных. Если данные необъективны или неполны, это может повлиять на результаты анализа.

К преимуществам метода анализа текста и автокодера можно отнести:

– эффективность: автокодеры эффективны при обработке больших объемов данных, что делает их подходящими для анализа текста;
– извлечение признаков. Автокодеры могут автоматически извлекать важные признаки из текстовых данных, которые можно использовать для дальнейшего анализа;
– сжатие данных: автокодеры могут сжимать данные в пространство меньшего размера, что может снизить вычислительные требования для анализа.

Недостатки:

– сложность. Автокодеры – это сложные модели, требующие большого количества обучающих данных и вычислительной мощности;
– переобучение. Иногда автокодеры могут переопределять данные, что означает, что они могут стать слишком специфичными для обучающих данных и могут плохо обобщать новые данные;
– ограниченная применимость: автокодеры подходят не для всех типов задач анализа текста, и их применимость может быть ограничена некоторыми специфическими задачами.

Можно заметить, что каждый из имеющихся методов имеет зависимость от качества текста, который будут анализировать, а значит, что созданный метод анализа текста с применением автокодера может дать более точное прогнозирование, чем имеющиеся четыре метода.

4.3 Алгоритм LSLR

В этом разделе мы представим и поэкспериментируем с методом выбора объектов. Сначала мы рассмотрим алгоритм выбора, затем мы будем использовать этот метод выбора объектов, чтобы изучить последствия ограничения типов используемых объектов (т.е. фундаментальных факторов, GICS или настроений) из набора данных. И потом, рассмотрим производительность функций, извлеченных из текста с помощью автокодеров.

«Поскольку существует большое количество возможных алгоритмов, которые можно попробовать, будем использовать только некоторые из наиболее популярных протестированных алгоритмов. Алгоритмы, подлежащие тестированию, следующие:

- линейная регрессия по методу наименьших квадратов (LSLR);
- стохастический градиентный спуск (SGD);
- повышение градиента (GB);
- нейронные сети с прямой связью (NN);
- adaboost с линейной регрессией (ALR)» [34].

Алгоритмы регрессии будут применены к функциям фундаментального фактора, функциям кодирования GICS и простым функциям настроения. Для каждого из алгоритмов выполняется быстрая оптимизация гиперпараметров, и для представления этого метода выбирается набор гиперпараметров с наилучшей производительностью проверки.

Таблица 2 – Базовые результаты

Алгоритм	Оценка валидации	Стандартное отклонение	Комбинированный балл валидации	Стандартное отклонение в показателях
LSLR	0,0231	0,0696	0,3314	0,0000
ALR	0,0212	0,0300	0,7063	0,0061

Продолжение таблицы 2

Алгоритм	Оценка валидации	Стандартное отклонение	Комбинированный балл валидации	Стандартное отклонение в показателях
SGD	0,0501	0,0713	0,7026	0,0024
GB	0,0011	0,0618	0,0178	0,0190
RF	0,0179	0,0388	0,4624	0,0388
NN	0,0252	0,0334	0,7531	0,0227

Как мы видим в таблице 2, ни одна из протестированных моделей не показывает наилучших результатов во всех областях, но алгоритм NN обладает наивысшей эффективностью проверки.

Алгоритм LSLR будет использоваться для испытаний из-за более низкой вариабельности, поэтому ее производительность здесь будет изучена немного более подробно.

Алгоритм LSLR, также известный как алгоритм линейной регрессии по методу наименьших квадратов, представляет собой статистический метод, используемый для поиска наиболее подходящей линейной зависимости между зависимой переменной и одной или несколькими независимыми переменными.

Он минимизирует сумму квадратов разностей между наблюдаемыми значениями и прогнозируемыми значениями из линейной модели.

Этот алгоритм обычно используется для задач прогнозирования и моделирования в различных областях, включая экономику, финансы и машинное обучение.

На рисунке 5 показаны баллы валидации для LSLR.

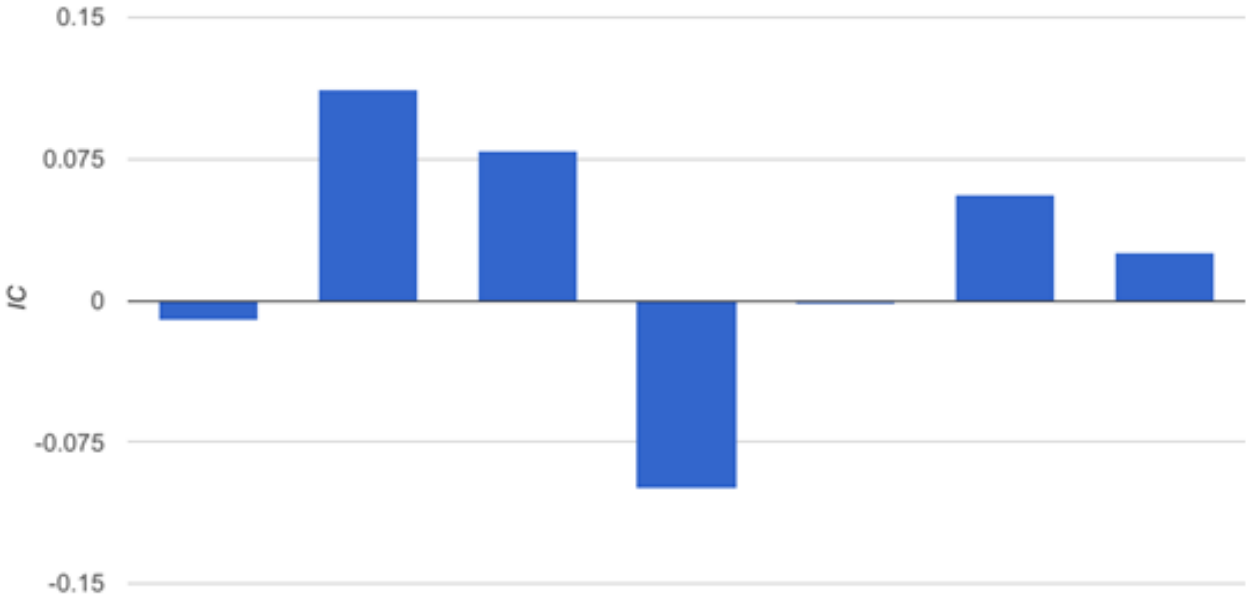

Рисунок 5 – Баллы валидации для LSLR

На рисунке 6 показана производительность LSLR за каждую неделю проверки, мы видим, что производительность обычно положительная, с плотными периодами плохой производительности. Этот тип дихотомической производительности ранее наблюдался в прогностических моделях фондового рынка.

