**Содержание**

[Введение 7](#_Toc167908548)

[1 Аналитическая часть 9](#_Toc167908549)

[1.1 Исследование и анализ предметной области 9](#_Toc167908550)

[1.2 Анализ моделей, методов и программных продуктов для решения подобных задач 18](#_Toc167908551)

[1.3 Постановка и формализация задачи на разработку 28](#_Toc167908552)

[2 Проектная часть 32](#_Toc167908553)

[2.1 Разработка методов и алгоритмов для решения поставленной задачи 32](#_Toc167908554)

[2.2 Проектирование моделей базы данных: концептуальной, логической, физической 34](#_Toc167908555)

[2.3 Проектирование программного обеспечения с обоснованием выбора методологии и среды проектирования и программирования 34](#_Toc167908556)

[2.4 Вычислительные эксперименты и анализ их результатов 49](#_Toc167908557)

[2.5 Возможные улучшения программы 52](#_Toc167908558)

[Заключение 53](#_Toc167908559)

[Список литературы 54](#_Toc167908560)

[Приложение А 57](#_Toc167908561)

# Введение

В современном мире компьютеры играют все более важную роль в повседневной жизни, а большинство информации сохраняется в электронном формате, что значительно облегчает ее хранение, доступ и обработку. Несмотря на цифровой формат, анализ многих данных вызывает определенные трудности.

Различают следующие типы данных: структурированные, частично структурированные и неструктурированные. Структурированные данные характеризуются определенной организацией и легкостью в обработке, например, базы данных. Частично структурированные данные, включают такие элементы, как электронная почта и файлы разметки (например, HTML, XML). Не структурированные данные представляют собой примерно 80% информации и вызывают сложности в извлечении и поиске нужной информации, требуя применения специализированных методов и алгоритмов обработки.

Интеллектуальный анализ текста (Text Mining) играет важную роль в обработке таких данных. Этот процесс объединяет несколько дисциплин, включая обработку веб-данных, информационный поиск, компьютерную лингвистику и обработку естественного языка. В сравнении с анализом структурированных данных, интеллектуальный анализ текста имеет большее значение, поскольку большая часть информации сохраняется в текстовом формате, и, следовательно, требует более сложных подходов для обработки.

Анализ новостей имеет широкий спектр практических применений, включая финансовые рынки, маркетинг, мониторинг общественного мнения и прогнозирование событий.

Современная финансовая среда предполагает, что инвесторы имеют доступ к разнообразной информации о компаниях, включая новости, финансовые отчеты, аналитические обзоры и другие материалы. Эта информация является ключевым фактором в процессе принятия инвестиционных решений, поскольку она может влиять на оценку рисков и потенциальной доходности компаний.

Инвесторы, имея доступ к актуальной и релевантной информации, могут более точно оценить финансовое состояние компании, ее перспективы на рынке, возможные риски и потенциальную доходность. Это позволяет принимать более обоснованные инвестиционные решения и минимизировать финансовые риски.

В связи с этим, сегодня большое внимание уделяется изучению взаимосвязи между информацией, доступной инвесторам через СМИ, и общим риском-доходностью компаний. Анализ этой взаимосвязи позволяет понять, какие новости и события могут повлиять на цену акций компании, ее капитализацию и инвестиционную привлекательность.

Таким образом, доступ к информации играет критическую роль в современной финансовой среде, и инвесторы активно используют ее для принятия обоснованных инвестиционных решений.

Методы машинного обучения позволяют эффективно извлекать, классифицировать и анализировать информацию из СМИ для выявления связей между новостными сюжетами и рыночной динамикой.

# Аналитическая часть

## **Исследование и анализ предметной области**

Разработка программного обеспечения для анализа влияния информации в СМИ на цены акций компаний с использованием нейросетей представляет собой важную область исследований для финансовых организаций и инвесторов. Нейросети могут использоваться для обработки и анализа большого объема данных, что позволяет прогнозировать и предсказывать рыночные тенденции с высокой точностью.

Основной задачей такого программного обеспечения является выявление связей между новостями, публикуемыми в СМИ, и изменениями в рыночной ситуации. Нейросети могут быть обучены на большом объеме данных, включая тексты статей, фото, видео и социальные медиа, для выявления паттернов и трендов.

Преимущества использования нейросетей в данном контексте включают встроенную способность к извлечению признаков и обработке неструктурированных данных, а также способность к обучению на большом объеме информации без явного программирования.

Однако, следует отметить, что нейросети требуют больших объемов данных для обучения и могут быть сложны в интерпретации результатов. Также важно учитывать этические и юридические аспекты использования данных из масс-медиа для принятия инвестиционных решений.

В целом, разработка такого программного обеспечения может быть мощным инструментом для принятия обоснованных инвестиционных решений, учитывающих широкий спектр факторов и событий.

Предметная область представляет собой комплексное исследование, сочетающее в себе элементы финансовой аналитики, информационных наук, искусственного интеллекта и программной разработки.

Данная тема актуальна в современном мире, где информация становится ключевым ресурсом для принятия инвестиционных решений. СМИ играют значительную роль в формировании представлений инвесторов о компаниях, рынке ценных бумаг, макроэкономических событиях и т.д. Однако не всегда информация, представленная в СМИ, является объективной и достоверной, что может влиять на риск и доходность инвестиций.

Анализ влияния информации в СМИ требует разработки специализированного программного обеспечения, которое будет уметь собирать, анализировать и интерпретировать данные из различных источников (новостных сайтов, социальных сетей, форумов и т.д.), а также проводить статистический анализ и прогнозирование.

Программное обеспечение должно иметь функционал для определения влияния конкретных новостей и медиа-событий на изменение цен на рынке ценных бумаг, а также оценки рисков и доходности инвестиций. Это может включать в себя разработку алгоритмов машинного обучения для выявления паттернов и трендов в информационных потоках, а также интеграцию с финансовыми данными и индикаторами.

Такой анализ позволит инвесторам и трейдерам принимать обоснованные решения, учитывая влияние информации в СМИ на финансовые рынки. Также это способствует более глубокому пониманию динамики рынка и повышению эффективности инвестиционной стратегии.

Для решения поставленных задач необходимо использовать методы анализа текстов из различных источников СМИ с использованием методов обработки естественного языка, что позволит выделить актуальные темы и определить тональность информации. Этот подход позволит извлечь значимую информацию о компаниях из новостных материалов, аналитических обзоров, интервью и других источников, которые влияют на рыночные параметры.

Также необходимо использовать статистический анализ данных финансовых рынков для выявления корреляций между опубликованной информацией в СМИ и рыночными показателями компаний. Это включает в себя проведение корреляционного анализа между временными рядами финансовых данных и параметрами новостной информации, а также оценку влияния различных событий на изменения котировок акций, объемов торгов и других финансовых показателей.

Для обработки больших объемов информации из СМИ необходимо применить методы машинного обучения, такие как классификация текстов, кластеризация новостных статей для выявления тематических групп, анализ сентимента для оценки тональности информации либо другие методы, которые позволят эффективно обрабатывать и анализировать массу данных из различных источников. Это поможет выявить взаимосвязь между информацией в СМИ и динамикой рыночных параметров компаний, а также углубить понимание воздействия новостей и медийной активности на поведение инвесторов и рыночную реакцию.

Таким образом, изучение влияния информации, представленной в СМИ, представляет большой интерес, поскольку может помочь инвесторам и управляющим портфелями принимать более обоснованные инвестиционные решения. Кроме того, данное исследование имеет практическую значимость, поскольку может помочь компаниям понять, как их известность влияет на их финансовые показатели.

### Финансовые определения

Финансовый рынок представляет собой совокупность инструментов и экономических отношений между участниками, которые связаны с созданием, поддержанием и обращением капитала. Этот термин является довольно абстрактным, и под ним часто подразумеваются более конкретные рынки, такие как:

* рынок купонных и бескупонных облигаций;
* фондовый рынок (рынок акций);
* валютный рынок.

Несмотря на выделение отдельных составляющих, каждый из этих рынков является частью единой системы, в которой финансы перемещаются между всеми конкретными рынками.

Каждый из финансовых рынков служит посредником между первоначальными владельцами финансов и их конечными пользователями. Если рынок основан на финансах как на капитале, он называется фондовым рынком и выступает в качестве составной части всего финансового рынка.

В России финансовые рынки функционируют под влиянием ряда специфических характеристик:

* Инвестиции в экономику страны. Это один из ключевых факторов, определяющих развитие финансовых рынков. Инвестиции способствуют росту экономики и созданию новых рабочих мест, что, в свою очередь, влияет на спрос и предложение на финансовых рынках.
* Международные рынки и тенденции глобализации. Россия является частью мировой финансовой системы, и её рынки подвержены влиянию глобальных тенденций. Это может проявляться в виде колебаний валютных курсов, изменения процентных ставок и других макроэкономических показателей.
* Современные компьютерные технологии. Развитие информационных технологий оказывает значительное влияние на финансовые рынки. С одной стороны, это упрощает доступ к информации и снижает издержки, связанные с операциями на рынке. С другой стороны, это создаёт новые риски, связанные с кибербезопасностью и защитой данных.
* Уровень компьютерной и информационной развитости участников рынков. От этого зависит скорость и эффективность проведения операций на финансовом рынке, а также уровень доверия между участниками.

Финансовый рынок может быть классифицирован по различным признакам:

* По стадиям обращения ценных бумаг:
  + Первичный рынок – рынок, на котором происходит первичное размещение ценных бумаг.
  + Вторичный рынок – рынок, где обращаются ранее выпущенные ценные бумаги.
* По степени организованности:
  + Организованный рынок – это рынок, который функционирует по определённым правилам и стандартам. Он включает в себя биржи и другие торговые площадки.
  + Неорганизованный рынок – это рынок, на котором сделки заключаются вне организованных площадок.
* По месту торговли:
  + Биржевой рынок – торговля ценными бумагами осуществляется на фондовых биржах.
  + Внебиржевой рынок – сделки совершаются вне бирж, например, через дилеров или брокеров.
* По типу торговли:
  + Традиционный рынок – операции проводятся в режиме реального времени с использованием торговых терминалов.
  + Компьютеризированный рынок – торги осуществляются с помощью электронных систем, таких как интернет-трейдинг.
* По способу исполнения сделок:
  + Кассовый рынок – расчёты по сделкам производятся немедленно.
  + Срочный рынок – исполнение сделок откладывается на определённый срок.

Участники рынка ценных бумаг – это физические лица или компании, которые участвуют в операциях с ценными бумагами, обеспечивая их оборот и расчёты по ним.

Основные участники рынка – эмитенты, выпускающие акции или облигации для привлечения финансирования или размещения свободных денежных средств. Эмитентами могут быть государство, субъекты государства или коммерческие предприятия. На первичном рынке цель эмитентов – разместить запланированный транш по максимальной цене.

Инвестор – это лицо, которое заинтересовано во вложении капитала в ценные бумаги. Цель инвесторов – максимально выгодно приобрести ценные бумаги наиболее перспективных компаний.

### Литературный обзор

За последние несколько десятилетий стало проводиться все больше исследований на тему прогнозирования цен на акции, поскольку фондовый рынок продолжает расширяться. В этих исследованиях делается попытка проанализировать и спрогнозировать колебания цен на акции. Цены на акции отличаются высокой динамичностью, нелинейностью и высокой степенью зашумленности. На цену отдельных акций влияют различные факторы, включая мировую экономику, политические события, политику правительства, природные и техногенные катастрофы, поведение инвесторов и другие факторы. Это делает прогнозирование временных рядов сложной задачей.

Li, Bu & Wu (2017) и Liu и др. (2017) описали некоторые исследования фондового рынка, основанные на теории случайных блужданий и гипотезе эффективного рынка. Эти исследования показали, что колебания цен на акции носят случайный характер и не поддаются прогнозированию. Согласно гипотезе эффективного рынка, на цены акций влияют новости, а не текущие или прошлые цены, поскольку новости непредсказуемы. Это означает, что цены на акции следуют случайной схеме, и их трудно предсказать с точностью более 50%. Однако более поздние исследования показали, что рыночная цена акций не является полностью случайной и в определенной степени может быть спрогнозирована.

