МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное образовательное учреждение

высшего образования

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ

ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Школа: ИЯТШ

Направление: Прикладная математика и информатика

Отделение: Отделение математики и математической физири

**Влияние новостной информации на котировки ценных бумаг**

Пояснительная записка к курсовой работе  
по дисциплине «Финансовая математика»

Выполнил:

Студент группы 0ВM41 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Д. С. Саматов

Проверил:

Профессор ОММФ ИЯТШ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А. A. Мицель

Томск 2024 г.

Table of Contents

[Определения, обозначения, сокращения, нормативные ссылки 3](#_Toc181554187)

[Введение 4](#_Toc181554188)

[Датасеты 6](#_Toc181554189)

[1. Теоретическая часть 7](#_Toc181554190)

[1.1. Многоуровневая модель для поддержки торговых решений 7](#_Toc181554191)

[1.2. Общая концепция многоуровневой архитектуры 7](#_Toc181554192)

[1.3. Модель долгосрочной и краткосрочной памяти (LSTM) 8](#_Toc181554193)

[**1.3.1.** **Структура модели LSTM** 8](#_Toc181554194)

[**1.3.2.** **Преимущества и недостатки LSTM** 10](#_Toc181554195)

[Преимущества модели LSTM: 10](#_Toc181554196)

[Недостатки модели LSTM: 10](#_Toc181554197)

[1.4. Языковая модель BERT 11](#_Toc181554198)

[**1.4.1.** **Классическая Bidirectional Encoder Representations from Transformers модель** 11](#_Toc181554199)

[**1.4.2.** **Financial Bidirectional Encoder Representations from Transformers модель** 13](#_Toc181554200)

[1.5. Методы корреляционного анализа 13](#_Toc181554201)

[1.6. Статистический тесты проверки значимости различий признаков 16](#_Toc181554202)

[**1.6.1.** **Одновыборочный критерий Стьюдента (t-критерий)** 16](#_Toc181554203)

[**1.6.2.** **Двухвыборочный критерий Стьюдента (t-критерий)** 17](#_Toc181554204)

[**1.6.3.** **U-критерий Манна-Уитни** 18](#_Toc181554205)

[1.7. Оценка значимости факторов методом перестановок (Permutation Importance) 19](#_Toc181554206)

[2. Практическая часть 21](#_Toc181554207)

[2.1. Использованные модели 21](#_Toc181554208)

[2.2. Сбор данных 22](#_Toc181554209)

[2.3. Обоснование реализуемости 23](#_Toc181554210)

[2.4. Предобработка данных 23](#_Toc181554211)

[2.5. Описание использования LSTM 23](#_Toc181554212)

[2.6. Подбор гиперпараметров модели 23](#_Toc181554213)

[2.7. Описание полученных результатов 23](#_Toc181554214)

# **Определения, обозначения, сокращения, нормативные ссылки**

**LSTM (Long Short-Term Memory) – разновидность рекуррентных нейронных сетей, способных эффективно обрабатывать последовательности данных с учетом долгосрочных зависимостей, что особенно актуально при работе с временными рядами и текстовой информацией.**

**BERT (Bidirectional Encoder Representations from Transformers) – модель обработки естественного языка, анализирующая контекст слова как слева, так и справа, обеспечивая глубокое понимание семантики предложений и повышая точность обработки текста.**

**FinBERT – специализированная финансовая адаптация модели BERT, дополнительно обученная на обширных финансовых текстах. Это позволяет ей точно анализировать тональность и содержание новостей в контексте финансовых рынков.**

**Helsinki-NLP/opus-mt-ru-en – нейросетевая модель машинного перевода, предназначенная для высокоточного преобразования текста с русского на английский язык.**

**MAE (Mean Absolute Error) – средняя абсолютная ошибка, измеряющая среднее значение абсолютных отклонений прогнозируемых значений от фактических, что позволяет оценить точность модели без учета направления ошибок.**

**MSE (Mean Squared Error) – среднеквадратичная ошибка, вычисляемая как среднее значение квадратов разностей между предсказанными и реальными значениями, акцентируя внимание на крупных отклонениях благодаря возведению в квадрат.**

**R² (Коэффициент детерминации) – статистическая метрика, отражающая долю дисперсии зависимой переменной, объясненную моделью, и характеризующая качество подгонки модели к данным.**

# **Введение**

**В эпоху стремительного развития технологий и глобализации финансовые рынки становятся все более динамичными и чувствительными к информационным потокам. Новостные события, распространяющиеся в режиме реального времени, оказывают мгновенное влияние на настроения инвесторов и, следовательно, на волатильность и направление движения цен на финансовые активы. В таких условиях эффективный анализ и интерпретация новостного фона становятся критически важными для принятия обоснованных торговых решений.**

**Современные методы машинного обучения и искусственного интеллекта предлагают уникальные возможности для автоматизации и повышения точности анализа новостей. Интеграция этих методов в торговые системы позволяет не только оперативно обрабатывать большие объемы информации, но и выявлять скрытые закономерности и тренды, недоступные при традиционном анализе.**

**Цель исследования – разработка интеллектуального аналитического инструмента в виде торгового робота, который, используя продвинутые алгоритмы анализа новостного контента, способен формировать рекомендации по покупке и продаже финансовых активов. Данный инструмент призван стать надежным помощником для трейдеров, обеспечивая глубокий и всесторонний анализ рыночных данных и новостных событий. Важно подчеркнуть, что торговый робот выполняет функции поддержки принятия решений и не осуществляет сделки автоматически, что сохраняет контроль за инвестиционными решениями за человеком.**

**Для достижения поставленной цели необходимо решить следующие задачи:**

1. **Собрать релевантные данные новостей и котировок, обеспечивая репрезентативность и полноту информации для анализа.**
2. **Провести очистку и подготовку исходных данных, включая обработку пропусков, нормализацию и преобразование текстовой информации для последующего анализа.**
3. **Cпроектировать и настроить эффективную архитектуру нейронной сети, способной анализировать новостной контент и предсказывать движение цен на основе выявленных закономерностей.**
4. **Обучить нейронную сеть на подготовленных данных, провести валидацию и тестирование модели для оценки ее производительности и точности прогнозов.**
5. **Использовать метрики качества, такие как MAE, MSE и R², для оценки эффективности модели и определения областей для дальнейшего улучшения.**

**Актуальность исследования обусловлена возрастающей сложностью финансовых рынков и необходимостью быстрого реагирования на изменяющийся информационный ландшафт. Традиционные методы анализа часто не справляются с объемом и скоростью поступающей информации, что создает спрос на интеллектуальные системы, способные автоматизировать и повысить эффективность процесса принятия решений.**

**В рамках данной работы предлагается инновационный подход к использованию анализа тональности и содержания новостных сообщений с применением моделей глубокого обучения. Это позволит более точно прогнозировать поведение финансовых активов, учитывая эмоциональную окраску и контекст новостей, что является важным фактором в динамике рыночных цен.**