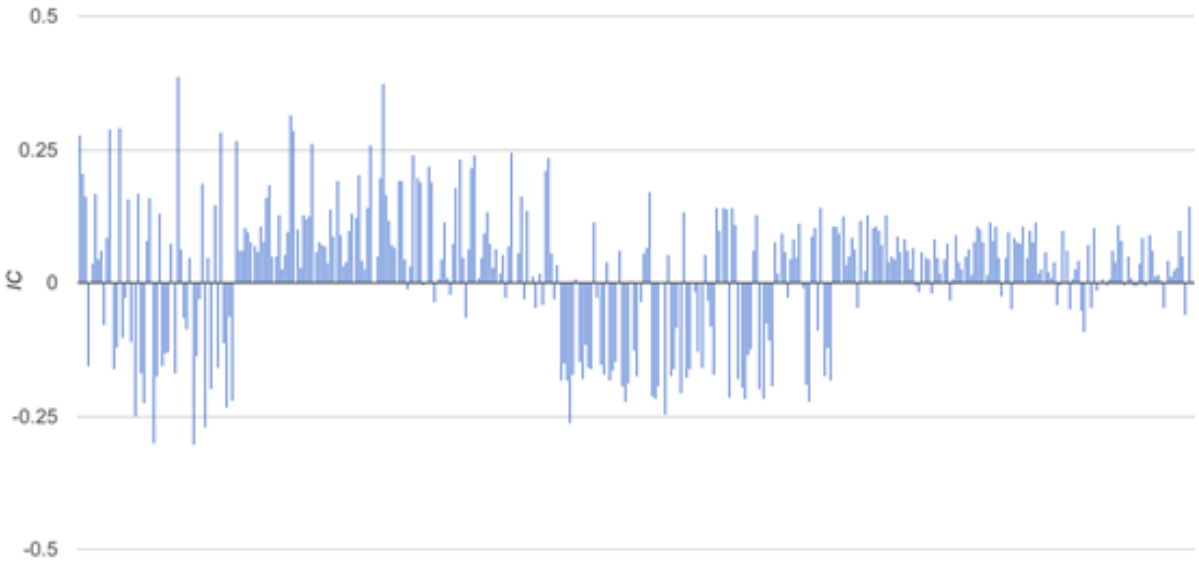

Рисунок 6 – IC для каждой недели проверки для LSLR

Здесь мы видим, что базовые модели не смогли достичь максимального значения IC 0,05.

Важно отметить, что линейная регрессия по методу наименьших квадратов работает лучше, если данные соответствуют предположениям модели, таким как линейность зависимости и нормальное распределение остатков. Если данные не соответствуют этим предположениям, может потребоваться использование других методов регрессии или преобразование данных для достижения линейности.

В этом пункте мы представим и поэкспериментируем с методом выбора объектов. Сначала мы рассмотрим алгоритм выбора, затем мы будем использовать этот метод выбора объектов, чтобы изучить последствия ограничения типов используемых объектов (т.е. фундаментальных факторов, GICS или настроений) из набора данных. Наконец, мы рассмотрим производительность функций, извлеченных из текста с помощью автокодеров.

При выполнении выбора функций необходимо выбрать алгоритм регрессии для оценки производительности набора функций. Поскольку этот процесс занимает очень много времени, полезно использовать алгоритмы с быстрым временем обучения и прогнозирования, а также с небольшими различиями в производительности в разных испытаниях, поэтому будет использоваться детерминированный алгоритм LSLR.

Метод выбора объектов, который мы выбрали для использования, представляет собой алгоритм прямого последовательного выбора в стиле поиска по лучу и состоит из итеративного расширения набора объектов, которые будут использоваться со следующим лучшим вариантом. Чтобы выполнить выбор функции, выполняется следующее до тех пор, пока не будет добавлено определенное количество функций или пока оценка не начнет уменьшаться.

Прогнозирование фондового рынка с использованием текстового анализа на Python.

Шаги кода:

– сбор данных;

– предварительная обработка данных;

– извлечение функций;

– анализ настроений;

– функциональная инженерия;

– обучение модели;

– оценка модели;

– развертывание модели.

Код:

Импорт NLTK

импортировали PD

от TextBlob Import TextBlob

от sklearn.model_selection import train_test_split

от Sklearn.linear_model Import Linearregression

Листинг 1 – Сбор данных

def collect_news_data (тикер):

Храним статьи в списке или dataframe

Листинг 2 – Предварительная обработка данных

def preprocess_text (text):

Удалили Стоп Слова

stop_words = nltk.corpus.stopwords.words ('английский')

tokens = nltk.word_tokenize (текст)

Filtered_tokens = [токен для токенов в токенах, если токен не в stop_words]

```
# Stemming или Lemmatization
stemmer = nltk.porterstemmer ()
Lemmatizer = nltk.wordnetlemmatizer ()
Processed_text = " .join ([stemmer.stem (token) для токена в фильтровании_tokens])
вернуть обработанную_xtext
```

Листинг 3 – Извлечение функции

```
# Извлекли результаты настроений из новостных статей
def get_sentiment (текст):
Анализ = TextBlob (текст)
 return Analysis.sentiment.polarity
news_data ['sentiment'] = news_data ['upported_text']. Применить (get_sentiment)
# Извлекли другие соответствующие функции из новостных статей
# проанализировали ключевые слова, распознавание сущности и т. Д.
```

Листинг 4 – Анализ настроений

```
def analyze_sentiment (текст):
Анализ = TextBlob (текст)
настроение = анализ
вернуть мнение
```

Листинг 5 – Инженерия функций

```
def extract_features (текст):
# Извлекли ключевые слова с использованием TF-IDF или других методов из организации (например, компании)
# В сочетании с оценкой настроений
```

функции = {

«чувства»: Analyze_sentiment (Text),

}

Листинг 6 – Обучение модели

Обучили модели

```
# Объединили данные фондового рынка и новостей на основе даты
merged_data = pd.merge (data, news_data, on = 'date')
# Выбрали функции для модели
функции = [«настроение», «open», «High», «low», 'volume']
target = 'close'
# Разделили данные на наборы обучения и тестирования
X_train, x_test, y_train, y_test = train_test_split (merged_data [features], merged_data [target], test_size = 0,2)
# Инициализировали модель линейной регрессии
модель = linearRegression ()
# Обучили модель на учебные данные
model.fit (x_train, y_train)
```

Листинг 7 – Оценка модели

```
# Сделали прогнозы по данным тестирования
y_pred = model.predict (x_test)
# Рассчитали среднюю квадратную ошибку (MSE)
mse = mean_squared_error (y_test, y_pred)
Печать (квадратная ошибка f'mean: {mse} ')
```

Листинг 8 – Развертывание модели

```
# Сохранили обученную модель для будущего использования
filename = 'stock_prediction_model.sav'
Pickle.dump (Model, Open (имя файла, 'wb')))
# Загрузили сохраненную модель и сделали прогнозы на новые данные
loaded_model = pickle.load (open (filename, 'rb'))
new_data = pd.dataframe ({'sentiment': [0.5], 'open': [100], 'High': [105], 'low': [95], 'том': [10000]})
Прогнозирование = loaded_model.predict (new_data)
Print (F'prediced Closing Price: {прогноз [0]} ')
```

Объяснение:

- сбор данных: собрали новостные статьи, финансовые отчеты и посты в социальных сетях, связанные с конкретными компаниями или более широким рынком;
- предварительная обработка данных: Очистили и подготовили текстовые данные для анализа, включая удаление стоп-слов, стемминга и лемматизации;
- извлечение функций: Извлекли оценки настроений из новостных статей с использованием TextBlob;
- анализ настроений: Применили методы анализа настроений (например, Vader, TextBlob), чтобы классифицировать общее настроение, выраженное в тексте;
- функциональная инженерия: Извлекли соответствующие функции из текста, такие как ключевые слова, сущности и оценки настроений;
- обучение модели: Объединили данные фондового рынка и данные новостей на основе даты;
- оценка модели: Сделали прогнозы на тестовых данных;

– развертывание модели: Сохранили обученную модель для будущего использования с помощью Pickle.

На практике выполнение каждой итерации может занять довольно много времени. Для результатов в этом тезисе была добавлена оптимизация, при которой после каждой итерации функции с наихудшим результатом (нижние 20%) в наборе F удалялись.

Использование Ss с этим алгоритмом выбора функций, для оптимизации работы функций и создания стабильного набора функций, которые демонстрируют высокую надежность в течение продолжительного времени, необходимо провести оценку оптимального значения s в метрике Ss. Это значение влияет на баланс между согласованностью и производительностью.

Для калибровки было использовано всего 5 функций, и проверенные значения s составляли 0, 0,05, 0,1 и 0,2.