Shah, Isah & Zulkernine (2019) и Selvin и др. (2017) провели исследование фондового рынка и попытались предсказать цены на акции. Они использовали два широко используемых метода: фундаментальный анализ и технический анализ. Фундаментальный анализ – это инвестиционный подход, который оценивает стоимость акций на основе профиля компании, перспектив отрасли, политических и экономических факторов, а также данных новостей и социальных сетей. Эта информация, как правило, неструктурирована и лучше всего подходит для долгосрочных прогнозов. Технический анализ, с другой стороны, пытается использовать исторические курсы акций для прогнозирования будущих тенденций. Он включает в себя регистрацию ежедневных колебаний цен в виде графиков и определение оптимальных точек покупки и продажи на основе наблюдаемых изменений. Обычно используемые алгоритмы при таком подходе включают K-линии, скользящие средние и индекс относительной силы, которые подходят для краткосрочных прогнозов.

Обзор литературы, проведенный Shah, Isah & Zulkernine (2019) и Bustos & Pomares-Quimbaya (2020), показал, что технический анализ является одним из наиболее часто используемых методов прогнозирования фондового рынка, и он широко изучался и использовался в качестве сигнала о том, когда покупать или продавать акции. Однако некоторые исследования показали, что доходность, получаемая от торговых стратегий, основанных на техническом анализе, на самом деле ограничена. Фундаментальный анализ редко использовался в традиционных исследованиях из-за сложности построения моделей на основе соответствующей информации. Тем не менее, с развитием обработки естественного языка и текстовой аналитики в некоторых недавних исследованиях была проанализирована неструктурированная информация, связанная с запасами, для повышения точности прогнозирования. Эта информация может включать официальные документы, финансовые новости и сообщения в социальных сетях.

Благодаря быстрому развитию искусственного интеллекта, Bustos & Pomares-Quimbaya (2020) и Nelson, Pereira & de Oliveira (2017) достигли многообещающих результатов, используя методы опорных векторов (SVM) и искусственные нейронные сети (ANN) для прогнозирования фондового рынка.

SVM, инструмент машинного обучения, может быть применен для решения задач классификации и регрессии. Его также можно назвать регрессией с опорными векторами (SVR), которая определяет гиперплоскость в пространствах объектов большой размерности для точного прогнозирования распределения данных.

Xia, Liu и Chen (2013) использовали SVR для разработки подхода к прогнозированию запасов, основанного на регрессионной модели, использующей информацию о временных рядах за прошлые периоды, сегодняшние цены открытия, самые высокие цены, самые низкие цены, цены закрытия, объемы торгов и скорректированные цены закрытия для прогнозирования цен открытия на следующий день.

RNN часто используются наряду с другими ANN для решения задач, связанных с временными рядами, и они также являются одним из наиболее подходящих методов для прогнозирования динамических временных рядов.

Liu и др. (2017) использовали RNN для прогнозирования волатильности акций и обнаружили, что точность модели RNN превосходит точность моделей MLP и SVM. Тем не менее, RNN могут испытывать проблемы с исчезновением градиента и резким ростом в течение нескольких итераций. Нейронные сети LSTM были разработаны для повышения производительности RNN в области искусственного интеллекта.

Gao, Chai & Liu (2017) собрали исторические торговые данные из индекса Standard & Poor's 500 (S&P 500), охватывающие последние 20 торговых дней, в качестве входных переменных, включая цену открытия, цену закрытия, самую высокую цену, самую низкую цену, скорректированную цену и объем сделки.

Они использовали нейронную сеть LSTM в качестве модели прогнозирования и оценивали ее эффективность, используя среднюю абсолютную ошибку (MAE), среднеквадратичную ошибку (RMSE), среднюю частоту абсолютных ошибок (MSE) и среднюю абсолютную процентную ошибку (MAPE). Их подход превзошел другие методы прогнозирования.

Lee & Soo (2017) объединили нейронную сеть LSTM с ANN для прогнозирования цен на тайваньские акции на основе исторических ценовых данных и финансовых новостей и обнаружили, что это уменьшает ошибку прогнозирования.

Li, Bu & Wu (2017) применили сеть LSTM для прогнозирования взлетов и падений китайского индекса CSI 300, введя данные об открытии, закрытии и объеме торгов за последние десять торговых дней, достигнув точности в 78,57%.

Khare и др. (2017) аналогичным образом обнаружили, что LSTM могут точно прогнозировать движение цен на акции.

Lu (2018) проанализировал положительные и отрицательные оценки настроений, полученные из публикаций PTT, и объединил исторические данные о биржевых транзакциях в качестве входных векторов для нейронной сети с долговременной кратковременной памятью (LSTM). Предложенный метод повысил точность прогнозирования цен на акции.

В своей статье Рябых А., Суржко Д., Коновалихи М. и Кольцов С. представили подход STTM (Sentiment-based Topic Model), который использует тематические модели и тональные потоки тем для анализа текстовых данных с целью прогнозирования направления движения финансового рынка.

На основе обработанных текстовых данных они генерируют вероятностные темы. Тематическая модель позволяет выделить основные темы в тексте и определить их вероятность появления. Тональность темы определяется функцией оценки регрессии или метрикой близости временных рядов. Это позволяет оценить эмоциональную окраску каждой темы и её влияние на рынок.

Затем в работе тональный поток тем вычисляется как произведение тональностей тем и тематических потоков для каждого момента времени. Это даёт возможность оценить общую тональность всех тем в определённый период времени. Выдвинутый ими индекс STTM представляет собой общую тональность всех тем за определённый период времени и может использоваться для прогнозирования направления движения рынка.

Однако стоит отметить, что подход STTM требует тщательной настройки параметров и проверки результатов на реальных данных. Также необходимо учитывать возможные ограничения и недостатки подхода, связанные с использованием текстовых данных и их интерпретацией.

## Анализ моделей, методов и программных продуктов для решения подобных задач

В данном обзоре рассмотрим существующие методы, используемые для разработки программного обеспечения задачи анализа влияния информации в СМИ на риск и доходность инвестирования в ценные бумаги.

Прогнозирование цен на акции является сложной задачей, требующей учёта множества факторов. Одним из подходов к решению этой задачи является использование новостных потоков. Новостные потоки содержат информацию о событиях, которые могут повлиять на цены акций. Анализируя эти данные, можно попытаться предсказать будущие изменения цен.

Для решения задачи прогнозирования цен на акции с использованием новостных потоков можно использовать различные модели, методы и программные продукты. Рассмотрим некоторые из них:

### Обзор алгоритмов машинного обучения

Регрессионные модели.

Существует множество регрессионных моделей, таких как линейная регрессия, логистическая регрессия, деревья решений, нейронные сети и другие. Каждая модель имеет свои преимущества и недостатки, поэтому важно выбрать модель, которая наилучшим образом соответствует данным и целям исследования.

Линейная регрессия – это статистический метод, который используется для моделирования взаимосвязи между двумя или более переменными. В контексте предсказания цен акций линейная регрессия может быть использована для определения зависимости между ценой акции и различными факторами, такими как новости, макроэкономические показатели и прошлые цены акций.

В линейной регрессии предполагается, что существует линейная зависимость между зависимой переменной (в данном случае, ценой акции) и независимыми переменными (например, новостями, макроэкономическими показателями и прошлыми ценами акций). Эта зависимость может быть выражена следующим уравнением:

где – зависимая переменная (цена акции), – свободный член, – коэффициент регрессии, а – независимая переменная (например, новость или макроэкономический показатель).

Линейная регрессия является простым и широко используемым методом для анализа данных. Однако она имеет некоторые ограничения. Например, она предполагает, что зависимость между переменными является линейной, что может не всегда соответствовать действительности. Кроме того, линейная регрессия чувствительна к выбросам, которые могут исказить результаты. Поэтому при использовании линейной регрессии для предсказания цен акций необходимо учитывать эти ограничения.

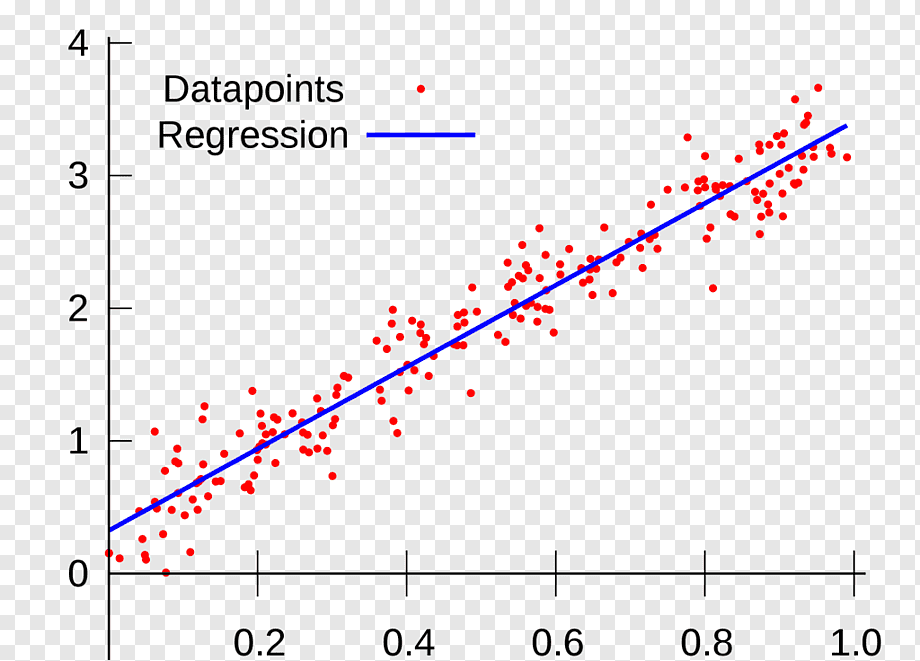


Рисунок 1.1 – пример использования линейной регрессии

Логистическая регрессия – это тип регрессионного анализа, который подходит, когда зависимая переменная принимает только два возможных значения (т.е. дихотомическое или бинарное). Этот метод часто имеет место при прогнозировании событий, когда есть только два исхода: событие произойдет (обозначается как 1) и не произойдет (обозначается как 0). Логистическая регрессия имеет некоторое сходство с линейной регрессией, поскольку она также предполагает определение значений коэффициентов для входных переменных. Однако результаты логистической регрессии преобразуются с использованием нелинейной или логистической функции, а не линейной. Как и другие методы регрессионного анализа, такая регрессия может использоваться для прогнозирования. Она используется для анализа данных и определения взаимосвязи между одной двоичной зависимой переменной и одной или несколькими независимыми переменными, которые могут быть номинальными, порядковыми или интервальными по своей природе.

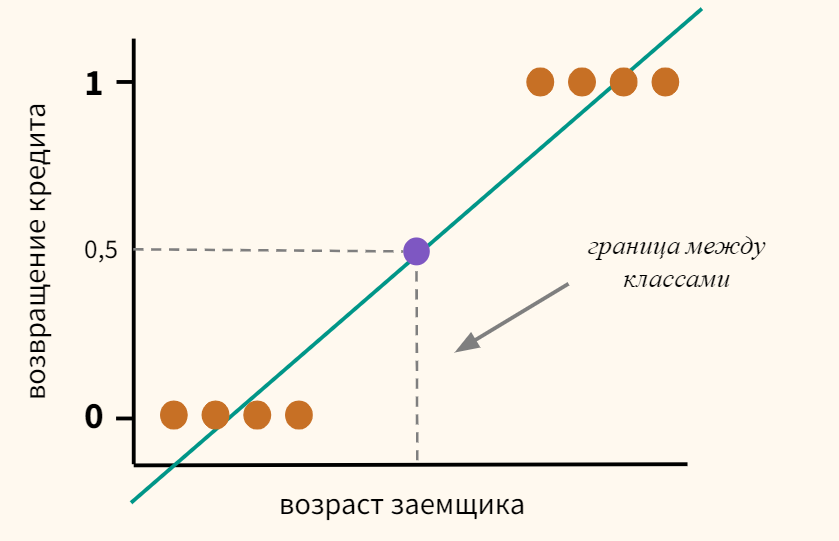


Рисунок 1.2 – пример использования логистической регрессии

Решающие деревья – это метод машинного обучения, который используется для решения задач классификации и регрессии. В контексте предсказания цен акций решающие деревья могут быть использованы для создания правил, которые определяют, как будет изменяться цена акции в зависимости от различных факторов.

Решающее дерево представляет собой иерархическую структуру, состоящую из узлов и листьев. Узлы представляют собой вопросы или условия, а листья – результаты или прогнозы. Процесс построения решающего дерева начинается с корневого узла, который содержит все данные. Затем данные разделяются на подмножества на основе вопросов или условий, заданных в узлах. Этот процесс повторяется до тех пор, пока не будут достигнуты листья, содержащие прогнозы или результаты.