**Практическая значимость исследования заключается в создании инструмента, который может быть интегрирован в торговые платформы и системы аналитики, предоставляя трейдерам и финансовым аналитикам конкурентное преимущество в условиях высокой волатильности рынков. Разработанный подход способствует повышению эффективности торговых операций и снижению рисков, связанных с человеческим фактором и субъективной интерпретацией новостей.**

**Обзор источников литературы**

**С развитием методов машинного и глубокого обучения возможности автоматизированного анализа текстовой информации значительно расширились. Современные исследования демонстрируют возрастающий интерес к разработке подходов для анализа новостей с целью прогнозирования поведения финансовых рынков. В данном разделе рассматриваются ключевые работы, посвященные анализу новостного контента, использованию релевантных датасетов и применению моделей глубокого обучения для предсказания динамики цен на активы.**

**В последние годы было проведено множество исследований, сфокусированных на изучении связи между новостями и финансовыми рынками.**

**1. В работе Боллена Дж., Мао Х. и Цзэна С. (2010) исследуется взаимосвязь между общественным настроением, оцененным по сообщениям в Twitter, и индексом Dow Jones Industrial Average (DJIA). Авторы анализируют возможность прогнозирования рыночных изменений на основе колебаний общественного настроения, используя методы, такие как причинно-следственный анализ Грейнджера и нелинейные модели, в частности, самоорганизующиеся нечеткие нейронные сети (SOFNN). Результаты исследования показывают, что настроение, определяемое с помощью Twitter, может служить индикатором для предсказания изменений на фондовом рынке. Особенно такие настроения, как «спокойствие» и «внимательность», обладают значительной прогностической ценностью для DJIA.**

**2. В исследовании Куликовой Т.Д., Ковтун Е.Ю. и Буденного С.А. (2023) анализируется влияние тематической категоризации новостей на точность прогнозирования цен акций. Учёные разделили новости на 20 тематических групп с использованием модели BERT и определяли тональность сообщений посредством модели FinBERT. Выяснилось, что учёт настроений внутри одной тематической группы улучшает точность предсказаний по сравнению с анализом всего новостного потока или его отсутствием. Эмоциональная насыщенность новостей играет ключевую роль, демонстрируя, что как положительные, так и отрицательные сообщения могут существенно влиять на точность прогнозов. Таким образом, тематическая классификация и анализ настроений оказываются эффективными инструментами для более глубокого понимания рыночных тенденций и поведенческих факторов, подтверждая возможность повышения точности прогнозирования при учёте специфических тематических категорий.**

**Эти исследования подтверждают, что сочетание анализа настроений с тематической категоризацией новостей может значительно повысить точность моделей прогнозирования цен акций и предоставить более глубокое понимание динамики рынка.**

# **Датасеты**

Для эффективного анализа финансовых новостей и обучения моделей прогнозирования цен на акции необходимы качественные и актуальные наборы данных. В ходе исследования были рассмотрены несколько доступных датасетов, каждый из которых имеет свои особенности и ограничения:

1. Sentiment Analysis for Financial News [3]: Этот датасет содержит анализ тональности 4837 статей, разделенных на три категории: положительные, нейтральные и отрицательные. Несмотря на ценность такого распределения, отсутствуют даты публикации, что затрудняет установление прямой связи между новостями и динамикой цен конкретных акций. Без временной метки сложно проводить временной анализ и связывать новостные события с реакцией рынка.
2. Daily Financial News for 6000+ Stocks [4]: Данный набор данных включает 843062 записи с указанием дат публикации и тикеров акций, что обеспечивает связь между новостями и конкретными ценными бумагами. Однако он охватывает информацию только по трем компаниям и не содержит меток тональности, необходимых для анализа настроений. Отсутствие оценки тональности ограничивает возможности использования датасета для моделей, учитывающих эмоциональную окраску новостей.
3. Financial News Sentiment Dataset (FiNeS) [5]: FiNeS предоставляет оценку тональности заголовков новостей по шкале от -1 до 1 и содержит данные о российских акциях. Это делает его особенно ценным для локальных исследований. Однако небольшой объем данных – всего 532 записи – может быть недостаточным для обучения сложных моделей глубокого обучения, требующих больших объемов данных для эффективного обучения и генерализации.
4. Twitter Financial News Sentiment [6]: Этот датасет включает 11932 записи, классифицирующие новости как "медвежьи", "бычьи" или "нейтральные". Хотя он предоставляет информацию о тональности, отсутствуют данные о дате публикации и связи с конкретными акциями. Это ограничивает возможности временного анализа и установления корреляций между настроениями в социальных медиа и движением цен на рынке.

Каждый из рассмотренных датасетов имеет свои ограничения, такие как отсутствие временных меток, недостаточный объем данных или отсутствие целевых меток для тональности. Эти факторы затрудняют проведение полноценного анализа и обучения моделей, способных точно прогнозировать движения цен на основе новостного контента. В связи с этим возникает необходимость разработки собственного парсера для сбора более детализированных и релевантных данных.

# **Теоретическая часть**

## **Многоуровневая модель для поддержки торговых решений**

**В рамках данного исследования разработан торговый робот, представляющий собой интеллектуальную полуавтоматическую систему для анализа финансовых данных и формирования рекомендаций по сделкам на фондовом рынке. Робот не осуществляет торговые операции самостоятельно; его основная функция – предоставление аналитической поддержки пользователю, помогая оценить текущую рыночную ситуацию и принимать информированные и взвешенные решения.**

**Концепция многоуровневой архитектуры модели основана на интеграции нескольких передовых технологий машинного обучения, включая:**

* **Языковая модель BERT (Bidirectional Encoder Representations from Transformers);**
* **Рекуррентная нейронная сеть LSTM (Long Short-Term Memory).**

## **Общая концепция многоуровневой архитектуры**

**Предложенная многоуровневая модель (см. рисунок 1.1) представляет собой последовательность взаимосвязанных этапов обработки данных, где каждая модель выполняет специализированную задачу. Такой подход позволяет эффективно использовать преимущества каждой технологии, обеспечивая более точные прогнозы благодаря комплексной обработке информации.**

**Архитектура модели включает следующие ключевые уровни:**

* 1. **Перевод новостных текстов: На первом этапе осуществляется перевод собранных новостных сообщений с русского на английский язык. Это необходимо для последующего использования англоязычных моделей анализа текста, таких как BERT, которые обладают высокой точностью в интерпретации тональности и смысловых нюансов на английском языке.**
  2. **Анализ тональности текста: Второй уровень предусматривает применение модели BERT для детального анализа тональности новостного контента. Здесь определяется эмоциональная окраска каждого сообщения – позитивная, негативная или нейтральная – что позволяет оценить общий эмоциональный фон и потенциальное влияние новостей на рыночные настроения.**
  3. **Прогнозирование ценовых трендов на основе временных рядов: Заключительный этап включает обработку временных данных с использованием модели LSTM. Эта рекуррентная нейронная сеть способна учитывать долгосрочные зависимости в данных и прогнозировать будущие изменения цен активов на основе исторических трендов и выявленных закономерностей.**