Учитывая значительную изменчивость производительности модели от года к году, оптимальное значение s для одного года может значительно отличаться от оптимального значения для другого года. Поэтому точная оценка наилучшего значения s на этапе калибровки затруднена.

Предполагаемая производительность теста нелинейно меняется при увеличении s, что указывает на потенциальную пользу более тщательной оптимизации s в будущих экспериментах.

Из рисунка 7 видно, что оптимальное значение s вероятно находится в диапазоне от 0,1 до 0,2. Для последующих экспериментов будет использовано значение 0,15.

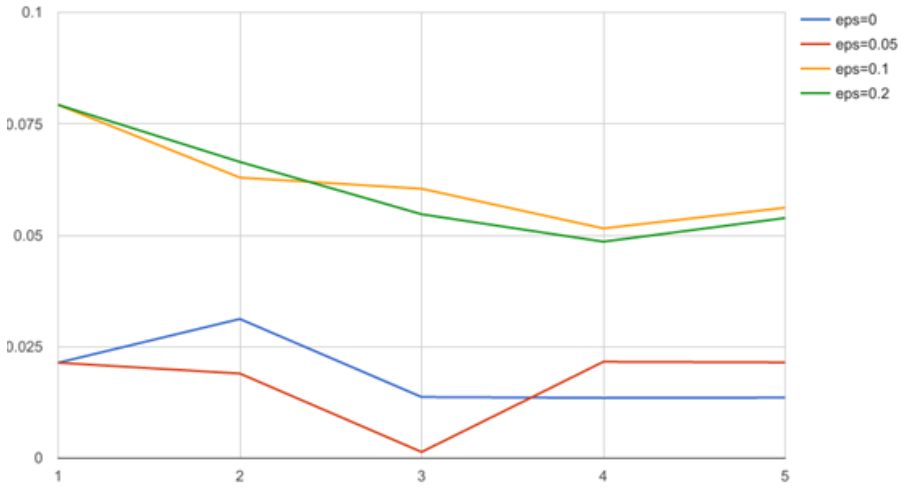

Рисунок 7 – Калибровка эпсилона

Мы обновили основные индикаторы в таблице 3 после определения значения S. Это показывает, что модель SGD лучше всего работает с этой новой метрикой. Мы повторно оптимизировали гиперпараметры, чтобы максимизировать новую метрику.

Таблица 3 – Базовые результаты с $S_{0,15}$

Базовая модель	Комбинированный балл валидации
LSLR	0,1050507
ALR	0,1441878
SGD	0,2327753
GB	0,0052034
RF	0,0949984
NN	0,1578199

Здесь мы видим результаты выбора функций при использовании всех исходных функций (фундаментальные факторы, кодировки GICS и простые функции настроения) в процессе выбора. Мы видим, что максимальный балл проверки 0,528 достигается после выбора 16 функций. В нашем наборе данных метод выбора объектов обеспечивает явное улучшение по сравнению с базовыми показателями, которые представлены в таблице 4 (0,528 против

0,232). Используя выбор функций, мы смогли достичь 0,232 только после одной итерации. Также интересно отметить, что IC проверки достигает максимума при 14 функциях, а затем немного снижается, в то время как оценка $S_{0,15}$ увеличивается из-за уменьшения различий в производительности между годами.

Таблица 4 – Производительность модели по мере добавления функций

Добавленная функция	Оценка валидации	Стандартное отклонение	Комбинированный балл проверки
Factor 137	0,05300124254	0,0347991508	0,2868045784
Factor 109	0,064387089	0,02586942825	0,3661073425
Factor 21	0,07069910576	0,02258910364	0,4096382927
Factor 99	0,07432728367	0,02408175369	0,4269676867
GICS_4_70	0,07687572059	0,02475900619	0,4398956155
GICS_4_25	0,07890748818	0,0240066607	0,4534739525
Factor 89	0,08108455964	0,02204046373	0,4713109805
Factor 100	0,08326501888	0,01720894412	0,4979698863
Factor 4	0,08411040933	0,01565303091	0,5077505004
GICS_3_20	0,08505637833	0,01631039674	0,511431516
Factor 13	0,08701098891	0,01885927511	0,5152869977
Factor 25	0,08793348136	0,01972159868	0,5181042486
Factor 65	0,08820532234	0,01940277967	0,520684032
GICS_4_80	0,08845472234	0,01897161449	0,5234886499
SCORE 10K	0,08740185402	0,01621277461	0,5258431804
Factor 15	0,08714494319	0,0149387765	0,5283472149
GICS_4_35	0,08714494319	0,0149387765	0,5283472149
GICS_4_45	0,08714494319	0,0149387765	0,5283472149
Factor 124	0,08682311881	0,01448618171	0,5278444542
Factor 2	0,08539276856	0,01263666027	0,5250523985

На рисунке 8 синяя кривая – оценка валидации, красная – комбинированный балл проверки.

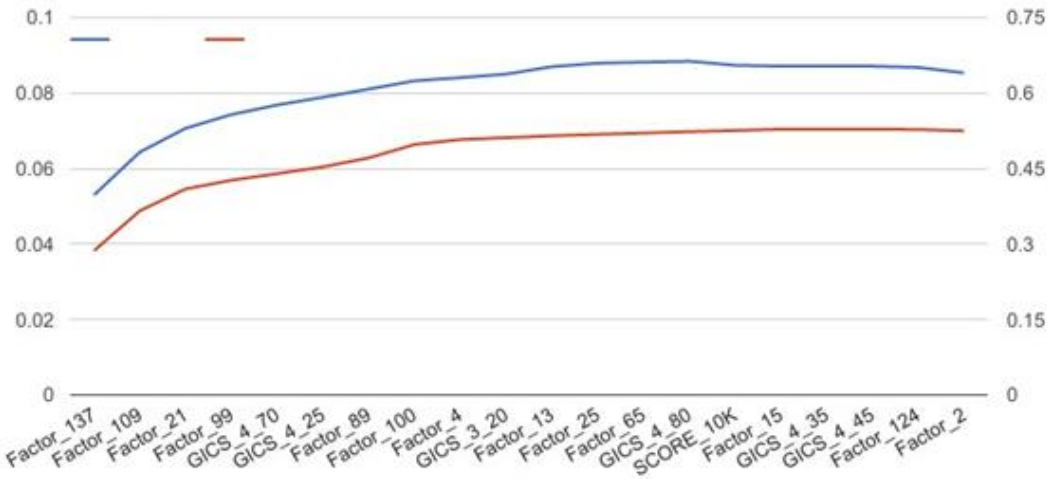

Рисунок 8 – Проверка IC и $S_{0,15}$ по мере добавления функций

Интересный вопрос, который следует задать об этом методе выбора функций: ”Насколько выгодно итеративное свойство?”. То есть, что, если функции были отсортированы по их проверке на 0,15 балла при индивидуальном использовании, и были выбраны лучшие из этого списка? На рисунке 9 мы можем видеть именно это. Добавление функций в порядке уменьшения $S_{0,15}$ балла при самостоятельном измерении. Левая вертикальная ось - это IC проверки, а правая вертикальная ось - оценка проверки $S_{0,15}$.