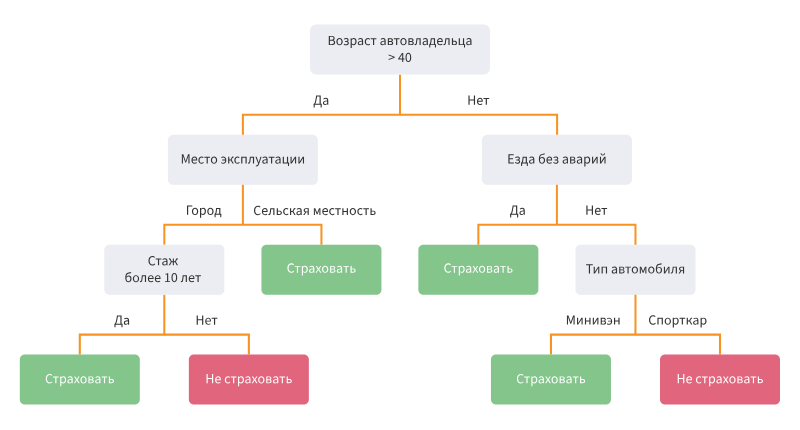


Рисунок 1.3 – пример использования решающих деревьев

Выбор модели зависит от конкретных целей исследования и доступных данных. Например, если необходимо предсказать вероятность изменения цены акции, то логистическая регрессия может быть более подходящей. Если же требуется создать правила для определения изменения цены на основе новостей, то деревья решений могут быть предпочтительнее.

Нейронные сети.

В контексте предсказания цен акций нейронные сети могут быть использованы для создания моделей, которые учитывают множество факторов, таких как новости, макроэкономические показатели и прошлые цены акций. Эти модели могут быть обучены на больших объёмах данных и способны давать точные прогнозы будущих цен.

Однако, несмотря на свою мощность, нейронные сети также имеют некоторые ограничения. Например, они могут быть чувствительны к переобучению, если модель слишком сильно адаптируется к обучающим данным. Поэтому важно тщательно выбирать параметры модели и проверять её на тестовом наборе данных перед использованием для принятия решений. Кроме того, интерпретация результатов нейронной сети может быть сложной задачей, поскольку она представляет собой «чёрный ящик», который не всегда позволяет понять, как именно были получены результаты.

Тем не менее, при правильном использовании нейронные сети могут стать мощным инструментом для прогнозирования цен акций. Они позволяют учитывать сложные зависимости между переменными и дают возможность получать точные прогнозы на основе большого объёма данных. Однако, прежде чем использовать нейронные сети для принятия инвестиционных решений, необходимо тщательно изучить их возможности и ограничения, а также провести тестирование на исторических данных.

Для задачи предсказания цен акций можно использовать различные архитектуры нейронных сетей:

Рекуррентные нейронные сети (RNN) – это тип нейронной сети, который может обрабатывать последовательные данные, такие как временные ряды. RNN могут быть использованы для анализа временных рядов цен акций и прогнозирования будущих значений.

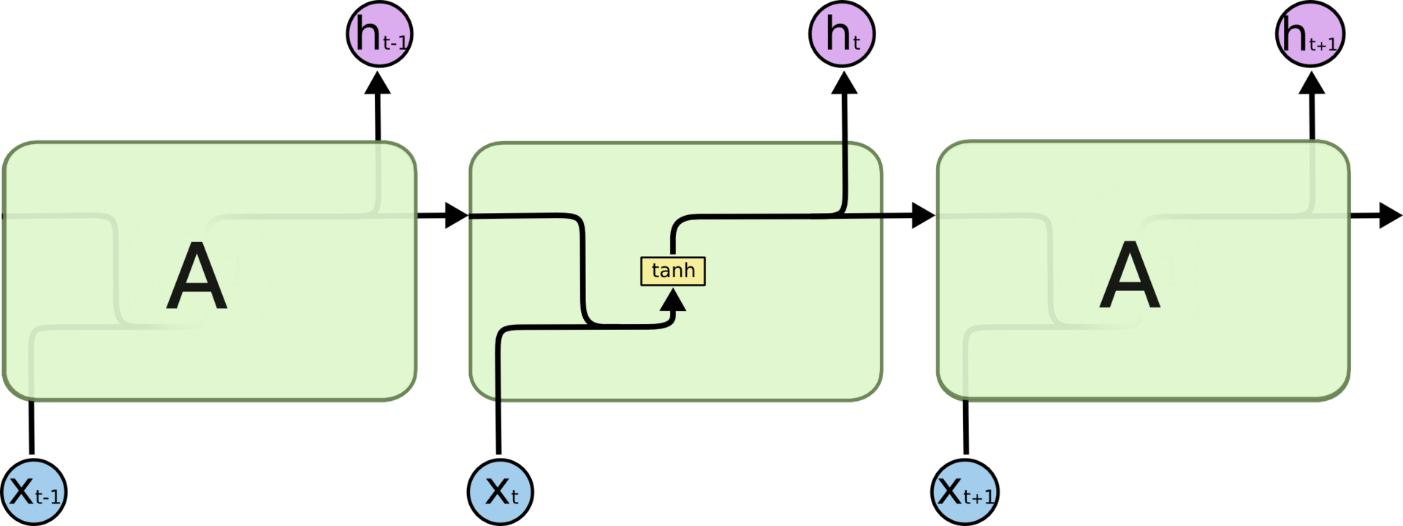


Рисунок 1.4 – архитектура рекуррентной нейронной сети

Свёрточные нейронные сети (CNN) – это ещё один тип нейронной сети, который хорошо подходит для обработки изображений. Однако CNN также могут быть адаптированы для работы с временными рядами, такими как цены акций. CNN могут использоваться для выявления паттернов в данных о ценах акций и прогнозирования их будущих значений.



Рисунок 1.5 – архитектура свёрточной нейронной сети

Нейронные сети с долгой краткосрочной памятью (LSTM) – это разновидность рекуррентных нейронных сетей, которая может эффективно обрабатывать долгосрочные зависимости в данных. LSTM могут быть полезны для анализа долгосрочных тенденций в ценах акций.

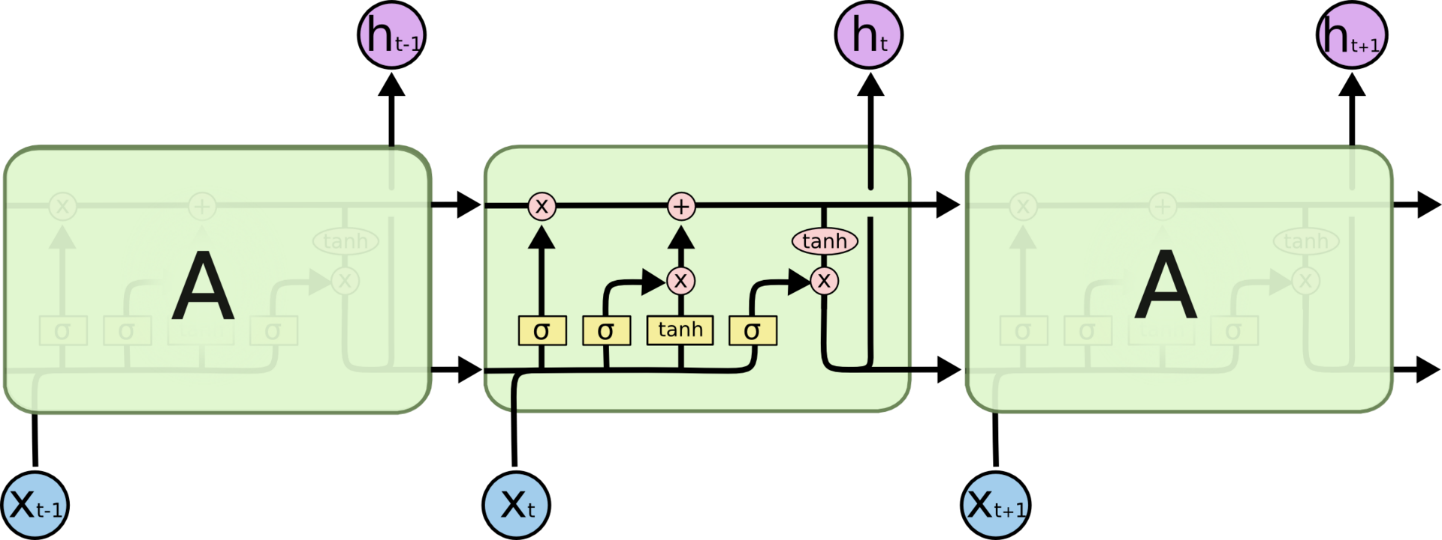


Рисунок 1.6 – архитектура сети с ячейками LSTM

Генеративно-состязательные сети (GAN) – это архитектура нейронных сетей, состоящая из двух подсетей: генератора и дискриминатора. Генератор создаёт новые данные, а дискриминатор пытается отличить эти данные от реальных. GAN могут быть использованы для создания синтетических данных о ценах акций, которые могут быть использованы для обучения других моделей.

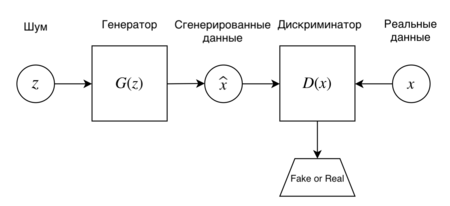


Рисунок 1.7 – архитектура генеративно-состязательной нейронной сети

Выбор конкретной архитектуры зависит от данных, целей и ресурсов. Если используется большой объём данных, то нейронные сети с ячейками LSTM или генеративно-состязательные сети могут дать наилучшие результаты. Иначе же более подходящими могут быть рекуррентные или свёрточные нейронные сети.

### Обзор методов обработки естественного языка (NLP)

Лемматизация.

Это метод приведения текста к его словарной форме (лемме). Лемма – это базовая форма слова, которая используется в словарях. Например, для глагола «бегу» леммой будет «бежать», а для существительного «столы» – «стол». В отличие от стемминга (обрезания суффикса слов), лемматизация является более лояльным методом обработки слов, так как она возвращает первоначальную форму слова.

Этот метод прост и эффективен, но он имеет некоторые недостатки. Во-первых, он может не работать с некоторыми словами, которые не имеют лемм в словаре. Во-вторых, он может изменить значение слова, если оно имеет несколько лемм.

Например, слово «бегут» может иметь две леммы: «бежать» и «бег». Если программа использует первую лемму, она изменит значение слова с множественного числа на единственное. Это может привести к ошибкам в обработке текста.

Чтобы избежать этих проблем, можно использовать более сложные методы лемматизации. Эти методы учитывают контекст слова и могут выбирать правильную лемму даже для слов, которых нет в словаре.

Для обработки русских слов существует множество пакетов, которые выполняют данную задачу: Pymystem, Pymorphy, Томита-парсер, Яндекс-спеллер и другие.

TF-IDF.

TF-IDF (Term Frequency – Inverse Document Frequency) – это статистическая мера, которая используется для оценки важности слова в контексте документа или корпуса документов. Она учитывает как частоту появления слова в документе, так и его редкость во всём корпусе.

Мера TF-IDF рассчитывается по следующей формуле:

где:

– частота слова в документе ;

– обратная частота документа слова

Частота слова () определяется как отношение количества вхождений слова к общему количеству слов в документе. Чем чаще слово встречается в документе, тем выше его важность.

Обратная частота документа () определяется как логарифм отношения общего количества документов к количеству документов, содержащих данное слово. Чем реже слово встречается во всём корпусе, тем выше его значимость.

Таким образом, TF-IDF представляет собой произведение этих двух показателей. Это позволяет учитывать как частое появление слова в одном документе, так и его редкую встречаемость во всём корпусе документов.

TF-IDF широко используется в различных задачах обработки естественного языка, таких как классификация текста, извлечение информации и машинный перевод. Например, в задаче классификации текста TF-IDF может быть использован для определения ключевых слов, которые характеризуют каждый класс.

Анализ тональности.

Сентимент-модели – это модели, которые используются для анализа тональности текста. Они позволяют определить, является ли текст положительным, отрицательным или нейтральным.

Явным примером для определения тональности текста является модель BERT (Bidirectional Encoder Representations from Transformers) – это модель машинного обучения, которая используется для обработки естественного языка. Она была разработана компанией Google в 2018 году и с тех пор стала одной из самых популярных моделей в этой области.

BERT основана на архитектуре трансформеров, которые представляют собой тип нейронных сетей, способных обрабатывать последовательности данных. Трансформеры состоят из двух основных компонентов: кодировщика и декодировщика. Кодировщик преобразует входные данные в скрытые представления, а декодировщик использует эти представления для генерации выходных данных.