**Каждый этап модели функционирует как самостоятельный модуль, при этом выход одного уровня служит входом для следующего. Такая модульная структура позволяет учитывать различные аспекты информации – от семантического содержания новостей до исторической динамики цен – что в совокупности повышает точность прогнозирования и обоснованность торговых рекомендаций.**

**На рисунке 1.1 представлена визуализация структуры многоуровневой модели. Исходные данные, такие как новости и котировки, обозначены голубыми трапециевидными блоками. Этапы предобработки данных отмечены красными прямоугольниками, нейронные модели представлены оранжевыми блоками, а итоговый прогноз модели отображен зеленым трапециевидным блоком. Стрелки указывают направление потока информации, а функциональные блоки иллюстрируют последовательность и взаимодействие между различными этапами обработки данных.**

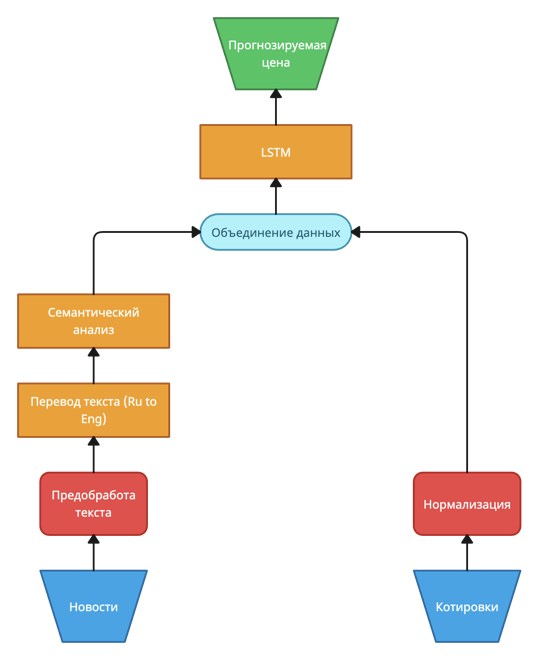


Рисунок 1.1. – Схема многоуровневой модели: стрелки указывают на направление потока информации, функциональные блоки отображают отдельные этапы обработки данных, а трапециевидные фигуры представляют исходные и итоговые данные.

## **Модель долгосрочной и краткосрочной памяти (LSTM)**

Модель Long Short-Term Memory (LSTM) [7] представляет собой разновидность рекуррентной нейронной сети, обладающую способностью эффективно работать с последовательными данными. Основное преимущество LSTM заключается в возможности сохранять долгосрочные зависимости, что делает её особенно востребованной для анализа временных рядов, обработки текстов и других типов данных, где последовательность имеет значение. Это свойство модели позволяет ей учитывать долгосрочные тенденции и выделять значимые паттерны, которые могут влиять на прогнозирование и анализ данных.

### **Структура модели LSTM**

Модель LSTM состоит из ряда специализированных компонентов, каждый из которых имеет несколько входов и выходов, а также так называемые «вентильные» механизмы (gates), регулирующие поток информации внутри сети. Эти вентильные механизмы позволяют сети избирательно забывать или сохранять информацию из предыдущих временных шагов. Основные элементы структуры LSTM включают следующие компоненты (см. Рисунок 1.2):

* Вентиль забывания (Forget Gate): На первом этапе LSTM решает, какую часть информации из предыдущего состояния нужно забыть. Вентиль принимает на вход текущее значение данных xtx\_txt​ и предыдущее скрытое состояние ​. На основе этих данных создается вектор забывания ​, который принимает значения от 0 до 1 для каждой ячейки памяти, указывая, какая информация должна быть удалена.
* Вентиль обновления (Input Gate): Этот вентиль регулирует, какая новая информация будет добавлена в память. Входными данными для него служат текущее значение ​ и предыдущее скрытое состояние ​. В результате работы вентиля генерируется вектор обновления ​, принимающий значения от 0 до 1, и кандидат  ​, который может быть добавлен в память.
* Обновление состояния памяти: На этом этапе модель определяет, какая информация будет обновлена в памяти. Процесс обновления зависит от векторов забывания ​, обновления ​ и кандидата , что позволяет сети сохранить только ту информацию, которая имеет значение для текущего шага. Итоговое состояние памяти обозначается как .
* Вентиль вывода (Output Gate): Этот вентиль отвечает за передачу информации в следующее скрытое состояние. Входными данными служат текущее значение ​ и предыдущее скрытое состояние . На выходе формируется вектор вывода ​, который указывает, какая информация будет передана в следующее скрытое состояние и определяет новое скрытое состояние .

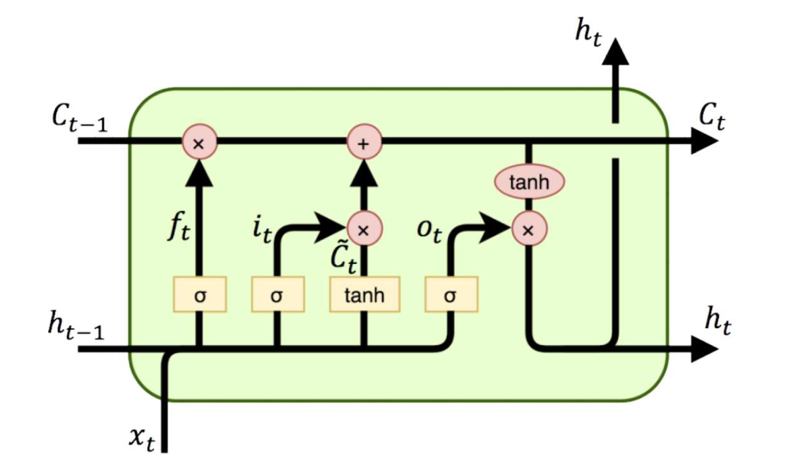


Рисунок 1.2. – Структура LSTM сети.

Каждый этап работы модели LSTM может быть выражен следующими формулами:

* 1. **Вентиль забывания**:
  2. **Вентиль обновления**:
  3. **Обновление состояния памяти**:
  4. **Вентиль вывода**:

где:

* – сигмоидная функция, преобразующая значения в диапазон от 0 до 1.
* – функция гиперболического тангенса, преобразующая значения в диапазон от -1 до 1.
* ​ – весовые матрицы для каждого вентиля.
* – конкатенация предыдущего скрытого состояния и текущего значения входного вектора.

### **Преимущества и недостатки LSTM**

Преимущества модели LSTM:

1. Работа с последовательными данными: LSTM модели предназначены для работы с последовательными данными, такими как временные ряды, текстовые данные и другие данные с временными зависимостями, что делает их универсальным инструментом для анализа информации во временном контексте.
2. Сохранение информации на длительные периоды: Благодаря вентильным механизмам LSTM решает проблему исчезающих градиентов, позволяя эффективно сохранять данные на долгосрочных интервалах. Это критически важно для задач, где ранние шаги последовательности могут существенно влиять на прогноз в будущем.