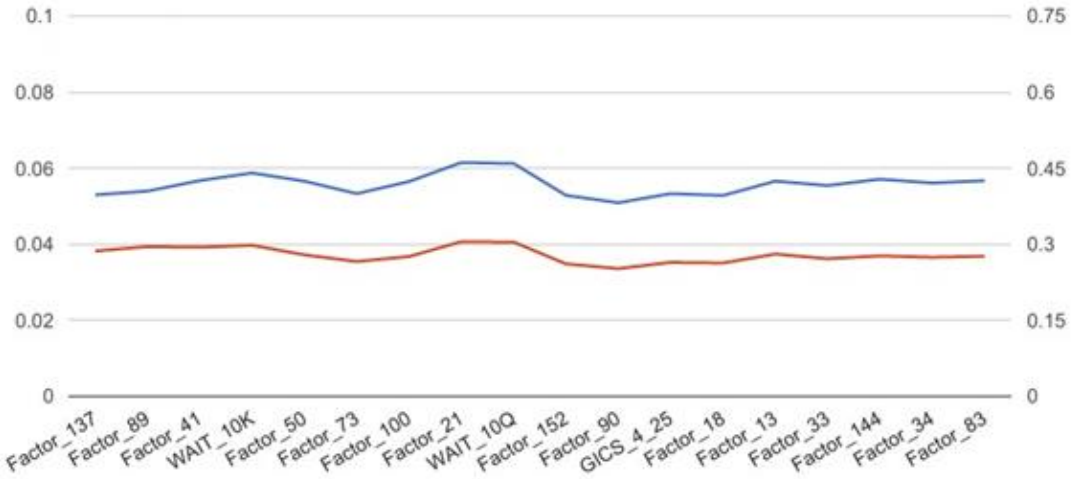

Рисунок 9 – Проверка IC и $S_{0,15}$ балла по мере добавления функций.

Хотя производительность хорошо начинается, она остается относительно постоянной по мере добавления новых лучших функций. На рисунке 10 мы можем видеть, что произойдет, если вместо сортировки по $S_{0,15}$ мы будем сортировать по IC проверки. Производительность лишь немного ниже, а набор функций вверху имеет большое пересечение. Очевидно, что итеративный аспект выбора функций важен, поскольку он гарантирует, что выбранные функции хорошо работают вместе.

Проверка IC и $S_{0,15}$ балла по мере добавления функций. Значения представляют собой среднее значение двух запусков с наборами функций, сгенерированных сетью с одинаковыми гиперпараметрами. Левая вертикальная ось - это IC проверки, а правая вертикальная ось - оценка проверки 0,15 балла.

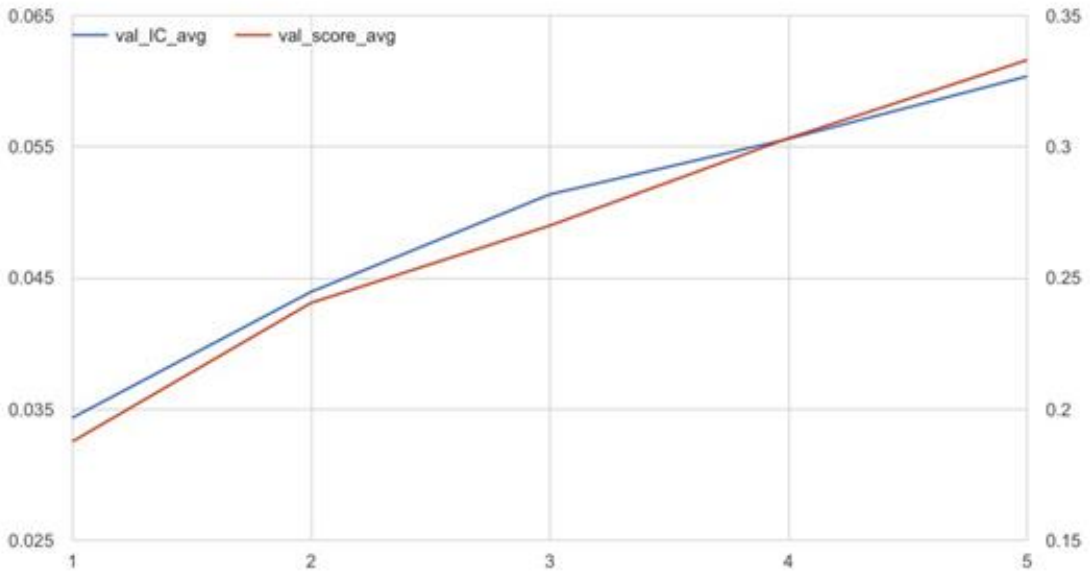

Рисунок 10 – Проверка IC и $S_{0,15}$ балла по мере добавления функций

Сначала мы рассмотрим настройку используемых автокодеров, а затем рассмотрим результаты использования новых текстовых функций.

Оптимизация гиперпараметров нейронной сети может занять очень много времени, так как необходимо учитывать множество факторов.

Некоторые важные из них включают продолжительность обучения, размер партии, оптимизатор веса, структуру графика и размеры слоев, метод инициализации веса, тип и вес регуляризации, а также функции активации.

Для наших экспериментов мы ограничим наши автокодеры простыми полностью связанными графиками прямой связи. Чтобы определить точную структуру, регуляризацию и вспомогательный вес потери, будут проведены несколько экспериментов.

Для нашей работы мы планируем провести три эксперимента для выбора размеров слоя AE, регуляризации и вспомогательного веса потерь. Первый эксперимент будет проводиться с небольшой сетью без регуляризации и вспомогательным весом потерь 0. Второй эксперимент предполагает использование более широкой и глубокой сети с умеренной регуляризацией и вспомогательным весом потерь 0. В третьем эксперименте мы будем использовать лучшие аспекты из первых двух экспериментов, но с добавлением ненулевого вспомогательного веса потерь. Учитывая стохастический характер обучения AE, каждый эксперимент будет повторен дважды для создания двух наборов объектов, и выбор объектов будет осуществляться для обоих наборов, после чего будут показаны их средние результаты, до 5 объектов в каждом. Набор функций, используемый для выбора, будет состоять только из тех, которые были сгенерированы AE.

В первом эксперименте будет использован один скрытый слой размером 25. Это позволит получить 25 функций из 10K заявок и 25 функций из 10Q заявок. На рисунке 11 мы видим, что оценка валидации достигает 0,24 после 5 функций, что немного ниже производительности функции Factor_137. IC 0,048 примерно такой же, как и для Factor_137. Производительность также, по-видимому, монотонно улучшается по мере выбора функций.

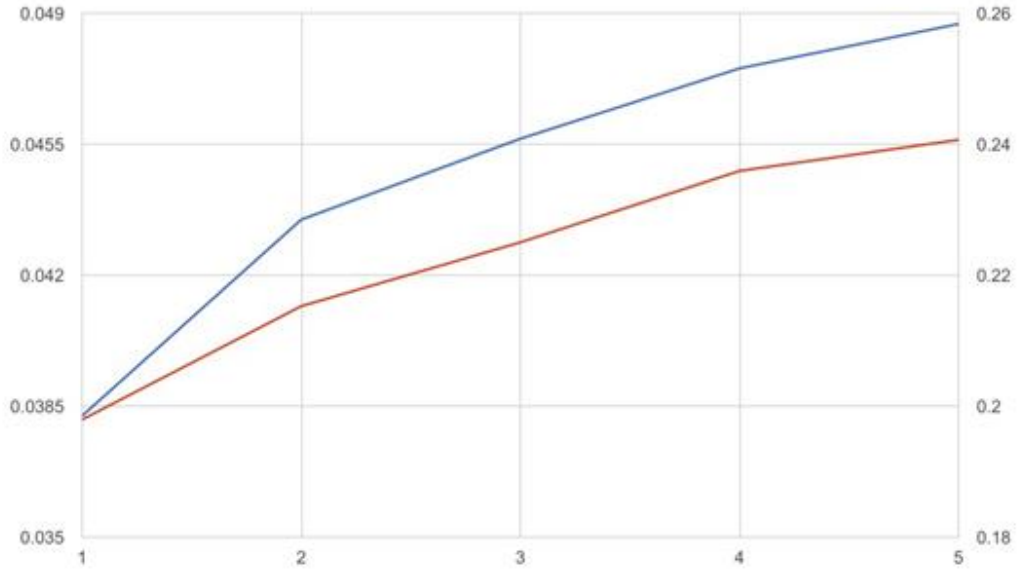

Рисунок 11 – Проверка IC и $S_{0,15}$ как функции, созданные небольшим AE

«Далее мы попробуем более глубокий автокодер. Размеры скрытых слоев составляют 5000-500-50-25-50-500-5000, с отсевом 0,4 для первых двух и двух последних слоев, и 0,2 для слоев непосредственно перед и после слоя кодирования. Как мы видим на рисунке 12, производительность после первой функции является наилучшей. В более ранних экспериментах с нейронной сетью мы обнаружили, что они очень легко подстраиваются под данные. Возможно, именно поэтому меньшая AE без регуляризации работает лучше, чем глубокая.