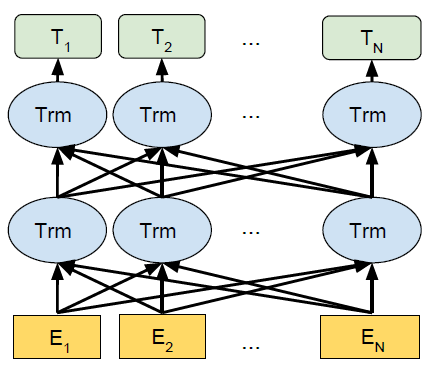


Рисунок 1.8 – архитектура модели BERT

В случае BERT кодировщик является двунаправленным, то есть он обрабатывает входные данные как слева направо, так и справа налево. Это позволяет модели учитывать контекст слов во входных данных и лучше понимать их смысл.

Для обучения BERT используются два метода:

Маскированный язык моделирования (MLM): В этом методе часть слов во входных данных маскируется, и модель должна предсказать эти слова на основе контекста.

Предсказание следующего предложения (NSP): В этом методе модель обучается предсказывать, какое предложение следует за данным предложением во входном тексте.

Эти методы позволяют модели научиться понимать контекст слов и предложений, что делает её полезной для различных задач обработки естественного языка, таких как классификация текста, ответы на вопросы и перевод.

### Обзор программных продуктов

Существует множество программных продуктов, которые можно использовать для анализа новостных данных и прогнозирования цен акций. Рассмотрим некоторые из них.

Python: является популярным языком программирования, который широко используется для машинного обучения и анализа данных. Также в нём существует набор библиотек Python, таких как pandas, numpy и scikit-learn, которые можно использовать для создания моделей прогнозирования цен акций.

R: является ещё одним популярным языком программирования для статистического анализа и машинного обучения. Существуют пакеты R, такие как tidyverse и caret, которые можно использовать для построения моделей прогнозирования цен акций.

Специализированные платформы: существуют также специализированные платформы, предназначенные для анализа финансовых данных, такие как Quandl, Alpha Vantage и Yahoo Finance. Эти платформы предоставляют доступ к финансовым данным, таким как цены акций, объёмы торгов и новостные данные.

Выбор конкретной модели, метода и программного продукта зависит от конкретных требований и ограничений задачи. Важно провести тщательное исследование и тестирование различных подходов, чтобы определить наиболее эффективный метод прогнозирования для конкретной ситуации.

## Постановка и формализация задачи на разработку

На рисунках 1 и 2 представлены постановка задачи в виде контекстной диаграммы и её декомпозиция, соответственно.

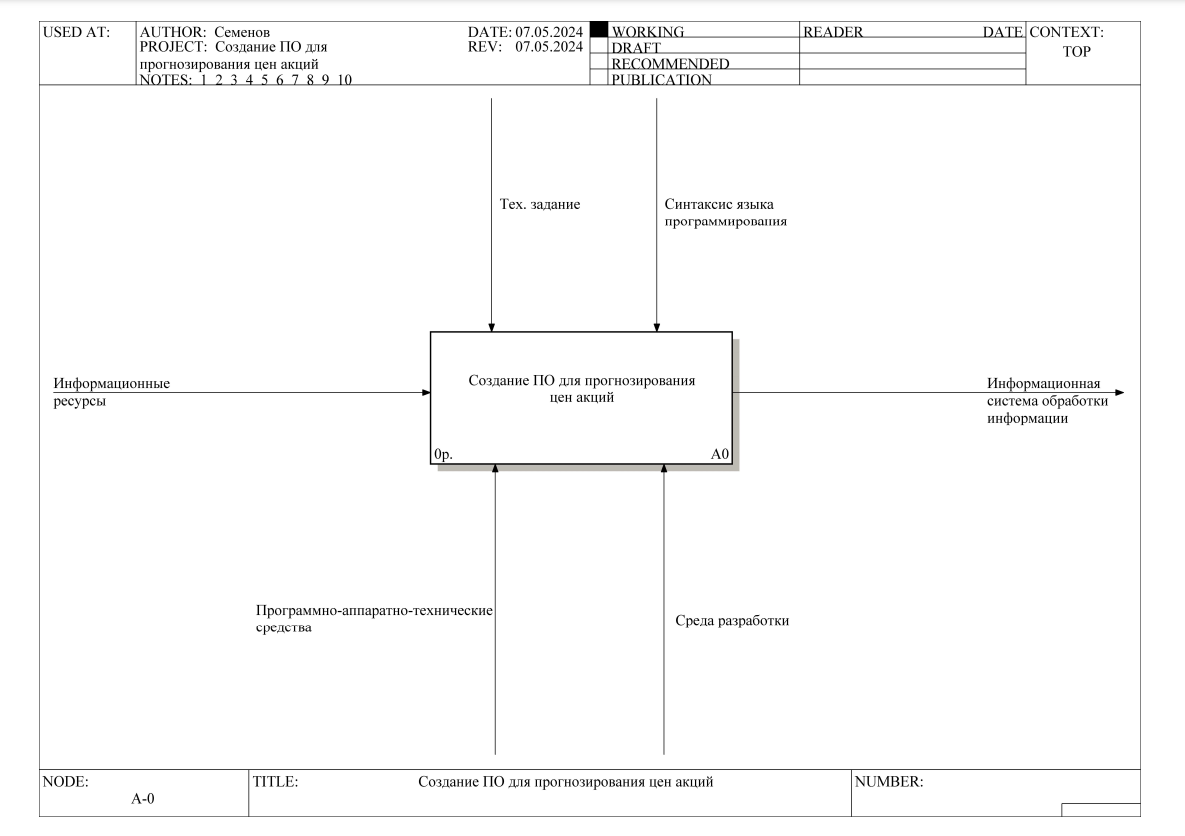


Рисунок 1.9 – Постановка задачи: контекстная диаграмма

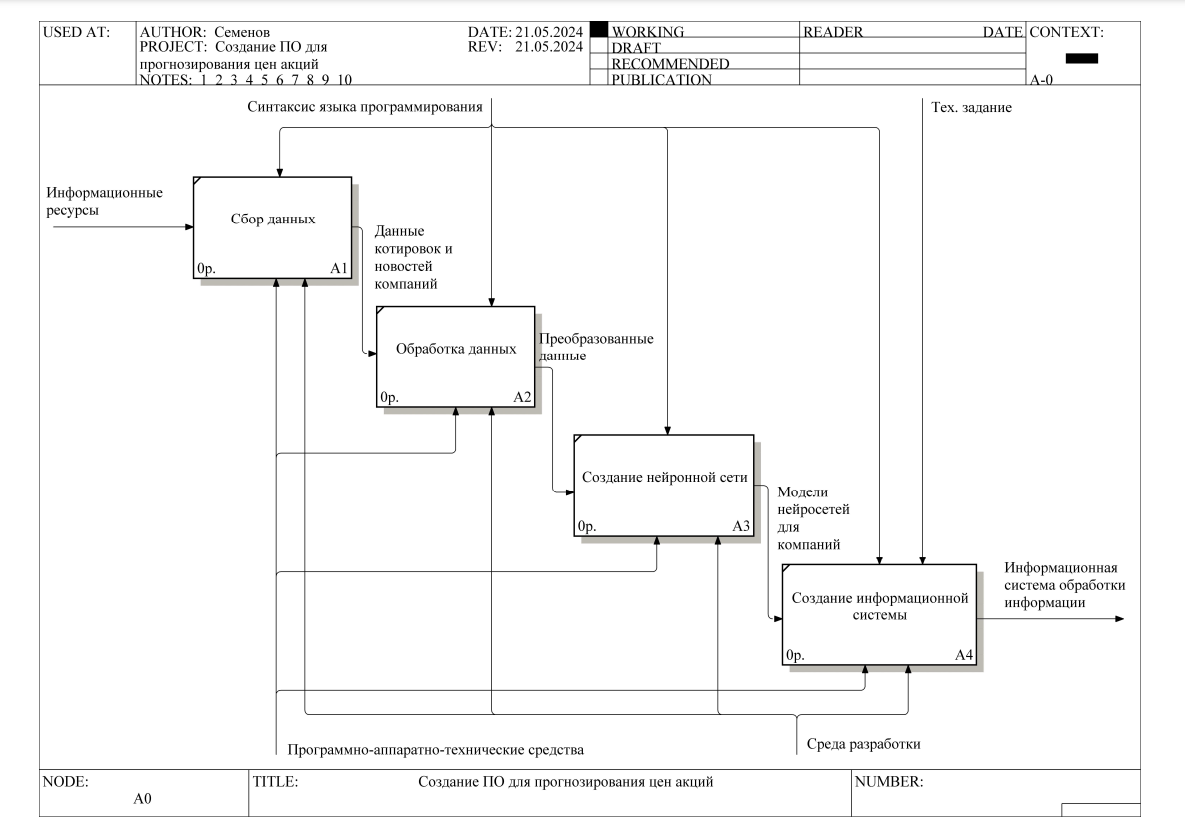


Рисунок 1.10 – Декомпозиция контекстной диаграммы

Цель проводимого исследования заключается в разработке ПО, способного анализировать воздействия информации, распространяемой в средствах массовой информации, на цены акций компаний на финансовых рынках. Реализация данной цели включает в себя решение целого ряда задач, направленных на полноценное исследование влияния информации в СМИ на финансовые рынки и компании.

Задачи:

* Провести сбор данных, получить новостные данные, а также котировки;
* Провести обработку текстовых данных новостей, привести в вид, понятный для модели нейронной сети;
* Провести анализ тональности новостных потоков для каждого дня, выявить ключевые слова позитивно и негативно настроенных потоков;
* Создать и обучить модель нейронной сети, способной предсказывать будущую цену акции, на основе прошлых данных котировок и новостного потока;
* Реализовать информационную систему, основанную на модели нейронной сети.

Таким образом, исследование данных задач позволит глубже понять механизмы взаимодействия информации, распространяемой в СМИ, с финансовыми рынками и компаниями.

# Проектная часть

Набор компаний для создания программного решения был выбран на основе индекса МосБиржи, а также некоторых популярных компаний, который не входили туда.

Таким образом был сформирован список из 49 российских компаний: Афк Система, Аэрофлот, Русагро, Алроса, Московский Кредитный Банк, Северсталь, Эн+, X5, Совкомфлот, Газпром, Глобалтранс, Норникель, Русгидро, Интер РАО, Лукойл, Ммк, Магнит, Мосэнерго, Мечел, Мтс, Нлмк, Новатэк, Озон, Фосагро, Пик, Полюс, Полиметал, Позитив, Роснефть, Ростелеком, Русал, Сбер, Селигдар, Сегежа, Самолет, Сургутнефтегаз, Татнефть, Тинькофф, Транснефть, Юнипро, Вконтакте, Втб, Яндекс, Фск Еэс ОАО, Qiwi, Тмк, Акрон, Мать и дитя, Ютэйр.

## Разработка методов и алгоритмов для решения поставленной задачи

Для решения поставленной задачи предлагается использовать алгоритм, представленный на рисунке ??? ниже.

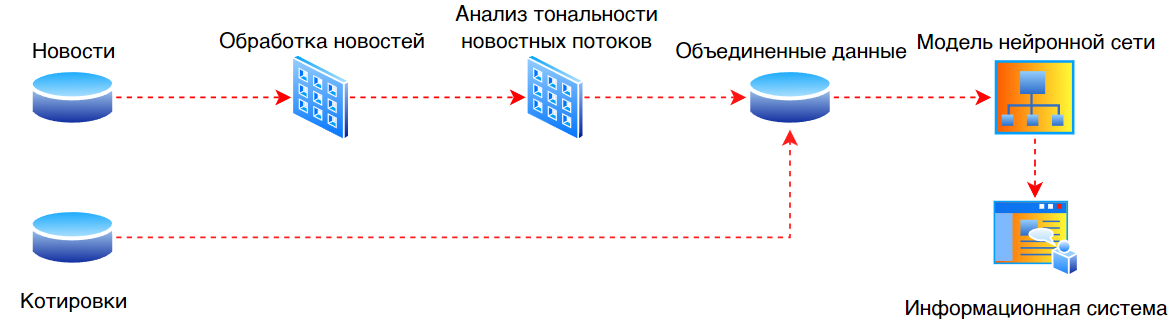


Рисунок 2.1 – схема алгоритма работы

В начале работы необходимо провести сбор данных новостей компаний. Для этого необходимо разработать программу, позволяющую автоматически собирать новостные публикации, обеспечивающую бесперебойную работу. В качестве информационного ресурса был выбран сайт finam.ru, так как в этом портале есть возможность собирать данные, разделенные по компаниям, и не имеющие временных ограничений по дате публикации.