Недостатки модели LSTM:

1. Высокая вычислительная сложность: LSTM содержит множество параметров и требует значительных вычислительных ресурсов, особенно для обработки больших объемов данных. Это может стать ограничением при работе на устройствах с низкой производительностью или при обработке данных в реальном времени.
2. Склонность к переобучению: LSTM может испытывать сложности при работе с небольшими объемами данных или в случае недостаточно оптимально подобранных гиперпараметров. Переобучение часто происходит, если модель слишком сложна для рассматриваемой задачи или использует слишком много параметров для малых данных.
3. Сложность для анализа кратковременных зависимостей: В случаях, когда зависимости во временных данных кратковременны или незначительны, LSTM может быть излишне сложной для данной задачи, что приводит к неоправданным вычислительным затратам.

В результате, LSTM является мощным инструментом для анализа временных рядов и других типов последовательных данных. Однако при её применении важно учитывать как достоинства, так и ограничения, подбирая соответствующую архитектуру и гиперпараметры для эффективного выполнения конкретных задач.

## **Языковая модель BERT**

### **Классическая Bidirectional Encoder Representations from Transformers модель**

**Bidirectional Encoder Representations from Transformers (BERT), разработанная компанией Google [8], представляет собой передовую модель для обработки естественного языка (Natural Language Processing, NLP). Она основана на архитектуре трансформеров и отличается от традиционных рекуррентных и сверточных нейронных сетей своим глубоким контекстным пониманием текста (см. рисунок 2.1).**

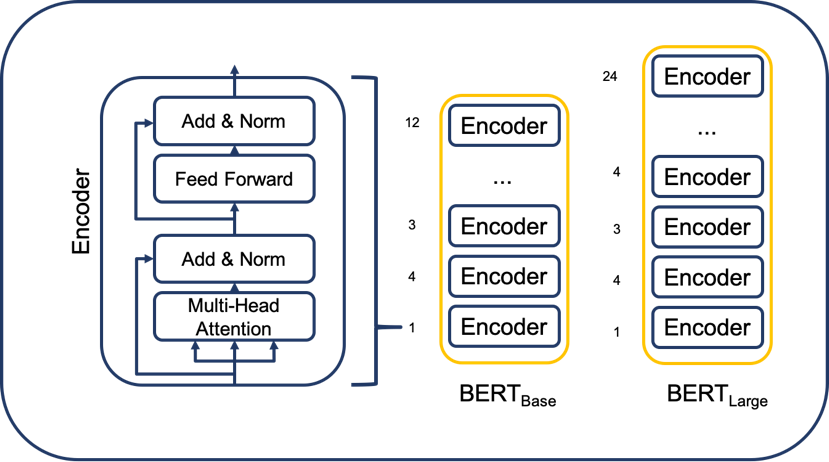


Рисунок 2.1. – Архитектура Bert

**Ключевые особенности модели BERT:**

* **Двунаправленный контекстный анализ: В отличие от моделей, которые обрабатывают текст последовательно слева направо или справа налево, BERT учитывает контекст с обеих сторон одновременно. Это позволяет модели более точно улавливать смысл слов, зависящий от их окружения в предложении.**
* **Архитектура трансформера: BERT использует механизм трансформеров, включающий энкодеры и слои внимания (attention layers). Эта архитектура позволяет модели фокусироваться на наиболее значимых частях текста, эффективно выделяя важные семантические связи между словами.**
* **Маскирование слов при обучении: В процессе обучения модель применяет метод маскированного моделирования языка (Masked Language Modeling). Некоторые слова в тексте заменяются специальным токеном [MASK], и задача модели – предсказать эти скрытые слова на основе контекста. Это способствует глубокому пониманию зависимостей между словами.**
* **Предсказание последовательности предложений: Помимо предсказания скрытых слов, BERT обучается определять, следует ли одно предложение за другим в исходном тексте. Модель получает пары предложений и должна установить их последовательность, что улучшает понимание связности текста.**

**Благодаря этим инновациям, BERT стала одной из наиболее мощных и широко используемых моделей в области NLP. Она способна эффективно обучаться на больших объемах данных и демонстрирует высокую производительность в различных задачах, включая анализ тональности, ответы на вопросы и распознавание именованных сущностей.**

### **Financial Bidirectional Encoder Representations from Transformers модель**

**FinBERT – это специализированная версия модели BERT, адаптированная для обработки текстов финансовой тематики. Модель была дополнительно обучена на большом корпусе финансовых документов, включая новости, отчеты, прогнозы и аналитические обзоры.**

**Особенности модели FinBERT:**

* **Специализация на финансовом контексте: За счет дообучения на специализированных финансовых данных, FinBERT эффективно распознает и интерпретирует терминологию и выражения, характерные для финансовой индустрии. Это позволяет модели точнее анализировать смысл текстов в контексте финансовых рынков.**
* **Точный анализ тональности: Финансовые тексты часто содержат сложные выражения и специфические термины, которые могут быть неправильно интерпретированы общими моделями NLP. FinBERT учитывает эти нюансы, обеспечивая более точный анализ тональности и эмоциональной окраски новостей.**
* **Применение в финансовых задачах: Модель успешно применяется для анализа новостных потоков, оценки рисков, прогнозирования рыночных тенденций и других задач, требующих глубокого понимания финансового контекста и точной интерпретации текстовой информации.**

**Использование FinBERT позволяет значительно повысить качество обработки текстов в финансовой сфере по сравнению с общими моделями, что делает ее незаменимым инструментом для аналитиков и трейдеров при анализе новостного фона и принятии инвестиционных решений.**

## **Методы корреляционного анализа**

Корреляционный анализ – это статистический подход для определения силы и направления связи между переменными. В исследовательской практике для анализа корреляций применяются различные коэффициенты, каждый из которых имеет свои особенности и области применения. Корреляционные методы помогают исследователям выбирать наиболее подходящие способы анализа данных, оценивая силу и характер зависимостей. Рассмотрим основные методы корреляционного анализа, наиболее часто используемые для анализа временных рядов и финансовых данных [9].

Коэффициент корреляции Пирсона – один из самых распространённых методов, который измеряет силу линейной зависимости между двумя количественными переменными. Он показывает, насколько изменения одной переменной могут объяснять изменения другой в линейной зависимости.

Формула для расчёта коэффициента Пирсона:

где:

* ​ и ​ – значения переменных,
* и – средние значения для и .
* Коэффициент Пирсона принимает значения от -1 до 1:
* Значение указывает на идеальную положительную линейную зависимость,
* Значение указывает на идеальную отрицательную линейную зависимость,
* Значение свидетельствует об отсутствии линейной зависимости.