Поскольку меньшая AE обеспечивает характеристики, по крайней мере, такие же хорошие, как и более глубокая, теперь мы будем использовать небольшую AE, чтобы определить, можно ли использовать вспомогательную потерю для улучшения. Это позволит сделать так, чтобы автокодер одновременно был обучен и воспроизводил свой основной ввод (гистограммы слов), а также использовать кодировку для прогнозирования относительной отдачи. Вес вспомогательных потерь будет установлен равным 0,25, а основных потерь – 0,75. Вспомогательный выходной слой напрямую связан со скрытым слоем с помощью линейных активаций, что явно поможет нам

понять скрытое представление содержания информации о том, насколько хорошо акции будут работать в будущем» [14].

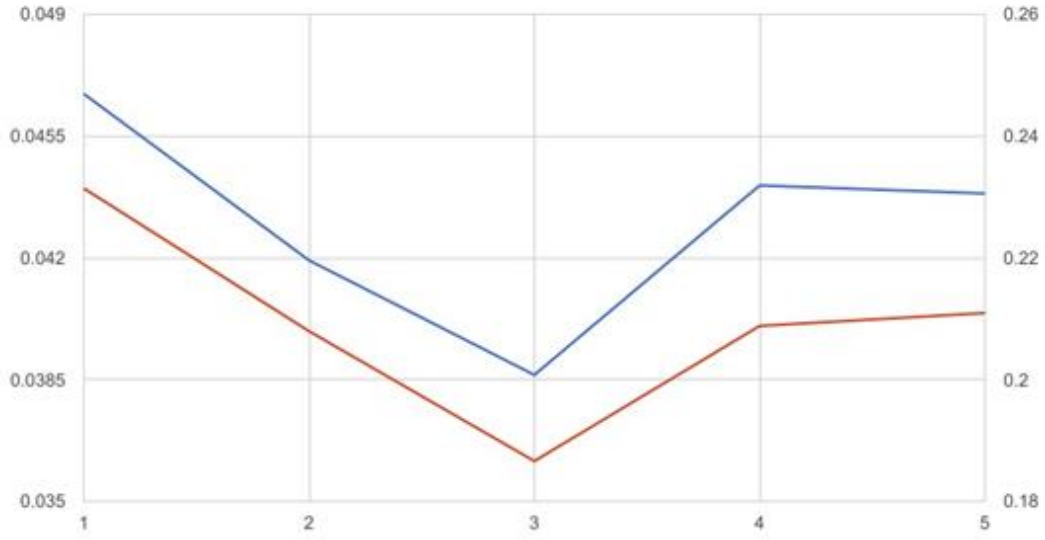

Рисунок 12 – Проверка IC и оценка $S_{0,15}$ в качестве объектов, созданных более крупными AE

Вместо прогнозирования только 120-дневного возврата значения возврата за 5, 10, 20, 60 и 120 дней были объединены в целевой вектор. Процесс сопряжения набора возвращаемых значений с каждой подачей был нетривиальным, и в итоге мы нашли все возвращаемые значения после выпуска заявки, но до следующей заявки, и усреднили их. Цель состоит в том, чтобы это дало краткое описание эффекта, который оказала подача заявки. На рисунке 13 мы можем видеть производительность проверки производительности модели по мере добавления функций. Производительность значительно выше, чем у небольшого AE с вспомогательными потерями.

После пяти функций он достигает $S_{0,15}$ и IC примерно 0,33 и 0,06 соответственно.

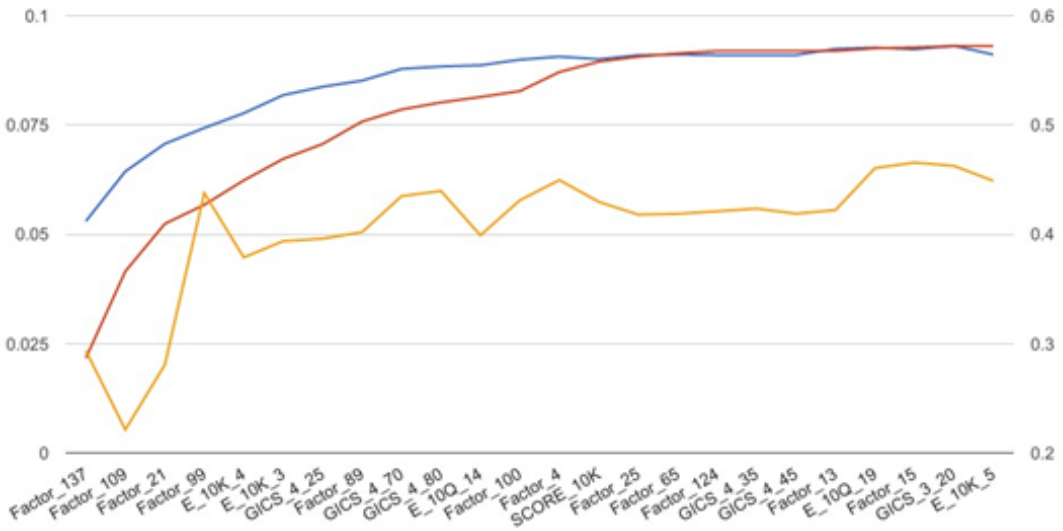

Рисунок 13 – Добавлены функции проверки IC, оценка $S_{0,15}$ и функции тестирования

Все протестированные до сих пор модели были "универсальными" моделями, поскольку они обучены делать прогнозы для акций во всех секторах. Другой подход заключается в создании моделей на уровне секторов, где каждый сектор или аналогичные секторы имеют свою собственную прогностическую модель. Это может быть желательным подходом по нескольким причинам. Наличие портфеля, в котором обязательно должны быть стратегии для многих секторов, гарантирует, что если в одном секторе неожиданно окажутся отрицательные результаты, то есть много других секторов, результаты которых могут перевесить потери. Вполне возможно, что закономерности, которые появляются в одном разделе и привели к убыткам, не связаны с закономерностями, возникающими в других секторах. У этого подхода также есть потенциальные недостатки. Это ограничивает общие знания – модели, которые появляются в одном секторе, могут помочь предсказать другой. Это также сокращает объем обучающих данных для каждой модели.

«Группировка секторов. GICS, используется для определения сектора

запаса на четырех уровнях возрастающей специфичности. На высшем уровне у нас есть акции в 9 различных секторах: 10, 20, 25, 30, 35, 40, 45, 50, 55. Однако количество запасов в каждом секторе не сбалансировано, и некоторые секторы могут содержать более разнообразный набор запасов. Чтобы справиться с этим, мы решили разделить вводные группы секторов. Для каждой из этих групп секторов (которые мы также будем называть просто секторами) мы подготовим модель» [50].