### Обработка текстовых данных новостей

Обработка текста новостей является неотъемлемой частью работы, поскольку неструктурированные данные при анализе не позволят провести их правильную интерпретацию.

Для того чтобы сделать это, сначала необходимо привести весь текст к нижнему регистру. Затем убираются специальные символы, то есть знаки пунктуации, а также числа. Так как работа основана на корпусе русских новостей, то латинские символы будут только затруднять понимание текста, поэтому от них тоже необходимо избавиться.

В большинстве текстов, для их правильного понимания человеком, присутствуют различные местоимения, предлоги и прочие части речи, которые не несут в себе определенной смысловой нагрузки для машины, поэтому на основе собранных новостей необходимо создать список стоп-слов, в который будут включены такие слова. На основании полученного набора, из каждого текстового представления данных необходимо вычленить эти слова.

После того как были выполнены ранее описанные действия, каждое слово необходимо привести к его нормальному виду используя алгоритм лемматизации.

Также новости часто состоят из слов, которые по тем или иным причинам повторяются в тексте не один раз. Для правильного определения тональности поступающего текста, леммы повторов будут создавать только лишний шум, поэтому от них также необходимо избавляться.

Затем, используя меру IDF – обратную частоту документа, являющейся частью метрики TF-IDF необходимо вычленить те слова, которые слишком часто, либо слишком редко появляются в корпусе новостей, так как при интерпретации такие слова будут только создавать помехи. Мера IDF в свою очередь вычисляется в соответствии со следующей формулой:

где – количество документов в корпусе,

– количество документов из корпуса , в которых встречается слово .

После этого необходимо убрать те слова, показатель IDF которых находится за пределами верхнего и нижнего квантиля заданного -уровня (). Это позволит избавиться от слов, которые не несут полезной информации для анализа тональности новостей.

Далее, для того, чтоб подготовить данные для обучения нейросети, необходимо объединить новости одного дня в единый набор, и вычислить тональность полученного текста для каждого дня. Этот параметр будет определять, как прошёл день для компании на основе новостей – позитивно или же негативно.

Каждому дню присваивается своё значение тональности, а также цена котировки на этот день. Таким образом составляется дата сет, для того чтобы нейронную сеть было возможно обучить.

## Проектирование моделей базы данных: концептуальной, логической, физической

2.2.1 Концептуальная модель базы данных

2.2.2 Логическая модель базы данных

2.2.3 Физическая модель базы данных

## Проектирование программного обеспечения с обоснованием выбора методологии и среды проектирования и программирования

Здесь нужно вставить диаграмму связей программных файлов отвечающих за ту или иную задачу, и описать её

### Разработка парсера новостей

В начале программы необходимо импортировать необходимые библиотеки. Основные это Selenium, через которую будет происходить связь и сбор информации с новостного сайта, а также Pyscopg2, которая создаёт подключение к базе данных и позволяет записывать данные в заранее созданную таблицу.

Работа парсера происходит следующим образом:

1.Задается URL ссылка общего вида на страницу, с которой будет происходить сбор новостей.

Пример: https://www.finam.ru/publications/section/companies/date/

2.Вносится диапазон необходимых дат начала и конца, т.е. с какого числа по какое необходимо будет собирать информацию (01.01.2014 – 01.01.2024).

3.Парсер запускается и проходится по необходимому диапазону дат добавляя в конец ссылки из п.1 ту или иную последовательную дату (Рисунок ).

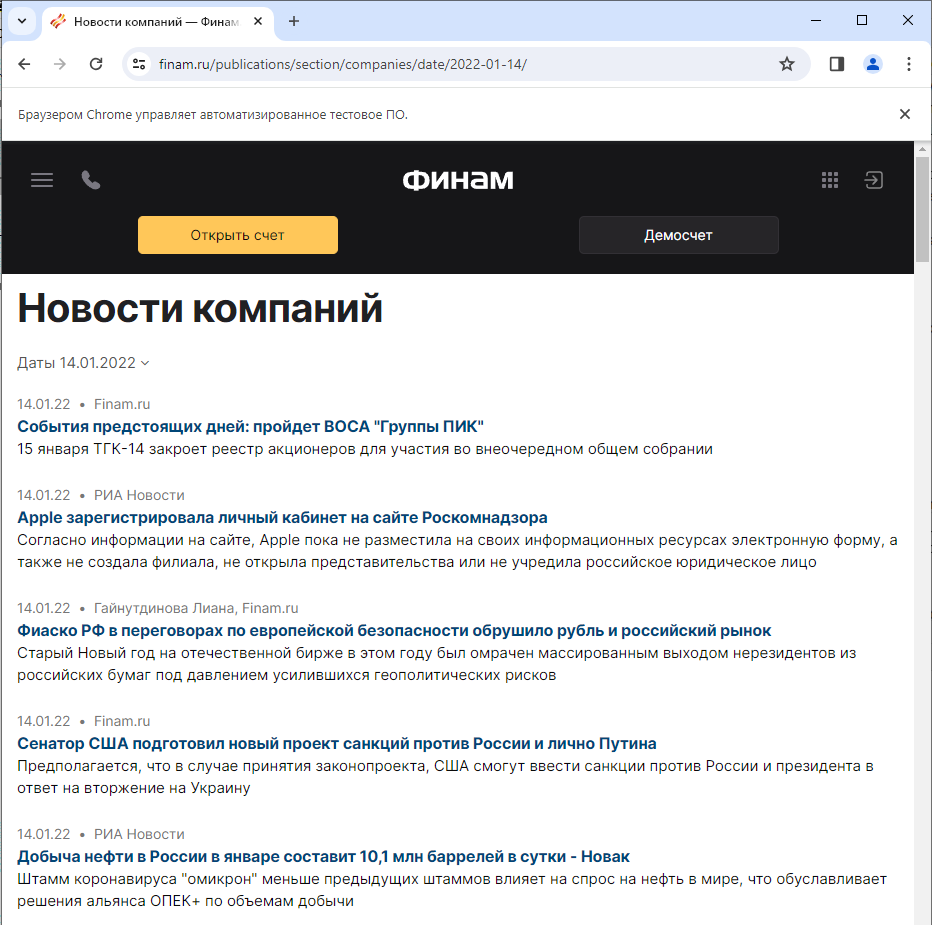


Рисунок 2.2 – Работа парсера новостей

4. При открытии страницы сайт ограничивает показ новостей количеством 20 штук, поэтому парсер сначала нажимает кнопку «Загрузить ещё» столько раз, сколько это возможно.

5. После того как все новости выбранного дня были выведены происходит запись ссылок на каждую новость в заранее определенный список.

6.После того как ссылки на новости в определенном дне были полностью собраны происходит повтор п.3 – п.5 до того момента пока не будет достигнута конечная дата из п.2.

7.Как только парсер прошел по всем датам из ранее заданного диапазона происходит непосредственный сбор информации из новостей: каждая ссылка из собранного списка последовательно открывается парсером и из новости отбираются все необходимые данные – ссылка на новость, текст заголовка, текст новости и дата выпуска новости (Рисунок ). Как только данные были получены они сразу записываются в базу данных.

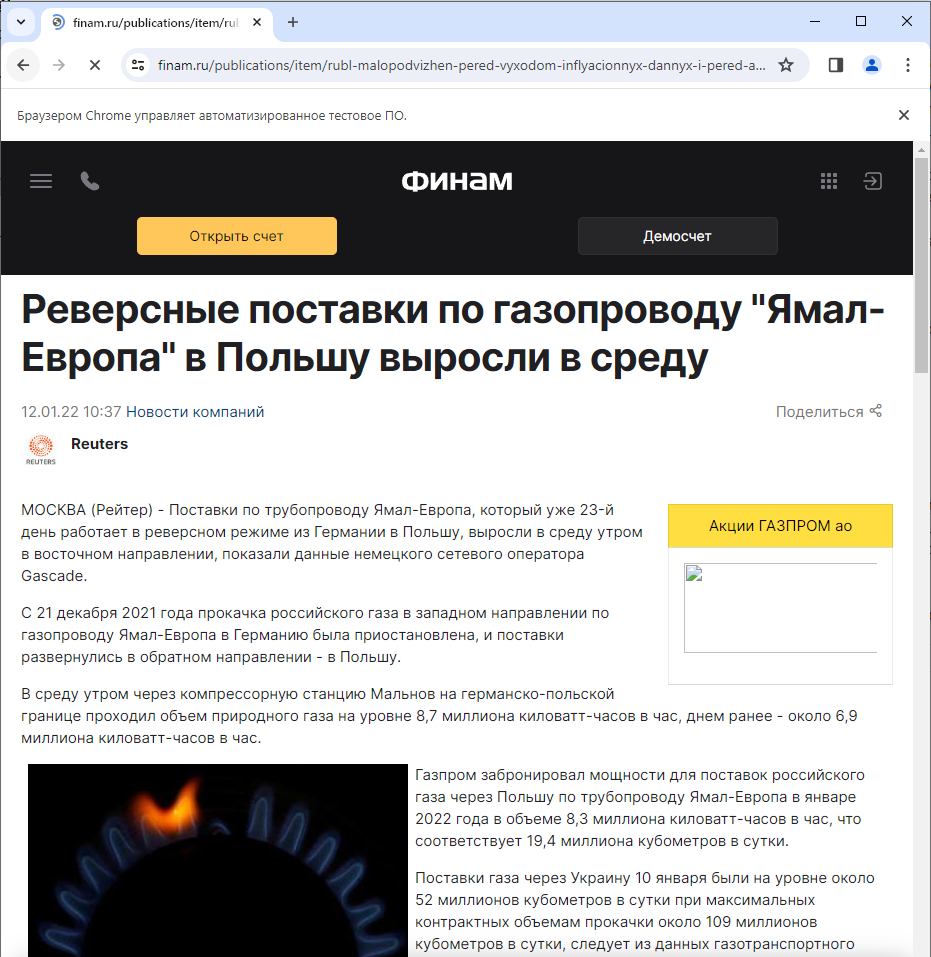


Рисунок 2.3 – Работа парсера новостей

8.Таким образом в базу данных за 2014 – 2024 гг. было собрано 97 572 новости.

Ниже приведена гистограмма распределения количества новостей по компаниям.

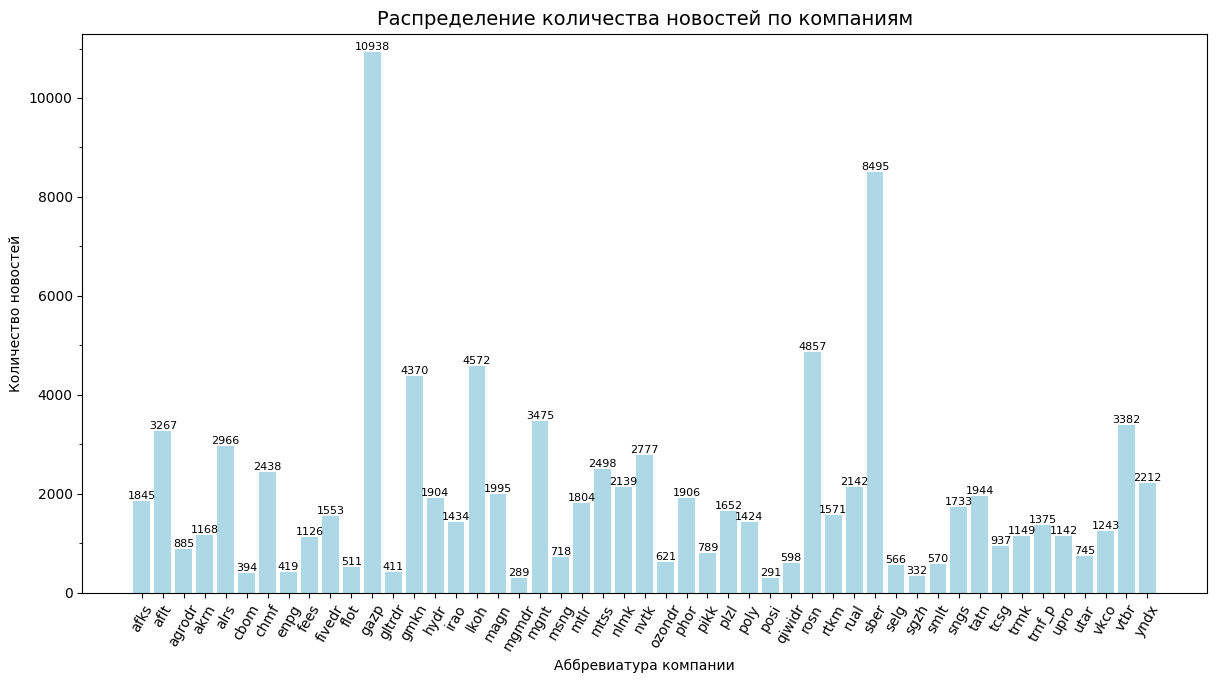


Рисунок 2.4 – Распределение количества новостей по компаниям

Исходя из представленного графика, самыми обсуждаемыми компаниями за последние 10 лет были: Газпром (10 938 записей), Сбербанк (8 495 записей) и Роснефть (4 857 записей).