**Применимость**: Коэффициент Пирсона подходит для анализа силы и направления линейной зависимости между двумя количественными переменными, при условии что связь между ними действительно линейная.

Коэффициент Спирмена используется для оценки монотонной зависимости между переменными. В отличие от коэффициента Пирсона, этот коэффициент учитывает ранговую (порядковую) связь, что делает его полезным для нелинейных зависимостей, где переменные изменяются монотонно, но не обязательно линейно.

Формула для расчёта коэффициента Спирмена:

где:

* ​ – разность между рангами значений переменных,
* – общее число наблюдений.
* Коэффициент Спирмена также принимает значения от -1 до 1:
* Значение указывает на идеальную положительную монотонную связь,
* Значение указывает на идеальную отрицательную монотонную связь,
* Значение указывает на отсутствие монотонной связи.

**Применимость**: Коэффициент Спирмена удобен для случаев, когда данные могут быть представлены в ранговой или интервальной шкале, и когда предполагается наличие нелинейных зависимостей между переменными.

Коэффициент Кенделла предназначен для анализа монотонной связи между переменными, используя их ранги. Этот коэффициент учитывает количество согласованных и несогласованных пар рангов, что позволяет оценить порядок значений без учета их фактических величин.

Формула для расчёта коэффициента Кенделла:

где:

* ​ – количество согласованных пар,
* ​ – количество несогласованных пар,
* – общее количество наблюдений.
* Значения коэффициента Кенделла также лежат в диапазоне от -1 до 1:
* свидетельствует о идеальной положительной монотонной связи,
* свидетельствует о идеальной отрицательной монотонной связи,
* указывает на отсутствие монотонной зависимости.

**Применимость**: Коэффициент Кенделла рекомендуется использовать для анализа связи между ранговыми переменными, особенно когда есть значительное количество повторяющихся значений (связанных рангов), и при наличии нелинейных связей.

Для наглядности в таблице 2.1 представлено соотношение между типами шкал и применимостью коэффициентов корреляции.

Таблица 1.1. Соотношение между типами шкал и мерами связи

|  |  |  |
| --- | --- | --- |
| Тип шкалы | | Мера связи |
| Переменная X | Переменная Y |
| Интервальная или отношений | Интервальная или отношений | Коэффициент Пирсона |
| Ранговая, интервальная или отношений | Ранговая, интервальная или отношений | Коэффициент Спирмена |
| Ранговая | Ранговая | Коэффициент Кенделла |

Таким образом, каждый из коэффициентов корреляции имеет свои особенности и применимость, что позволяет выбирать наилучший метод анализа для каждого конкретного типа данных и вида зависимости.

## **Статистический тесты проверки значимости различий признаков**

Для анализа различий между выборками в статистике применяются различные тесты, которые позволяют оценить значимость отклонений наблюдаемых данных от ожидаемых значений. Рассмотрим основные тесты, используемые для проверки гипотез в этом исследовании.

### **Одновыборочный критерий Стьюдента (t-критерий)**

Одновыборочный критерий Стьюдента [10], или t-критерий, применяется для проверки гипотезы о том, что среднее значение нормальной выборки совпадает с заданным значением mmm. Этот тест актуален для выборок с нормальным распределением .

Для выборки размером и заданного значения t-статистика рассчитывается следующим образом:

где:

* – среднее значение выборки ,
* ​ – несмещённая оценка стандартного отклонения,
* – размер выборки,
* – предполагаемое среднее значение для проверки гипотезы.

Гипотеза ​: с альтернативными гипотезами:

1. Двусторонняя проверка: Если альтернативная гипотеза ​ предполагает, что , то критическая область для проверки гипотезы выглядит следующим образом:

где – квантиль нормального распределения на уровне значимости .

1. Односторонняя проверка (для ): Если альтернативная гипотеза , критическая область:
2. Односторонняя проверка (для ): Если , критическая область:

где ​ – квантиль нормального распределения на уровне .

### **Двухвыборочный критерий Стьюдента (t-критерий)**

Двухвыборочный t-критерий используется для проверки гипотезы о равенстве средних значений двух независимых выборок и , каждая из которых предполагает нормальное распределение. Данный тест позволяет проверить, являются ли средние значения двух выборок статистически равными.

Для выборок () и () размером ​ и соответственно t-статистика вычисляется по формуле:

где:

* и – средние значения выборок и ,
* ​ и ​ – выборочные дисперсии для и ,
* и ​ – размеры выборок.

Гипотеза с альтернативными гипотезами:

1. Двусторонняя проверка: Если , то критическая область:
2. Односторонняя проверка (для ​): Если ​, критическая область:
3. Односторонняя проверка (для ​): Если ​, критическая область:

Число степеней свободы для данного t-критерия рассчитывается как .

### **U-критерий Манна-Уитни**

U-критерий Манна–Уитни [11] – это непараметрический тест, применяемый для сравнения двух независимых выборок с целью оценки различий в их распределениях. Этот метод не требует предположения о нормальности распределения данных и является альтернативой двухвыборочному t-критерию.

Постановка гипотез

* **Нулевая гипотеза (**​**)**: распределения обеих выборок одинаковы, т.е. они взяты из одной генеральной совокупности.
* **Альтернативная гипотеза (** ​**)**: распределения выборок отличаются.

Пошаговый алгоритм проверки гипотезы

1. Пусть и – независимые выборки размером ​ и ​.
2. Объедините обе выборки и присвойте каждому элементу ранг в общей последовательности, сортируя значения по возрастанию. Совпадающим значениям присваивается средний ранг.
3. Определите суммы рангов для каждой выборки: ​ и ​.
4. Рассчитайте статистики:
5. Вычислите стандартизированную статистику ​:

при условии, что ​ имеет нормальное распределение .

1. Используйте двустороннюю критическую область для проверки гипотезы:

где – квантиль нормального распределения на уровне значимости .

## **Оценка значимости факторов методом перестановок (Permutation Importance)**

Метод перестановок (permutation importance) [12] является одним из методов, позволяющих оценить влияние каждого признака на точность прогноза модели. Суть подхода заключается в том, что при случайной перестановке значений важного признака качество модели ухудшается. В основе этого метода лежит принцип перестановочного тестирования, также известного как тест Монте-Карло.

Перестановочный тест предполагает случайное перемешивание значений признака или целевой переменной, с последующим анализом изменений качества модели. Если такие изменения значительны, то признак считается важным.

Для оценки значимости каждого признака формулируются следующие гипотезы:

* **Нулевая гипотеза (** ​**)**: Перемешивание значений признака ​ не оказывает влияния на качество модели, что означает, что данный признак не является важным.

где:

* – функция потерь (например, RMSE),
* – модель,
* – набор всех признаков, кроме ​,
* ​ – случайно перемешанный признак ​.
* **Альтернативная гипотеза (** ​**)**: Перемешивание значений признака ​ ухудшает качество модели, что указывает на его значимость.