«Модельное Обучение. Используя 16 предопределенных секторов, мы создадим 16 моделей. То есть, используя данные обучения и проверки, связанные с каждым сектором, мы будем использовать линейную регрессию методом наименьших квадратов и прямой выбор объектов с S_s в качестве функции оценки набора объектов. Однако вместо s = 0,15 используется 0,10. Для каждой из моделей отраслевого уровня выбор функций выполнялся до тех пор, пока оценка валидации не начала снижаться или не выровнялась, максимум до 15 функций» [52].

Выполнив выбор функций для каждой модели сектора, мы видим в таблице 5, что количество функций, выбранных для модели каждого сектора, варьируется от 1 до 15.

Таблица 5 – Количество функций, выбранных для каждой модели сектора

Сектор групп	Выбранные функции
10	6
15	11
2010+2020	1
2030	13
2510+2520	10
2530+2540+2550	12
30	1
3510	15
3520	12
4010	11
4020+4030	15
4040	5
4510	15

Продолжение таблицы 5

Сектор групп	Выбранные функции
4520+4530	12
50	12
55	6

«Мы также можем увидеть тестовую микросхему для наших моделей (SLR-FS).

Производительность модели, предоставленной Highstreet.

Наиболее заметным аспектом результатов является большая разница в показателях сектора для каждой из моделей, а также между моделями.

Однако характеристики этих двух моделей не являются полностью независимыми; корреляция между показателями секторов составляет примерно 0,6, что указывает на то, что некоторые секторы в целом могут быть более предсказуемыми, чем другие.

Размер секторов также может влиять на производительность модели; корреляция между количеством тестовых выборок для каждого сектора и соответствующей производительностью модели сектора составляет около 0,35 как для наших моделей, так и для моделей Highstreet (рисунок 14)» [27].

Точно сравнить производительность нашей модели и модели Highstreet непросто.

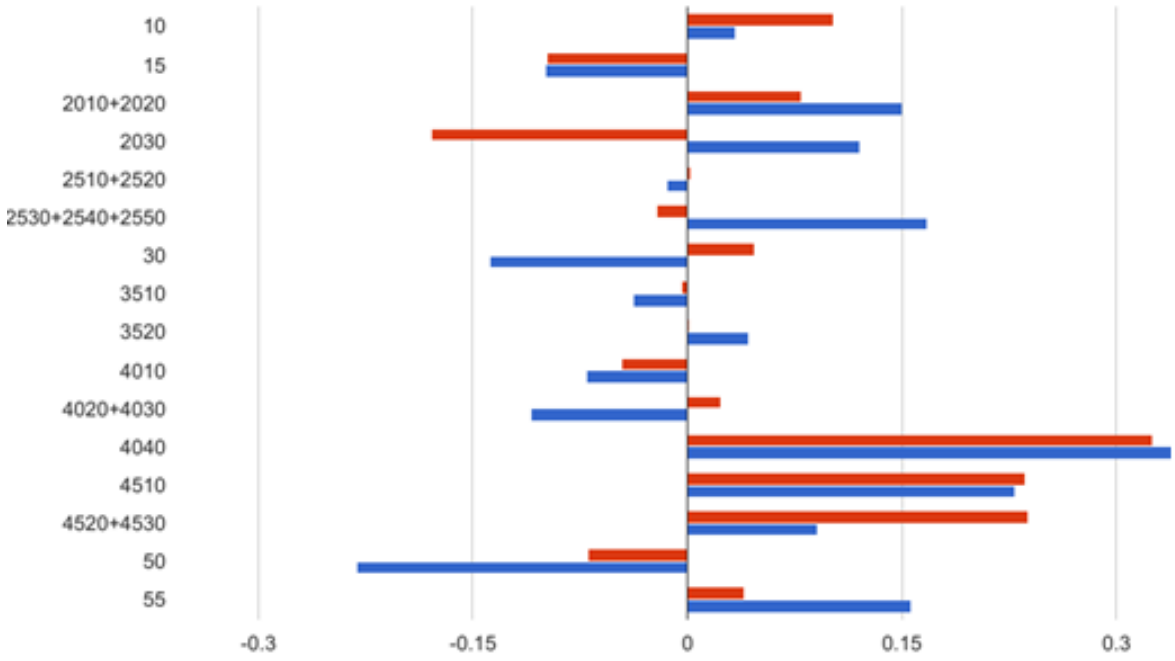

Рисунок 14 – Сравнение модели на уровне сектора и модели Highstreet

«Взять среднее значение показателей по секторам нереально, поскольку размеры секторов значительно различаются. Использование средневзвешенного значения по размеру сектора также не очень точно, поскольку размер секторов не является постоянным с течением времени. Чтобы получить приблизительное значение, мы будем использовать средневзвешенное значение, где вес - это количество тестовых образцов для каждого сектора. Выполнение этого показывает, что общие характеристики довольно близки: IC для модели Highstreet составляет 0,0625, а IC для нашей модели составляет 0,0618. Учитывая разрозненные характеристики двух моделей на отраслевом уровне, было бы полезно их объединить. Есть некоторые сектора, где каждая модель имеет явную выгоду, что говорит о том, что использование нашего алгоритма для одних секторов, а не для других, может быть полезным, однако требуется более тщательное тестирование, чтобы подтвердить, что показатели сектора одинаковы в разные годы (таблица 6)» [44].

Таблица 6 – Сравнение модели отраслевого уровня

Сектор	Highstreet модель	SLR-FS
10	0,1020	0,0341
15	-0,0980	-0,0998
2010+2020	0,0795	0,1499
2030	-0,1795	0,1202
2510+2520	0,0018	-0,0143
2530+2540+2550	-0,0216	0,1673
30	0,0471	-0,1385
3510	-0,0040	-0,0379
3520	0,0009	0,0434
4010	-0,0461	-0,0705
4020+4030	0,0239	-0,1096
4040	0,3251	0,3382
4510	0,2359	0,2291
4520+4530	0,2379	0,0907
50	-0,0695	-0,2309
55	0,0400	0,1565

В этой работе мы рассмотрели проблему прогнозирования динамики запасов. Несмотря на значительный объем исследований по этой теме, очень мало исследований направлено на долгосрочное прогнозирование с использованием методов машинного обучения и текстовых источников данных. Мы подготовили данные о запасах за более чем десять лет и предложили решение, которое сочетает в себе функции текстовых годовых и квартальных отчетов с фундаментальными факторами для долгосрочного прогнозирования динамики запасов. Кроме того, мы разработали новый метод извлечения объектов из текста с целью прогнозирования производительности и применили выбор объектов с помощью новой функции оценки.

Для создания эффективных моделей мы столкнулись с двумя основными проблемами. Первым был вопрос об эффективности рынка, который накладывает теоретические ограничения на то, как можно найти закономерности на фондовых рынках с целью прогнозирования. Это свойство может стать конкретной проблемой, поскольку в данных проявляются закономерности, которые бесполезны или даже вредны для прогнозирования

будущих значений. «То, как мы пытались справиться с этим, заключалось в тщательном разделении наших данных на данные обучения, проверки и тестирования с расширением окон, чтобы максимально использовать их, пытаясь, избегайте случайного переоснащения. Второй способ, которым мы справились с этой проблемой, заключался в использовании специально разработанной метрики производительности модели S_s, которая была направлена на обеспечение хорошей производительности моделей при тестировании не только за счет максимизации IC валидации модели, но и за счет минимизации различий между годами валидации этого значения. Третий способ, которым мы рассматривали эффективность рынка, заключался в выполнении выбора функций с использованием S_s, чтобы удалить те функции, которые работали плохо или ненадежно. Второй набор проблем был связан с составлением набора данных для использования в экспериментах и тестировании. Из-за большого объема данных необходимо было соблюдать осторожность при их очистке и подготовке, а неизбежные ошибки на этом пути требовали повторной обработки данных. Кроме того, выбросы и недостающие значения в данных были устранены несколькими способами. Используя экспертные знания, мы определили, как решать различные проблемы в данных, и в итоге использовали замену среднего значения и удаление объектов» [13].