### Разработка парсера котировок

Для того чтобы можно было определить влияние той или иной новости на цену акции компании, необходимо также иметь данные об изменениях цен на бумаги. Для этих целей создается парсер котировок цен акций компаний.

Принцип его работы заключается в сборе исторических данных каждой компании с сайта Investing.com. Работает он следующим образом:

1. Заранее создается список ссылок на страницы с историческими данными котировок для каждой компании.
2. При включении парсера последовательно открывается каждая ссылка из списка, созданного в п.1.
3. При открытии страницы показываются ежедневные данные лишь за последний месяц. Поэтому парсер автоматически открывает виджет в котором задается диапазон дат, и вносит туда необходимый диапазон (01.01.2014 – 01.01.2024).
4. После того как все данные подгрузятся и выведутся на страницу происходит их сбор: дата, цена закрытия, цена открытия, максимальная цена, минимальная цена, объем торговли и изменение цены.
5. После того как данные были собраны, в базе данных создается таблица с аббревиатурным названием компании, и в неё записываются собранная информация.
6. По окончании работы парсера в базе данных появляется 49 новых таблиц с котировками за 2014 – 2024 гг.

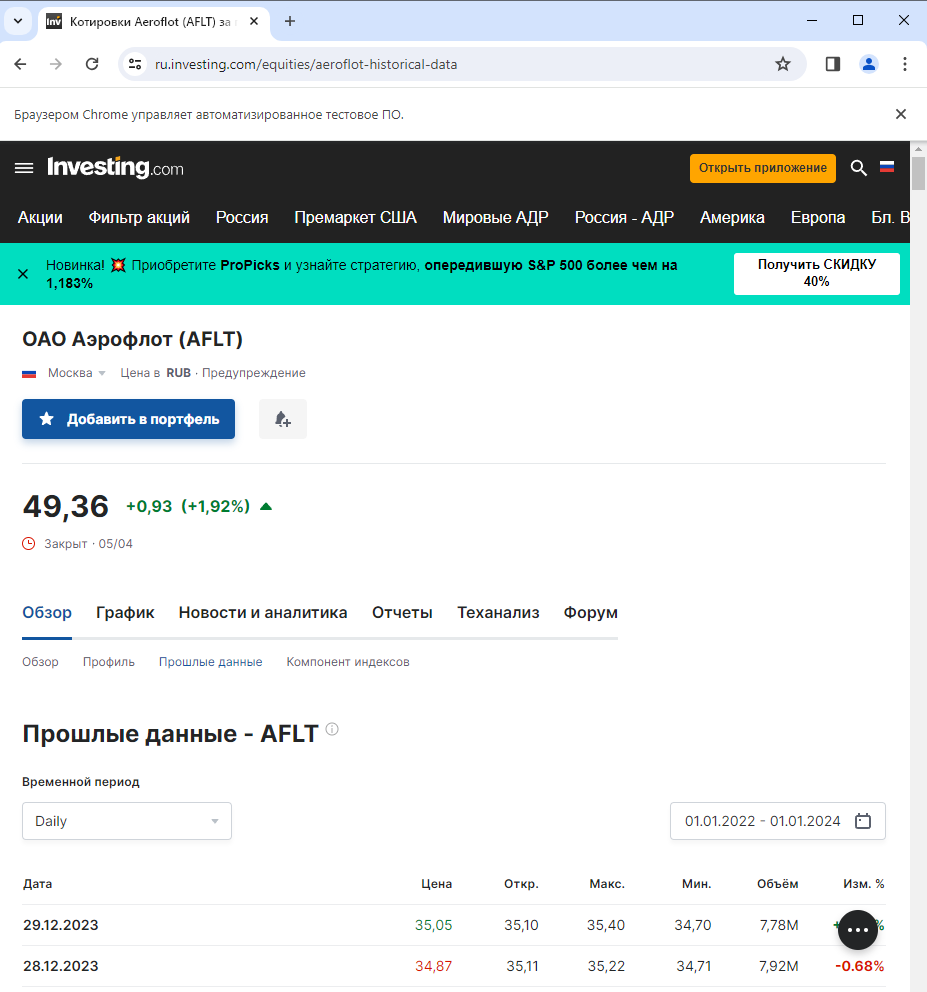


Рисунок 2.5 – Работа парсера котировок

Ниже приведена гистограмма распределения количества записей котировок для каждой компании.

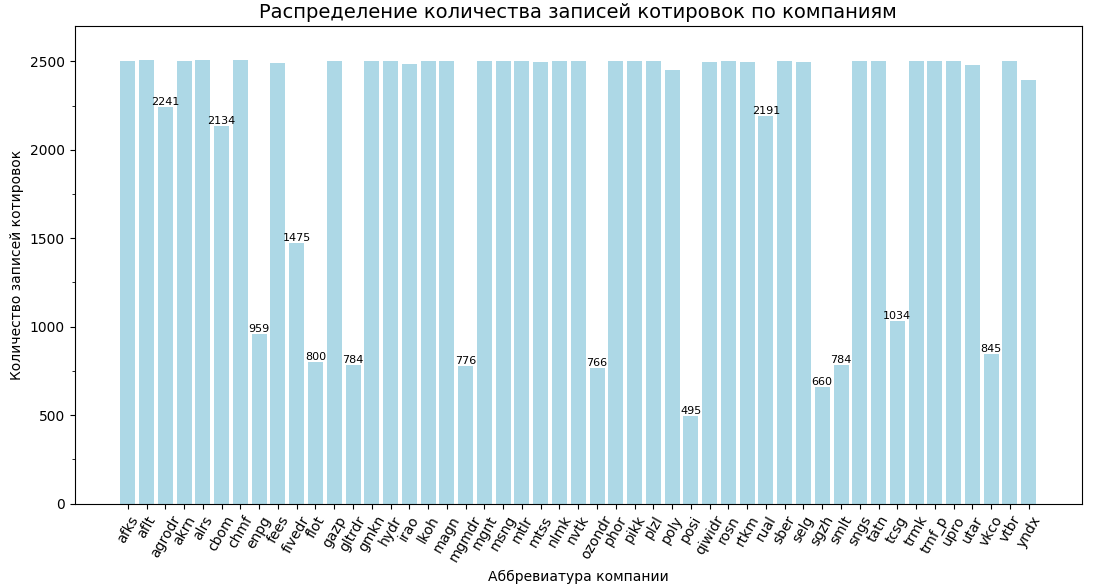


Рисунок 2.6 – Распределение количества записей котировок по компаниям

Как видно на графике, не все компании имеют исторические данные за данный период. Это может быть связано с тем, что компания не сразу начала выпускать акции. Таким образом, компаниями с наименьшим количеством записей являются: Позитив (495 записей, выпустили акции в декабре 2021 года), Сегежа (660 записей, выпустили акции в апреле 2021 года) и Озон (766 записей, выпустили акции в ноябре 2020 года).

### Обработка новостей

За обработку собранных новостей отвечает программа text\_processer\_init.py. Каждая новость поступает на вход программы и проходит ряд операций.

К одной из таких относится процесс лемматизации, при котором каждое слово заменяется на его первоначальную форму. За выполнение этого шага отвечает модуль Pymystem 3.1 от компании Яндекс. Преимуществом данного пакета является то, что он создан специально для корпуса русских слов, а также, что немаловажно, использует контекстные алгоритмы, позволяющие приводить слова к начальной форме учитывая контекст. То есть, если бы у нас был текст «Мама мыла раму» который мы бы хотели лемматизировать, то обычный алгоритм расшифровал бы его как «Мама мыло рама», а, в свою очередь, контекстный – «Мама мыть рама».

Рассмотрим алгоритм обработки текста на примере заголовка новости компании «АФК Система»: «Чистая прибыль АФК "Система" по US GAAP в 2013 году выросла на 11,1% - до $2 млрд».

1. Текст приводится к нижнему регистру: «чистая прибыль афк "система" по us gaap в 2013 году выросла на 11,1% - до $2 млрд»
2. Убирается вся пунктуация: «чистая прибыль афк система по us gaap в 2013 году выросла на 111 до 2 млрд»
3. Убираются все числовые символы: «чистая прибыль афк система по us gaap в году выросла на до млрд»
4. Убираются все латинские символы: «чистая прибыль афк система по в году выросла на до млрд»
5. Из слов выделяются их леммы: «чистый прибыль афк система по в год вырастать на до млрд»
6. Убираются стоп-слова из заранее созданного соответствующего списка stop-words.txt: «чистый прибыль афк система вырастать»
7. Далее убираются повторяющиеся слова.
8. Также убираются слова, которые излишне часто и крайне редко повторяются в корпусе: «чистый прибыль вырастать»
9. И наконец, используя метрику IDF убираются слова, которые выходят за пределы верхнего и нижнего квантилей 95% уровня.

В итоге, получаем обработанный текст – «чистый прибыль вырастать», первоначально имевший вид – «Чистая прибыль АФК "Система" по US GAAP в 2013 году выросла на 11,1% - до $2 млрд».

Таким образом из новости вычленяются самые полезные для её дальнейшего использования слова, по которым можно достоверно оценить её смысловой характер.

### Сентиментальный анализ текста

Применение сентиментального анализа текста является одним из ключевых аспектов работы. За проведение этой процедуры отвечает файл sentiment\_analyze.py.

При проведении данной операции используется нейросеть RuBERT-base-cased-sentiment, основанная на архитектуре трансформеров. Данная модель уже обучена на 351 797 русских текстах. Итогом работы данной нейросети является трехмерный вектор, компонентами которого являются следующие параметры: NEUTRAL, POSITIVE, NEGATIVE.

Рассмотрим алгоритм на примере обработанного текста компании АЛРОСА: «выручка группа включая прочий деятельность увеличиваться примерно прошлое составлять сообщаться материал алмазодобыча карат реализация запас нейтральный динамика алмазный квартал придерживаться прогноз число ювелирный качество средний технический».

После того как текст был обработан нейросетью, он имеет следующие вероятностные показатели принадлежности к сентимент-классам: Positive: 0.925, Neutral: 0.06, Negative: 0.014

Для того чтобы оценить итоговое сентиментальное настроение дня, все тексты обработанных новостей объединяются в единый текстовый поток. Затем, с помощью нейросети RuBERT, каждому дню присваивается значение вероятности принадлежности к классу позитивных, негативных или же нейтральных новостей.

Для того чтобы продемонстрировать пример работы данного алгоритма ниже приведены представления в виде облака слов для наборов наиболее встречающихся слов позитивных, негативных, а также нейтральных текстов для компании «Алроса».



Рисунок 2.7 – Word cloud для позитивных текстов



Рисунок 2.8 – Word cloud для негативных текстов



Рисунок 2.9 – Word cloud для нейтральных текстов

Как видно из данных рисунков, наиболее часто, в новостях с позитивным уклоном, встречаются слова «», «», «». Дни с преимущественно негативными настроениями чаще характеризуют слова: «», «», «».

С нейтральными текстами ситуация несколько сложнее, так как в них одинаково по количеству могут встречаться и слова из положительных наборов, и с отрицательных, а также те, по которым сложно охарактеризовать эмоциональный окрас.

### Создание модели нейронной сети

После того как все основные данные для создания и обучения нейронной сети были подготовлены, необходимо создать саму нейросеть.

Для решения поставленной задачи анализа текста отлично подойдет архитектура модели, включающая в себя ячейки LSTM. Такой выбор был сделан в связи с тем, что данный метод имеет возможность запоминать в долгосрочной перспективе данные, которые в неё вошли. Таким образом в модели будет учитываться контекст новостей.

Для начала подготовленные данные необходимо разделить на обучающую и тестовую выборки. Разделение данных производится по следующему принципу: 80% данных будут использоваться как тренировочные, а оставшиеся 20% для теста.