Пошаговая процедура метода

1. Базовая оценка качества модели: Рассчитываем исходное качество модели с использованием базовой метрики, такой как RMSE:

где – количество наблюдений, – истинное значение целевой переменной для -го наблюдения, – предсказание модели для -го наблюдения.

1. Перестановки признаков: Для каждого признака выполняем перестановок, создавая наборы данных с перемешанными значениями признака ​, где .
2. Расчет качества после перестановки: Применяем модель к каждому набору данных с перемешанным признаком и рассчитываем метрику качества:
3. Оценка разницы в качестве: Для каждой перестановки вычисляем разницу в качестве по сравнению с исходным значением. Также рассчитываем среднюю разницу и стандартное отклонение:
4. Вычисление t-статистики: Для оценки значимости признака рассчитываем t-статистику:

Признаки, для которых наблюдаются значительные положительные изменения в качестве модели при перестановке, считаются важными, так как их наличие в исходных данных существенно влияет на точность предсказаний. Признаки, которые вызывают незначительные изменения при перестановке, могут считаться менее значимыми.

Метод перестановок является удобным и интерпретируемым способом оценки значимости признаков, позволяя не только идентифицировать ключевые переменные для модели, но и принимать обоснованные решения при отборе признаков и интерпретации результатов.Bottom of Form

# **2. Практическая часть**

Для реализации цели прогнозирования цен акций с учётом новостного фона была разработана система, состоящая из нескольких этапов обработки и анализа данных. В ходе работы последовательно осуществляются сбор и предобработка данных, автоматическая разметка тональности новостей, а также анализ и прогнозирование цен акций с использованием нейронных сетей.

## **Использованные модели**

Для перевода текстов новостей с русского на английский язык использована модель Helsinki-NLP/opus-mt-ru-en [13, 14]. Эта модель базируется на архитектуре OPUS-MT (Open Parallel Unified System for Machine Translation), которая представляет собой нейронную сеть, построенную на архитектуре трансформера и предназначенную для задач машинного перевода. OPUS-MT использует методы глубокого обучения, обеспечивая качественный перевод, что необходимо для дальнейшего анализа текста с использованием англоязычных моделей.

После перевода новости передаются на анализ тональности, который выполняется с помощью модели ProsusAI/finbert [15]. Эта нейронная сеть предназначена для анализа текстов в финансовой сфере и возвращает вероятности принадлежности текста к одному из трёх классов тональности: «positive», «negative» или «neutral». Результат модели представлен в виде «логитов» – вектора вероятностей, сумма значений которого равна единице. После анализа каждое сообщение классифицируется по категориальному признаку “sentiment\_category”, принимающему одно из значений: «положительная», «отрицательная» или «нейтральная». Для удобства в дальнейших расчётах категориальные значения преобразуются в числовые: положительная тональность – 1, нейтральная – 0, отрицательная – -1.

Для прогнозирования цены акций на следующий временной шаг использовалась рекуррентная нейронная сеть LSTM (Long Short-Term Memory). Эта модель учитывает последовательность исторических данных и значимость тональности новостей, позволяя прогнозировать будущее поведение цены. Для повышения точности были подобраны оптимальные гиперпараметры, учитывающие особенности временных рядов и динамику новостного фона.

Таким образом, разработанная система сочетает модели для машинного перевода, анализа тональности и прогнозирования цен акций, обеспечивая комплексный подход к обработке и интерпретации данных.

## **Сбор данных**

Для формирования набора данных был разработан парсер, который с использованием библиотек Selenium и Beautiful Soup 4 осуществлял сбор информации о котировках акций с веб-ресурса БКС Экспресс. Парсер позволил извлечь данные о котировках акций ПАО «Сбербанк» за период с начала 2015 года по конец 2023 года, с ежедневной частотой обновлений.

Собранные данные были сохранены в формате CSV и включали следующие параметры: тикер, период, дата, время, цены открытия, максимума, минимума и закрытия, а также объем торгов. Дополнительно были собраны данные о новостях, связанных с акцией «Сбербанк» (тикер SBER) за аналогичный период времени, что позволило расширить контекст для последующего анализа.

В общей сложности по указанному периоду было собрано 3877 новостных сообщений. Новости распределялись по различным категориям, включая такие, как «Теханализ», «Российский рынок», «Среднесрочные тренды» и другие. Статистика по категориям представлена в таблице 2.1.

Таблица 2.1. Статистика категорий новостей

|  |  |
| --- | --- |
| Категория | Количество страниц |
| Теханализ | 2296 |
| Российский рынок | 881 |
| Среднесрочные тренды | 253 |
| Новости | 205 |
| Торговые рекомендации | 157 |
| Мнения | 57 |
| Дивиденды | 28 |

Наибольшее количество новостных статей приходилось на категории «Теханализ», «Российский рынок» и «Среднесрочные тренды», что подтверждает их актуальность для аналитики и прогнозирования цен на акции.

## **Обоснование реализуемости**

В основе нейросетевого подхода к прогнозированию котировок акций лежит идея о том, что новостной фон может напрямую или косвенно повлиять на движение цены актива. В этом исследовании проведен корреляционный анализ влияния тональности новостей на изменения цены акций ПАО «Сбербанк» (SBER) по дням. Результаты анализа представлены в таблице 3.2.

В таблице жирным шрифтом выделены статистически значимые коэффициенты корреляции (с уровнем значимости 0.05). Отметим, что наблюдается достаточно большое количество значений с высокой корреляцией, особенно в течение первых четырех дней после публикации новости. Это позволяет предположить, что новости действительно влияют на динамику актива, хотя связь, как правило, нелинейная и зависит от множества факторов, не ограничиваясь только содержанием новости. Долгосрочное влияние новостного фона в данном анализе не учитывается.

Таблица 2.2. Корреляция и уровень значимости изменения цены закрытия от новостного фона по дням



Для каждого типа новостей оценка их влияния на изменение цены проводилась с использованием всех трёх типов коэффициентов корреляции. Однако, учитывая специфику данных (одна переменная – непрерывная, другая – порядковая), согласно разделу 1.5 и таблице 1.1, наиболее целесообразно было использовать коэффициент корреляции Спирмена, который лучше всего подходит для анализа монотонных связей в условиях порядковых данных.