Наши результаты показывают, что, используя только линейную регрессионную модель наименьших квадратов, прямой выбор функций и метрику S_s в сочетании с основами, функциями GICS и функциями простой регистрации, мы значительно улучшили базовые показатели. «В то время как максимальный базовый IC составлял 0,05269, мы достигли 0,08714. Самый высокий тестовый IC, полученный в базовом режиме, составил 0,01560, тогда как мы достигли 0,06469. Наши эксперименты также показали, что, хотя наша новая методика извлечения объектов из текста с помощью автокодеров, позволила получить IC проверки 0,09111, она не улучшила тестовую IC, достигнув 0,06222. Мы сравнили наш набор моделей с собственными

моделями Highstreet, хотя сравнить характеристики непросто, обе модели достигают приблизительного значения тестовой IC 0,062» [41]. Поскольку наш метод позволяет создавать модели, которые работают сопоставимо с моделями ручной работы, не требуя при этом экспертных знаний, помимо подготовки данных, это делает модель привлекательным помощником для построения инвестиционных портфелей.

Выводы по главе 4

Мы подготовили данные о запасах за более чем десятилетний период и предложили решение, которое сочетает функции текстовых годовых и квартальных отчетов с фундаментальными факторами для долгосрочного прогнозирования динамики запасов. Кроме того, мы разработали новый метод извлечения объектов из текста с целью прогнозирования производительности и выбора прикладных объектов с помощью новой функции оценки.

Заключение

Научно исследовательская работа посвящена моделям и алгоритмам прогнозирования фондового рынка на основе интеллектуального анализа текста.

Исследований в области прогнозирования довольно много, но большая часть использует методы машинного обучения и фундаментального анализа. После анализа источников, заметно, что метод анализа текста используют не многие и в основном для долгосрочного или краткосрочного периода, где не учитываются факторы, например, кризис. Можно сделать вывод, что проблема исследования прогнозирования фондового рынка с помощью анализа текста актуальна.

Для исследования будем использовать прогнозирование фондового рынка в условиях кризиса с помощью анализа текста и автокодера, которые в совокупности должны дать благоприятный прогноз для компаний и организаций.

В этой работе мы рассмотрели проблему прогнозирования динамики запасов. Несмотря на значительный объем исследований по этой теме, очень мало исследований направлено на долгосрочное прогнозирование с использованием методов машинного обучения и текстовых источников данных. Мы подготовили данные о запасах за более чем десять лет и предложили решение, которое сочетает в себе функции текстовых годовых и квартальных отчетов с фундаментальными факторами для долгосрочного прогнозирования динамики запасов. Кроме того, мы разработали новый метод извлечения объектов из текста с целью прогнозирования производительности и применили выбор объектов с помощью новой функции оценки.

Наши результаты показывают, что, используя только линейную регрессионную модель наименьших квадратов, прямой выбор функций и метрику S_s в сочетании с основами, функциями GICS и функциями простой регистрации, мы значительно улучшили базовые показатели. В то время как

максимальный базовый IC составлял 0,05269, мы достигли 0,08714. Самый высокий тестовый IC, полученный в базовом режиме, составил 0,01560, тогда как мы достигли 0,06469. Наши эксперименты также показали, что, хотя наша новая методика извлечения объектов из текста с помощью автокодеров, позволила получить IC проверки 0,09111, она не улучшила тестовую IC, достигнув 0,06222. Мы сравнили наш набор моделей с собственными моделями Highstreet, хотя сравнить характеристики непросто, обе модели достигают приблизительного значения тестовой IC 0,062.

Задачи, определённые для достижения цели работы, были выполнены в полном объёме.

Цель научно исследовательской работы была достигнута – были построены модели и алгоритмы прогнозирования фондового рынка на основе интеллектуального анализа текста.

Гипотеза исследования подтверждена.

Список используемых источников

1. Анализ временных рядов: учебное пособие для бакалавриата и магистратуры/ О. А. Подкорытова, М. В. Соколов. – М.: Издательство Юрайт, 2016. – 266 с. – Серия: Бакалавр и магистр. Модуль.

2. Ицхоки О. Выбор модели и парадоксы прогнозирования // Квантиль. 2006. № 9. С. 43–51.

3. Лысенко, В. Д. Анализ тональности текста для прогнозирования цен на фондовом рынке / В. Д. Лысенко. – Текст: непосредственный // Молодой ученый. – 2018. – № 22 (208). – С. 420-423. – URL: https://moluch.ru/archive/208/51025/ (дата обращения: 21.12.2022).

4. Митина О.В., Евдокименко А.С., Методы анализа текста: методологические основания и программная реализация. – Текст: непосредственный // Молодой ученый. – 2014. – № 20 (301). – С. 29-38.

5. Молчанов А. А. Использование GARCH модели для исследования динамики курса валют./ А.А.Молчанов// Психолого-педагогический журнал Гаудеамус. №2 (20). – 2012. С. 222 – 229.

6. Морозова, В. И. Прогнозирование методом машинного обучения / В. И. Морозова, Д. И. Логунова. – Текст: непосредственный // Молодой ученый. – 2022. – № 21 (416). – С. 202-204. – URL: https://moluch.ru/archive/416/92048/ (дата обращения: 29.12.2022).

7. Николаев С.В., Пронина О.Ю., Баженов Р.И. Исследование методов интеллектуального анализа для формирования краткосрочного прогноза в программной среде Statistica // Экономика и менеджмент инновационных технологий. 2015. № 7 [Электронный ресурс]. URL: https://ekonomika.snauka.ru/2015/07/9500 (дата обращения: 15.12.2022).

8. Обучаем компьютер чувствам (sentimentanalysis по-русски) // habr.com. URL: https://habr.com/post/149605/ (дата обращения: 21.12.2022).

9. Современные методы анализа тональности текста // http://datareview.info. URL: http://datareview.info/article/sovremennyie-metodyi-

analiza-tonalnosti-teksta/ (дата обращения: 21.12.2022).

10. Яковлева К. Оценка экономической активности на основе текстового анализа // Серия докладов об экономических исследованиях Центральный банк Российской Федерации. 2017. № 25. Октябрь.

11. A. I. Ivanus, ”Text mining applied to patent mapping: a practical business case,” World Patent Information, 25, 335 (2020).

12. Anastasakis L., Mort N. Exchange rate forecasting using a combined parametric and nonpar-ametric self-organising modelling approach // Expert Systems with Applications. 2009. Vol. 36. P. 12001-12011.

13. Azad A.K., and Mahsin M. Forecasting Exchange Rates of Bangladesh using ANN and ARIMA models: A comparative study // International Journal of Advanced Engineering Science & Technologies, 10(1). 2011. p. 31-36.

14. Balahur A., Steinberger R., Goot E. v. d., Pouliquen B., Kabadjov M. Opinion Mining on Newspaper Quotations // Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology. 2009. Vol. 03. P. 523-526.

15. Berlanga-Llavori R, Anaya-Sánchez H, Pons-Porrata A, Jiménez-Ruiz E. Conceptual subtopic identification in the medical domain. In: Geffner H, Prada R, Machado Alexandre I, David N, editors. Advances in artificial intelligence – IBERAMIA 2008. Lecture notes in computer science, vol 5290. Springer; 2008. p. 312–21.