Для того чтобы сформировать выборки в программе существует специальная функция CreateDataset, которая предназначена для подготовки данных для обучения модели прогнозирования временных рядов. Эта функция принимает набор данных, целевую переменную и необязательный параметр обратного просмотра, который определяет количество предыдущих временных шагов, которые должна учитывать модель. Функция выполняет итерацию по набору данных, генерируя пары входных последовательностей и соответствующие им целевые значения.

Для каждой итерации она извлекает из набора данных окно временных шагов обратного просмотра и формирует последовательность из этих значений в список обучающих данных. Соответствующее целевое значение для последнего временного шага в этом окне добавляется в список проверочных данных. Наконец, функция возвращает два массива, которые можно использовать для обучения модели. Каждая строка в списке обучающих данных содержит последовательность предыдущих наблюдений, а каждый соответствующий элемент в списке проверочных данных представляет целевое значение, которое необходимо предсказать в данный момент.

Данный алгоритм наглядно описан на рисунке ниже. Здесь параметр n описывает параметр обратного просмотра, описанный выше.

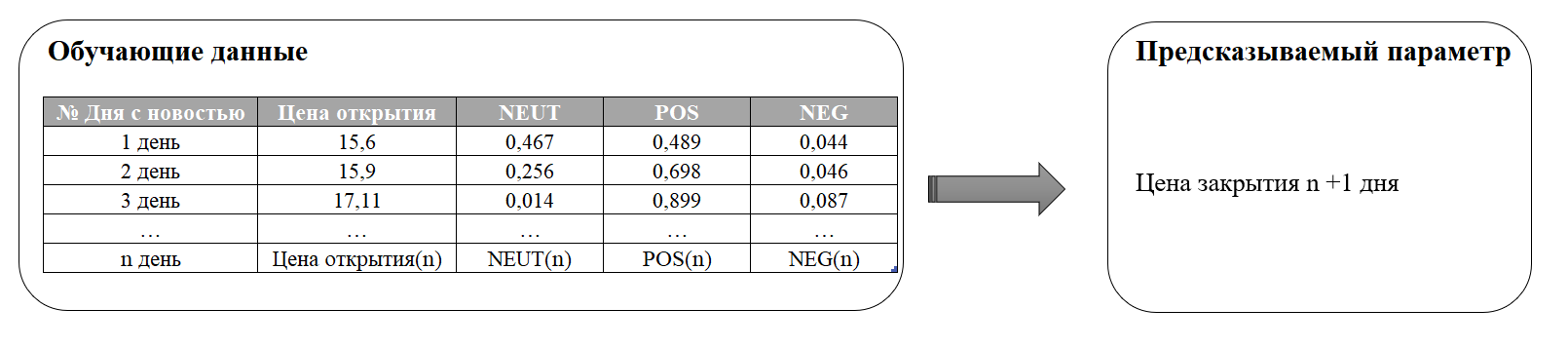


Рисунок 2.10 – Алгоритм предсказания параметра

Таким образом обучающий набор данных в общем случае представляет из себя вектор, состоящий из n записей, каждая из которых в свою очередь состоит из четырех компонент, которые характеризуют следующие показатели: «цена открытия», «показатель позитивного настроя», «показатель негативного настроя» и «показатель нейтрального настроя». На основании этих данных модель нейронной сети должна предсказывать «цену закрытия», которая и является целевой функцией.

Архитектура нейронной сети.

Составление нейросети проводилось с помощью пакета Keras, широко используемого в машинном обучении.

В общем виде архитектура созданной модели состоит из следующих слоев, приведенных на рисунке.

Набор слоев состоит из последовательно идущих трех слоев LSTM и Dropout, а завершается слоем Dense и Activation. Каждый слой LSTM состоит из 100 нейронов, а каждый слой Dropout имеет параметр, который определяет долю исключаемых нейронов, равный 0.6.

Таким образом, блоки LSTM принимают входные данные в виде последовательности, передают данные в слой Dropout, который используется для предотвращения переобучения модели путём случайного исключения некоторых нейронов во время обучения. После того как данные прошли обработку вышеуказанными слоями, они поступают в слой Dense, который преобразует выходные данные последнего LSTM блока в одномерный вектор с помощью функции активации linear. Для вывода конечного результата используется слой Activation, в котором применяется функция активации tanh, которая ограничивает значения выходного вектора между -1 и 1.

Также слой Dense включает в себя параметр регуляризации, который используется для ограничения весов модели, чтобы предотвратить их чрезмерное увеличение и улучшить обобщающую способность модели. В данном случае используется L1-регуляризация с коэффициентом 0.00001.

При компиляции модели используется функция потерь MSE (Mean squared error или же средняя квадратичная ошибка), оптимизатор Adam, а также дополнительно отслеживается параметр MAE (mean absolute error или же средняя абсолютная ошибка).

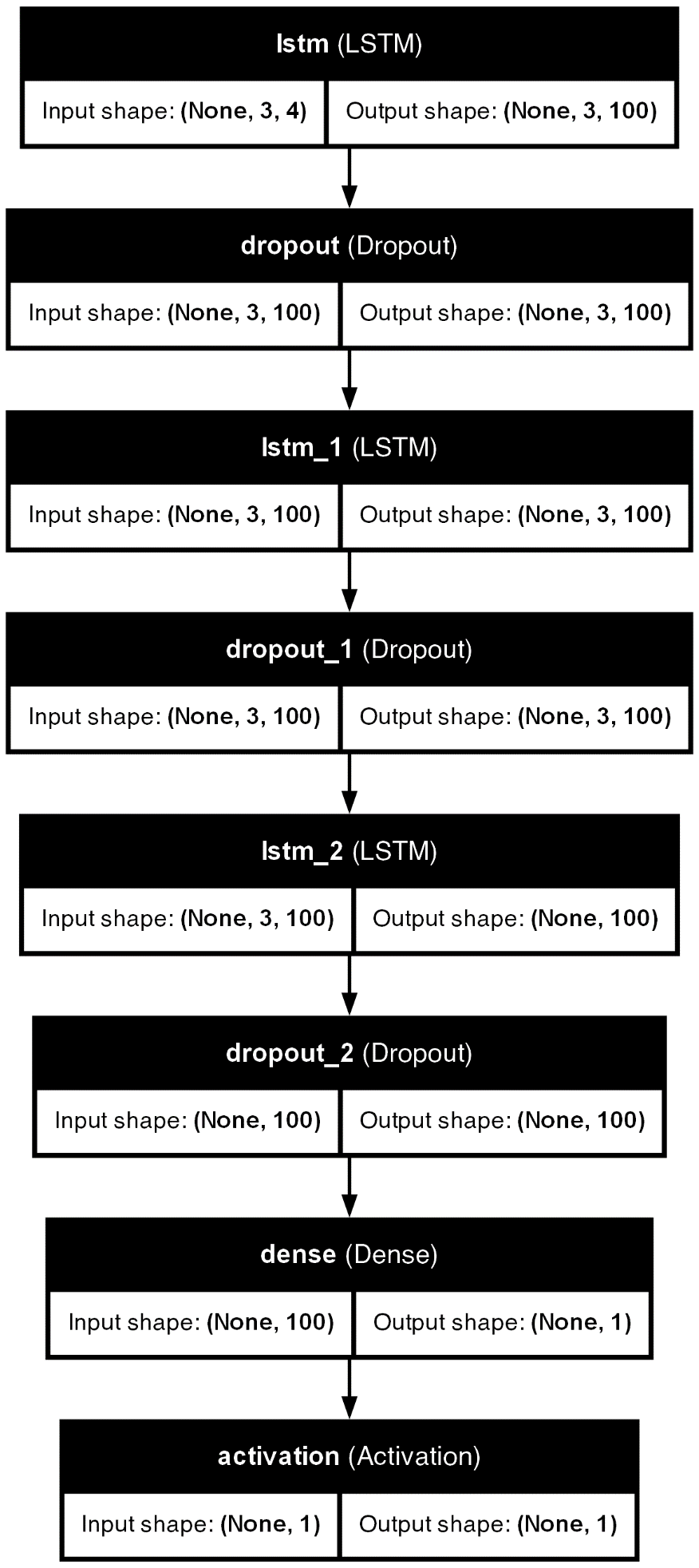


Рисунок 2.11 – архитектура нейросети

2.3.6 Создание информационной системы

## Вычислительные эксперименты и анализ их результатов

Рассмотрим процесс вычислительного эксперимента на примере компании «Алроса».

В ходе обучения модели по эпохам, как было сказано ранее, происходило отслеживание метрик MAE – Mean Squared Error (Средняя квадратичная ошибка) и MAPE – Mean Absolute Percentage Error (Средняя абсолютная процентная ошибка).

На рисунке ??? представлены данные показатели в зависимости от эпохи обучения.

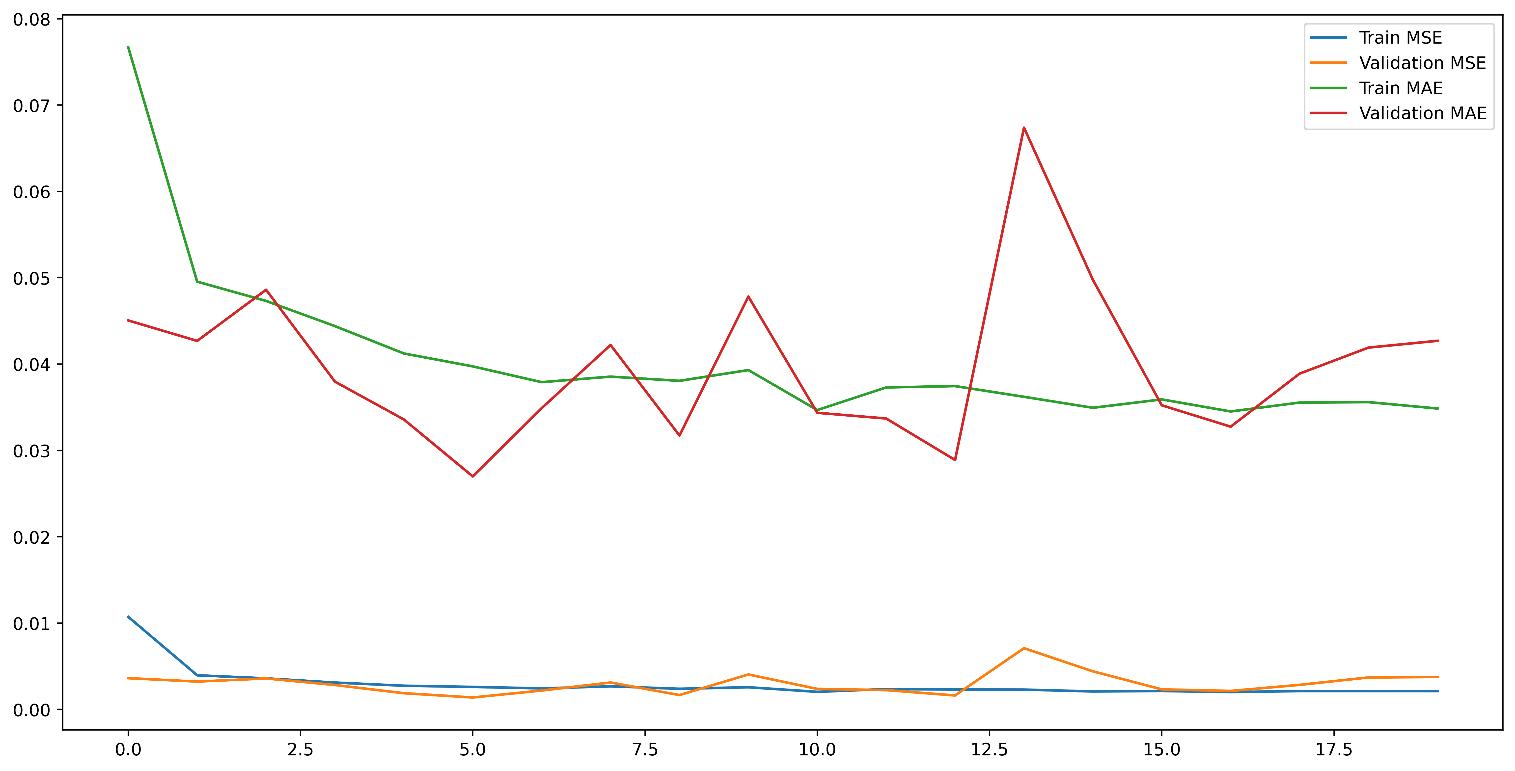


Рисунок 2.12 – Изменение наблюдаемых метрик по эпохам обучения

Как видно из графика, данные параметры последовательно уменьшаются до шестой эпохи, после чего их проверочные значения начинают меняться хаотичным образом, что говорит о том, что модель начала переобучаться и слишком сильно подстраиваться под тренировочные значения, выводя неверный результат при проверке. Таким образом для этой компании сохраняется и в дальнейшем будет применяться модель, соответствующая шестой эпохе обучения.