## **Предобработка данных**

**Анализ данных, представленных в таблице 2.2, показал, что не все категории новостей оказывают значительное влияние на изменения цен акций. Хотя некоторые категории представлены в малом количестве, они не были исключены из анализа, так как могут содержать важную информацию. В то же время, для предотвращения дублирования данных и минимизации возможных искажений были удалены повторяющиеся новости, опубликованные в пределах одного дня.**

**Итоговая выборка составила 1916 торговых дней, в течение которых проводилась оценка новостного фона. Распределение новостей по основным категориям представлено ниже:**

* **Теханализ: 1262 новости**
* **Российский рынок: 374 новости**
* **Среднесрочные тренды: 133 новости**
* **Новости: 49 новостей**
* **Торговые рекомендации: 40 новостей**
* **Мнения: 33 новости**
* **Обучение: 13 новостей**
* **Дивиденды: 12 новостей**

**Распределение новостей по тональности:**

* **Позитивные (1.0): 836 новостей**
* **Негативные (-1.0): 673 новости**
* **Нейтральные (0.0): 407 новостей**

**Таким образом, получена достаточно сбалансированная выборка по признаку тональности, что способствует повышению надежности и точности последующего моделирования и анализа.**

**Для обеспечения корректности и качества анализа были выполнены следующие этапы предобработки текстовой информации:**

* **Удаление специальных символов: исключены символы, не несущие смысловой нагрузки (например, %, $, # и т.д.).**
* **Удаление нумерационных знаков и маркеров списка: убраны элементы форматирования, такие как номера пунктов и буллеты, чтобы унифицировать текст.**
* **Устранение переносов строк: текст преобразован в непрерывный формат для корректной обработки моделями.**
* **Замена табуляций на пробелы: табуляции заменены для предотвращения разрывов в данных и обеспечения целостности текста.**

**Поскольку каждая новость могла содержать упоминания нескольких финансовых инструментов, для точной идентификации были сформированы списки ключевых слов для каждого тикера. Например, для тикера SBER (Сбербанк) был составлен следующий перечень ключевых слов: "SBER", "СБЕР", "Сбербанк", "Сбер", "Сберколл", "Сбербизнес", "Сбертек", "Сбермаркет", "Мегамаркет", "Греф", "СберЛогистика", "СберЛизинг", "СберНедвижимость", "Сбербанк-АСТ", "Сбербанк-Резерв", "Сбербанк-Спасение", "Сбербанк-Страхование", "Сбербанк-Технологии", "Сбербанк-Эквайринг", "Сбербанк-Эксперт", "Сбербанк-Контактный Центр".**

**Используя эти списки, был проведен детальный анализ новостных сообщений для выявления всех релевантных упоминаний. Фильтрация новостей осуществлена корректно, без возникновения пустых значений, что подтверждает качество проведенной работы.**

**Данные о ценах закрытия акций были получены с ресурса Finam. Для обеспечения сопоставимости и корректной работы моделей машинного обучения значения цен закрытия (close\_price) были нормализованы в диапазон от 0 до 1. Нормализация позволяет модели эффективно обучаться и улучшает точность прогнозирования.**

**После нормализации данные о котировках были объединены с новостными данными по дате, что позволило связать изменение цен с соответствующим новостным фоном. Из объединенного набора данных были исключены все несущественные столбцы, оставив только:**

* **close\_price: нормализованная цена закрытия акции.**
* **sentiment\_category: категориальная оценка тональности новости (позитивная, негативная, нейтральная).**

**Для построения и оценки модели данные были разделены на:**

* **Обучающую выборку: 90% данных, используемых для обучения нейронной сети.**
* **Тестовую выборку: 10% данных, предназначенных для проверки и валидации модели.**

**Такой подход к разделению данных обеспечивает баланс между обучением модели и ее способностью обобщать знания на новых данных.**

## **Описание использования LSTM**

**Модель, разработанная для прогнозирования цен на акции, построена на основе улучшенной архитектуры LSTM. Она включает несколько слоев LSTM, слой нормализации батчей (Batch Normalization) и завершается полносвязным слоем для окончательного предсказания. Такая комбинация позволяет эффективно обрабатывать временные ряды и улучшает обобщающую способность модели.**

**Описание архитектуры модели:**

* 1. **Входные данные:**
* **Размерность входных данных: Трехмерный тензор размера *(batch\_size, sequence\_length, input\_dim)*.**
* ***batch\_size* – размер пакета (может варьироваться).**
* ***sequence\_length* – длина последовательности временных шагов (например, 30 дней).**
* ***input\_dim* – количество входных признаков (в данном случае 1, представляющий цену акции).**
  1. **Первый LSTM-слой:**
* **Параметры слоя:**
* ***input\_dim* = 1.**
* ***hidden\_dim* = 64 (количество нейронов в скрытом состоянии).**
* ***num\_layers* = 3 (количество слоев LSTM).**
* ***dropout* = 0.3 (вероятность отключения нейронов для регуляризации).**
* **Выходные данные:**
* **Трехмерный тензор размера *(batch\_size, sequence\_length, hidden\_dim)*.**
  1. **Dropout-слой:**
* **Назначение:**
* **Снижает вероятность переобучения путем случайного отключения 30% нейронов во время обучения.**
* **Параметры слоя:**
* **Не имеет обучаемых параметров.**
* **Выходные данные:**
* **Сохраняет размерность входного тензора *(batch\_size, sequence\_length, hidden\_dim)*.**

1. **Извлечение последнего временного шага:**

**После прохождения через LSTM-слои и Dropout, из выходного тензора выбирается последний временной шаг для каждого элемента в пакете. Это позволяет модели фокусироваться на самой последней информации при прогнозировании.**

1. **Batch Normalization слой:**

* **Назначение:**
* **Нормализует выходы предыдущего слоя, ускоряя и стабилизируя процесс обучения.**
* **Параметры слоя:**
* **Имеет обучаемые параметры: γ и β для каждого из *hidden\_dim* нейронов.**
* **Выходные данные:**
* **Тензор размера *(batch\_size, hidden\_dim)*.**

1. **Полносвязный (Dense) слой:**

* **Назначение:**
* **Преобразует выходы из Batch Normalization слоя в окончательный прогноз.**
* **Параметры слоя:**
* **Входные нейроны: *hidden\_dim* = 64.**
* **Выходные нейроны: *output\_dim* = 1 (предсказание цены).**
* **Выходные данные:**
* **Тензор размера *(batch\_size, output\_dim)*.**

**Общее количество параметров модели:**

* **LSTM-слои: 50 688 параметров.**
* **Batch Normalization слой: 128 параметров.**
* **Полносвязный слой: 65 параметров.**
* **Итого: 50 688+128+65=50 881 параметров.**

**Представленная архитектура нейронной сети эффективно сочетает в себе преимущества глубоких рекуррентных слоев для обработки временных рядов с методами регуляризации и нормализации. Это позволяет модели точно прогнозировать цены на акции, учитывая исторические данные и выявляя скрытые закономерности. Использование таких технологий, как LSTM, Dropout и Batch Normalization, обеспечивает баланс между сложностью модели и её способностью к обобщению, что особенно важно в условиях волатильности финансовых рынков.**

## **Подбор гиперпараметров модели**

**Для повышения точности прогнозирования цен акций и оптимизации работы разработанной модели была проведена тщательная настройка гиперпараметров. Этот процесс, известный как гиперпараметрическая оптимизация, позволяет найти такие значения параметров модели, при которых она показывает наилучшую производительность на тестовых данных.**