16. Bollen J., Huina M., Zeng Xiao-Jun Twitter mood predicts the stock // Journal of Computational Science. 2010. Vol. 2. P. 1-8.

17. Cambria E., Schuller B., Yunqing X., Havasi C. New Avenues in Opinion Mining and Sentiment Analysis // IEEE Intelligent Systems. 2013. Vol. 28. P. 15-21.

18. Chatrath A., Miao H., Ramchander S., Villupuram S. Currency jumps, cojumps and the role of macro news // Journal of International Money and Finance. 2014. Vol. 40. P. 42-62.

19. Datta K. ARIMA Forecasting of Inflation in the Bangladesh Economy //

The IUP Journal of Bank Management, X(4). 2011. p. 7-15.

20. Fasanghari M., Montazer G. A. Design and implementation of fuzzy expert system for Tehran Stock Exchange portfolio recommendation // Expert Systems with Applications. 2010. Vol. 37. P. 6138-6147.

21. Friesen G., Weller P. A. Quantifying cognitive biases in analyst earnings forecasts // Journal of Financial Markets. 2006. Vol. 9. P. 333-365.

22. G. E. Hinton, R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. DOI: 10.1126/science.1127647. 2006. - https://pdfs.semanticscholar.org/7d76/b71b700846901ac4ac119403aa7 37a285e36.pdf

23. Hagenau M., Liebmann M., Neumann D. Automated news reading: Stock price prediction based on financial news using context-capturing features // Decision Support Systems. 2013. Vol. 55. No. 3. P. 685-697.

24. Heeyoung Lee, ”Forecasting Vacant Technology of Patent Analysis System using Self Organizing Map and Matrix Analysis,” Journal of the Korea Contents Association, 10, 462 (2010).

25. Hong Keel Sul, Alan R. Dennis, Lingyao (Ivy) “Trading on Twitter: The Financial Information Content of Emotion in Social Media”, 2014 47th Hawaii International Conference on System Science.

26. Huang C.-J., Liao J.-J., Yang D.-X., Chang T.-Y., Luo Y.-C. Realization of a news dissemination agent based on weighted association rules and text mining techniques // Expert Systems with Applications. 2010. Vol. 37. P. 6409-6413.

27. Huang S.-C., Chuang P.-J., Wu C.-F., Lai H.-J. Chaos-based support vector regressions for exchange rate forecasting // Expert Systems with Applications. 2010. Vol. 37. P. 8590-8598.

28. Hurtado J, Huang S, Zhu X. Topic discovery and future trend prediction using association analysis and ensemble forecasting. In: the 16th IEEE international conference on information reuse and integration. San Francisco, CA: 2015.

29. Kaltwasser P. R. Uncertainty about fundamentals and herding behavior in the FOREX market // Physica A: Statistical Mechanics and its Applications. 2010.

Vol. 389, No. 8. P. 12151222.

30. Kamaladdin Fataliyev, Aneesh Chivukula, Mukesh Prasad, and Wei Liu. 2021. Text-based Stock Market Analysis: A Review. 1, 4 (January 2023), 30 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

31. Khadjeh Nassirtoussi A., Aghabozorgi S., Ying Wah T., Ngo D. C. L. Text mining for market prediction: A systematic review // Expert Systems with Applications. 2014. Vol. 41. P. 76537670.

32. Khadjeh Nassirtoussi A., Ying Wah T., Ngo Chek Ling D. A novel FOREX prediction methodology based on fundamental data // African Journal of Business Management. 2011. Vol. 5. P. 8322-8330.

33. K. Kasravi and M. Risov, ”Patent Mining - Discover y of Business Value from Patent Repositories,” Proceedings of 40th Annual Hawaii International Conference on System Sciences, 54 (2007).

34. Kleinnijenhuis J., Schultz F., Oegema D. Atteveldt W.H. van. Financial News and Market Panics in the age of high frequency trading algorithms // Journalism. 2013. Vol. 14, No. 2. P. 271-291

35. Kontopoulos E., Berberidis C., Dergiades T., Bassiliades N. Ontologybased sentiment analysis of twitter posts // Expert Systems with Applications. 2013. Vol. 40. P. 4065-4074.

36. Loia V., Senatore S. A fuzzy-oriented sentic analysis to capture the human emotion in Web-based content // Knowledge-Based Systems. 2014. Vol. 58. P. 75-85.

37. Lupiani-Rui E., García-Manota I., Valencia-Garcí R., García-Sánchez F., Castellanos-Nieves D., Fernández-Breis J. T. et al. Financial news semantic search engine // Expert Systems with Applications. 2011. Vol. 38. P. 15565-15572.

38. Majumder D. Towards an efficient stock market: Empirical evidence from the Indian market // Journal of Policy Modeling. 2013. Vol. 35. P. 572-587.

39. Merh N., V. P. Saxena V.P. and Pardasani K.R. Next Day Stock market Forecasting: An Application of ANN & ARIMA // The IUP Journal of Applied Science, 17(1). 2011. p. 70-84.

40. Ortigosa-Hernández J., Rodríguez J. D., Alzate L., Lucania M., Inza I., Lozano J. A. Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers // Neurocomputing. 2012. Vol. 92. P. 98-115.

41. Peramunetilleke D., Wong R. K. Currency exchange rate forecasting from News Headlines // Australian Computer Science Communications. 2002. Vol. 24. P. 131-139.

42. Premanode B., Toumazou C. Improving prediction of exchange rates using differential EMD // Expert Systems with Applications. 2013. Vol. 40. P. 377-384.

43. Robertson C., Geva S., Wolff R. What types of events provide the strongest evidence that the stock market is affected by company specific news? // Proceedings of the fifth Australasian conference on Data mining and analystics. 2006. Vol. 61. P. 145-153.

44. Schumaker R. P., Zhang Y., Huang C.-N., Chen H. Evaluating sentiment in financial news articles // Decision Support Systems. 2012. Vol. 53. P. 458-464.

45. Sermpinis G., Laws J., Karathanasopoulos A., Dunis C. L. Forecasting and trading the EUR/USD ex-change rate with gene expression and psi sigma neural networks // Expert Systems with Applications. 2012. Vol. 39, No. 10. P. 8865-8877.

46. S. Jun and D. Uhm, ”Patent and Statistics, What’s the connection?” Communications of the Korea Statistical Society, 17, 205 (2010).

47. Sunghae Jun, ”Development of New Technology Forecasting Algorithm: Hybrid Approach for Morphology Analysis and Conjoint Analysis of Patent Information,” IEEE Transactions on Engineering Management, 54, 588 (2007).

48. Uko A.K. and Nkoro E. Inflation Forecasts with ARIMA, Vector Autoregressive and Error Correction Models in Nigeria // European Journal of Economics, Finance & Administrative Science. 2012. № 50. p. 71-87

49. Vanstone B., Finnie G. An empirical methodology for developing stockmarket trading systems using artificial neural networks // Expert Systems with Applications. 2009. Vol. 36. P. 6668-6680.

50. Vanstone B., Finnie G. Enhancing stockmarket trading performance with

ANNs // Expert Systems with Applications. 2010. Vol. 37. P. 6602-6610.

51. Wong H., Tu Y. and Wang C. Application of fuzzy time series models for forecasting the amount of Taiwan export // Experts Systems with Applications, 2010. p. 1456-1470.

52. Yu H., Nartea G. V., Gan C., Yao L. J. Predictive ability and profitability of simple technical trading rules: Recent evidence from Southeast Asian stock markets // International Review of Economics & Finance. 2013. Vol. 25. P. 356-371.

53. Yu L.-C., Wu J.-L., Chang P.-C., Chu H.-S. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news // Knowledge-BasedSystems. 2013. Vol. 41. P. 89-97.