После того как наилучшая модель была выявлена и сохранена, необходимо провести её тестирование и замер показателей, описывающих её метрик.

График результата тестирования предсказания цены представлен на рисунке ??? ниже. Голубой и зеленой линией отмечены цены акций, а оранжевой и красной – цены, спрогнозированные моделью.



Рисунок 2.13 – Результат предсказания цен акций компании «Алроса»

В общем случае полученные данные имеют вид:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Date | Price | Open | Neutral | Positive | Negative | Predict |
| 2014-02-14 | 36,20 | 36,01 | 0,692 | 0,261 | 0,047 | 36,80 |
| 2014-02-17 | 36,00 | 36,20 | 0,822 | 0,147 | 0,031 | 35,18 |
| 2014-02-19 | 36,54 | 35,51 | 0,181 | 0,068 | 0,751 | 36,10 |
| 2014-02-26 | 37,25 | 36,82 | 0,821 | 0,147 | 0,031 | 37,74 |
| 2014-03-03 | 35,52 | 35,52 | 0,815 | 0,149 | 0,035 | 39,18 |
| … | … | … | … | … | … | … |

Метрика MSE.

Данная метрика рассчитывается по следующей формуле:

где – количество наблюдений;

– фактическое значение зависимой переменной;

– предсказанное моделью значение зависимой переменной.

По результатам наблюдений, метрика MSE на обучающем наборе выдала значение 12,673, а на тестовом – 56,732. Так как данный параметр направлен на отражение влияния именно больших ошибок, то по результатам предсказания на обучающем наборе (12,673) при минимальном и максимальном значении ~40 и ~140 пунктов соответственно, данный показатель показывает себя достаточно хорошо, чего нельзя сказать про прогнозирование на тестовом наборе. Такой результат (56,732) учитывает помехи, основанные на крайней нестабильности российского рынка в целом, которые происходили в феврале 2022 года.

Метрика RMSE.

Данная метрика вычисляется по следующей формуле:

Параметр RMSE по своей сути вычисляется как корень из среднеквадратичной ошибки, то есть зависит от него напрямую. Таким образом для данной модели были получены следующие характеристики метрики: на обучающих данных RMSE = 3,559, на тестовых RMSE = 7,532.

Метрика MAE.

Средняя абсолютная ошибка (Mean Absolute Error) рассчитывается как среднее значение модуля разности предсказанного и фактического значений. Данный параметр, в отличие от двух предыдущих является линейной оценкой, то есть все ошибки предсказаний в среднем взвешиваются одинаково. Параметр имеет следующую формулу расчёта:

Чем меньше значение показывает MAE, тем лучше модель предсказывает фактические значения. Это означает, что модель имеет более высокую точность.

Преимущества метрики MAE включают простоту понимания и интерпретации, а также устойчивость к выбросам. Однако она не учитывает направление ошибки и может быть чувствительной к большим отклонениям.

При расчёте данного параметра на модели для компании «Алроса» были получены следующие результаты: для тренировочных данных: MAE = 2,679, для тестовых данных: MAE = 5,285.

Метрика MAPE.

Средняя абсолютная процентная ошибка (Mean Absolute Percentage Error) представляет собой среднее абсолютное процентное отклонение между фактическими значениями и предсказанными. Рассчитывается согласно формуле:

Данная метрика даёт более наглядное представление о модели и означает на сколько процентов в среднем модель ошибается в своих прогнозах.

По результатам обучения для рассматриваемой нейросети были получены следующие значения метрик: на тренировочных данных MAPE = 3,525%, на тестовых данных MAPE = 5,346%. Дословно расшифровывается так: в среднем модель при предсказании ошибается на 5,346%.

## Возможные улучшения программы

Здесь тоже надо сказать что-то

# Заключение

В ходе проведенной работы был успешно собран, обработан значительный объем данных и разработан программный продукт на основе новостных статей и котировок ценных бумаг для 49 компаний. Тексты новостей были очищены от стоп-слов, пунктуации и прошли лемматизацию, что позволило стандартизировать данные и подготовить их к анализу и построению модели.

Был проведен анализ влияния информации в СМИ на риск и доходность инвестирования в ценные бумаги, что имеет большое практическое значение для инвесторов и трейдеров. Это позволит им принимать более информированные решения и управлять инвестиционным портфелем с учетом внешних факторов. Был выполнен дополнительный анализ, включая статистическую обработку данных, построение моделей и проверку гипотез.

Исходя из проделанной работы, был сделан вывод, что выбранная тема исследования обладает потенциалом для дальнейшего исследования и построения моделей, полезных для принятия инвестиционных решений. Был проведен более глубокий анализ связей между новостной информацией и изменениями в рыночных показателях, а также разработаны прогностические модели для прогнозирования рисков и доходности инвестирования.

Благодаря проведенному анализу данных и предварительной обработке информации, исследование достигло поставленных целей и может служить основой для дальнейших исследований в области анализа влияния информации в СМИ на инвестиционные решения.

# Список литературы

1. Chollet François et al. Keras. –– https://github.com/fchollet/ keras. –– 2015.
2. Loper E. Bird S. NLTK: The Natural Language Toolkit // Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1. –– ETMTNLP ’02. –– Stroudsburg, PA, USA : Association for Computational Linguistics, 2002. –– P. 63–70.
3. M. Korobov. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. –– Springer International Publishing, 2015. –– Vol. 542 of Communications in Computer and Information Science. –– P. 320–332.
4. S. Hochreiter, J. Schmidhuber. Long short-term memory // Neural computation. –– 1997. –– Vol. 9, no. 8. –– P. 1735–1780.
5. Srivastava N. Hinton G. Krizhevsky A., I. Sutskever. Dropout: a simple way to prevent neural networks from overfitting. // Journal of Machine Learning Research. –– 2014. –– Vol. 15, no. 1. –– P. 1929–1958.
6. Sumathy K. L. Chidambaram M. Text Mining: Concepts, Applications, Tools and Issues – An Overview // International Journal of Computer Applications. –– 2013. –– October. –– Vol. 80, no. 4. –– P. 29–32.
7. V.P. Romanov. Information technology modeling of financial markets - (Applied Information Technology) / Informatsionnye tekhnologii modelirovaniya finansovykh rynkov - (”Prikladnye informatsionnye tekhnologii”). –– Finansy i statistika, 2010. –– ISBN: 5279034444.
8. Cutler D. M., Poterba J. M., Summers L. H. (1988) What moves stock prices? // Massachusetts Institute of Technology. N 487. URL: https://dspace.mit.edu / bitstream /handle / 1721.1 / 64351 / whatmovesstockpr00cutl.pdf.
9. Berinsky A. (2017) Rumors and HealthCare Reform: Experiments in Political Misinformation // The British Journal of Political Science. Vol. 47, No 2. P. 241–262.
10. Baum M. A., Groeling T. (2008) Shot by the messenger: Partisan Cues and Public Opinion Regarding National Security and War // Political Behavior. Vol. 31.2. P.157–186. URL; https://sites.hks.harvard.edu / fs / mbaum / documents / ShotByTheMessenger\_POBH.pdf.
11. De Bondt W. F. M., Thaler R. (1985) Does the Stock Market Overreact? // The Journal of Finance. Vol. XL, No 3. URL: https://onlinelibrary.wiley.com / doi / full / 10.1111 / j.1540–6261.1985.tb05004.x.
12. Eccles R. G., Newquist S. C., Schatz R. (2007) Reputation and Its Risks // Harvard Business Review. Febr. URL: https://hbr.org / 2007 / 02 / reputation-and-its-risks.
13. Flynn D. J., Nyhan B., Reifler J. (2016). The Nature and Origins of Misperceptions: Understanding false and Unsupported Beliefs about Politics // Advances in Pol. Psych. Vol. 38, No S1. P. 127–150.
14. Goel S., Anderson A., Hofman J. et al. (2015) The Structural Virality of Online Diffusion // Management Science. Vol. 62, No 1. P. 180–196.
15. How does Fake News affect corporate reputation? (2017) // Alva Group. URL: http://www.alva-group.com / en / fake-news-affect-corporate-reputation / .
16. Ikenberry D. L., Ramnath S. (1999) Underreaction to Self-Selected News Events: The Case of Stock Splits // Rice. URL: http://www.ruf.rice.edu / ~jgspaper / W\_Ikenberry\_underreactionv9.pdf.
17. Kolbel J. (2008) The effect of bad news on reputation and shareprice: An empirical survey // Swiss Federal Institute of Technology (ETH). URL: https://www.ethz.ch / content / dam / ethz / special-interest / mtec / chair-of-entrepreneurial-risks-dam / documents / dissertation / master%20thesis / Term\_Paper\_KoelbelJ\_0508-Grade6.pdf.
18. Media Tenor ( [s.a.]). URL: http://us.mediatenor.com / en / .
19. Ratkiewicz J., Conover M., Meiss M. et al. (2011) Detecting and tracking political abuse in social media // Proc. 5thInternational AAAI Conferenceon Weblogsand Social Media (ICWSM). P. 297–304.
20. Sornette D., Malevergne Y., Muzy J. F. (2002) Volatility Fingerprints of Large Shocks: Endogeneous Versus Exogeneous. URL: https://arxiv.org / pdf / cond-mat / 0204626.pdf.
21. Swire B., Berinsky A. J., Lewandowsky S. et al. (2017) Processing political misinformation: comprehending the Trump phenomenon // Royal Society Open Science. Vol. 4, N 3. URL: http://rsos.royalsocietypublishing.org / content / royopensci / 4 / 3 / 160802.full.pdf.
22. Tetlock P. C. (2007) Giving Content to Investor Sentiment: The Role of Media in the Stock Market // The Journal of Finance. Vol. LXII, No 3. URL: https://www.gsb.columbia.edu / faculty / ptetlock / papers / Tetlock\_JF\_07\_Giving\_Content\_to\_Investor\_Sentiment.pdf.
23. Tambuscio M., Ciampaglia G. L., Oliveira D. F. M. et al. (2017) Modeling the competition between the spread of hoaxes and fact checking / George Washington University. [Washington].

# Приложение А

(Список иллюстраций)

[Рисунок 1.1 – пример использования линейной регрессии 19](#_Toc167908517)

[Рисунок 1.2 – пример использования логистической регрессии 20](#_Toc167908518)

[Рисунок 1.3 – пример использования решающих деревьев 21](#_Toc167908519)

[Рисунок 1.4 – архитектура рекуррентной нейронной сети 22](#_Toc167908520)

[Рисунок 1.5 – архитектура свёрточной нейронной сети 23](#_Toc167908521)

[Рисунок 1.6 – архитектура сети с ячейками LSTM 23](#_Toc167908522)

[Рисунок 1.7 – архитектура генеративно-состязательной нейронной сети 24](#_Toc167908523)

[Рисунок 1.8 – архитектура модели BERT 27](#_Toc167908524)

[Рисунок 1.9 – Постановка задачи: контекстная диаграмма 29](#_Toc167908525)

[Рисунок 1.10 – Декомпозиция контекстной диаграммы 30](#_Toc167908526)

[Рисунок 2.1 – схема алгоритма работы 32](#_Toc167908527)

[Рисунок 2.2 – Работа парсера новостей 35](#_Toc167908528)

[Рисунок 2.3 – Работа парсера новостей 37](#_Toc167908529)

[Рисунок 2.4 – Распределение количества новостей по компаниям 38](#_Toc167908530)

[Рисунок 2.5 – Работа парсера котировок 39](#_Toc167908531)

[Рисунок 2.6 – Распределение количества записей котировок по компаниям 40](#_Toc167908532)

[Рисунок 2.7 – Word cloud для позитивных текстов 43](#_Toc167908533)

[Рисунок 2.8 – Word cloud для негативных текстов 43](#_Toc167908534)

[Рисунок 2.9 – Word cloud для нейтральных текстов 44](#_Toc167908535)

[Рисунок 2.10 – Алгоритм предсказания параметра 45](#_Toc167908536)

[Рисунок 2.11 – архитектура нейросети 48](#_Toc167908537)

[Рисунок 2.12 – Изменение наблюдаемых метрик по эпохам обучения 49](#_Toc167908538)

[Рисунок 2.13 – Результат предсказания цен акций компании «Алроса» 50](#_Toc167908539)