**Для автоматизации процесса подбора гиперпараметров был использован метод GridSearchCV из библиотеки scikit-learn. Данный инструмент позволяет перебрать все возможные комбинации заданных гиперпараметров и оценить модель по заданной метрике качества. В нашем случае в качестве метрики использовалась RMSE (Root Mean Squared Error), которая хорошо подходит для задач регрессии и отражает среднеквадратичное отклонение предсказанных значений от фактических.**

**В рамках оптимизации рассматривались следующие гиперпараметры модели:**

* **predicted\_days: количество предыдущих дней, используемых моделью для прогнозирования следующего значения. Значения: 15, 30, 60.**
* **learning\_rate: скорость обучения оптимизатора Adam. Значения: 0.001, 0.003, 0.01.**
* **dropout\_rate: коэффициент отсева в слоях LSTM для регуляризации модели и предотвращения переобучения. Значения: 0.1, 0.2, 0.3.**
* **EPOCHS: количество эпох обучения модели. Значения: 50, 100, 150.**
* **BATCH\_SIZE: размер пакета данных для одного шага обновления весов модели. Значения: 8, 16, 32, 64.**

**Всего было рассмотрено 3×3×3×3×4=324 различных комбинаций гиперпараметров. Каждая комбинация использовалась для обучения отдельной модели на данных акции SBER, что обеспечило всестороннее исследование пространства параметров.**

**Для оценки производительности каждой модели использовалась метрика RMSE, которая вычисляется по формуле:**

**где ​ – фактическое значение, ​ – предсказанное моделью значение, n – количество наблюдений.**

**Низкое значение RMSE указывает на высокую точность модели.**

**По результатам перебора было выявлено, что наилучшая производительность модели достигается при следующих значениях гиперпараметров:**

* **predicted\_days: 30**
* **dropout\_rate: 0.2**
* **learning\_rate: 0.001**
* **EPOCHS: 50**
* **BATCH\_SIZE: 32**

**Обоснование выбора данных параметров:**

* **predicted\_days = 30: Использование 30 дней исторических данных для прогнозирования оказалось оптимальным балансом между достаточным объемом информации и сложностью модели. Меньшее количество дней (15) не позволяло модели учесть все необходимые тенденции, тогда как большее (60) приводило к переобучению и усложнению модели без существенного прироста точности.**
* **dropout\_rate = 0.2: Коэффициент отсева 20% позволил эффективно бороться с переобучением, сохраняя при этом способность модели обучаться на значимых паттернах в данных.**
* **learning\_rate = 0.001: Низкая скорость обучения обеспечила стабильную сходимость модели. Более высокие значения приводили к колебаниям и возможному пропуску глобального минимума функции потерь.**
* **EPOCHS = 50: Оптимальное количество эпох позволило модели достичь необходимой точности без переобучения. Увеличение количества эпох не приводило к значительному улучшению результатов, а иногда даже ухудшало их из-за переобучения.**
* **BATCH\_SIZE = 32: Размер пакета в 32 наблюдения обеспечил баланс между скоростью обучения и стабильностью градиентного спуска. Меньшие значения увеличивали время обучения, а большие – снижали точность модели.**

**Модель, обученная с указанными выше параметрами, показала наилучшие результаты на тестовых данных, демонстрируя минимальное значение RMSE по сравнению с другими комбинациями гиперпараметров. Это свидетельствует о ее высокой точности и способности к обобщению на новых данных.**

**Проведенная гиперпараметрическая оптимизация позволила значительно улучшить качество прогнозирования модели. Использование метода GridSearchCV обеспечило систематический подход к подбору параметров и гарантировало, что выбранная комбинация гиперпараметров является оптимальной в рамках заданного пространства поиска.**

**Данные результаты подтверждают эффективность разработанной модели и ее потенциал для использования в качестве аналитического инструмента при поддержке торговых решений на финансовых рынках.**

## **Описание полученных результатов**

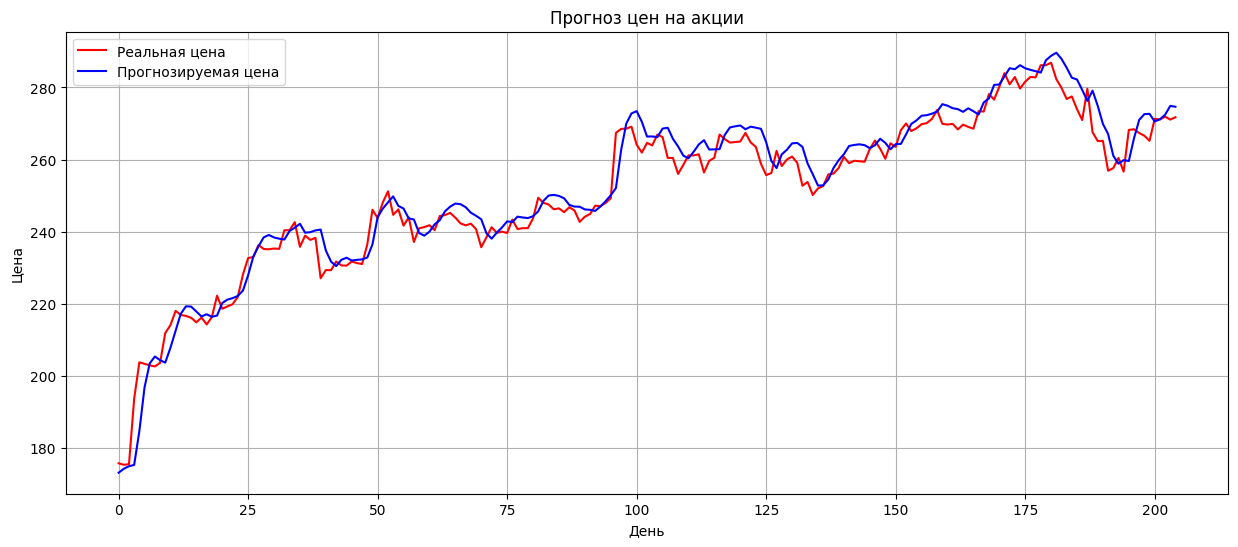
**На примере котировок акций SBER были проведены эксперименты для оценки влияния учета тональности новостей на точность прогнозирования цен акций с использованием разработанной нейронной сети. Были обучены две модели: одна с использованием признака тональности новостей (real\_score), другая – без него.**

**Нейронная сеть, обученная с учетом признака тональности новостей, показала следующие результаты на этапе обучения:**

* **Эпоха 0, Потери: 3.3148**
* **Эпоха 10, Потери: 0.6177**
* **Эпоха 20, Потери: 0.5161**
* **Эпоха 30, Потери: 0.5338**
* **Эпоха 40, Потери: 0.3588**

**После обучения модель была протестирована на тестовой выборке, где достигла значения RMSE (Root Mean Squared Error) равного 4.73 рубля. Это составляет примерно 1.8% от цены актива на момент прогноза, что свидетельствует о высокой точности модели.**

**Визуализация результатов модели с учетом тональности представлена на рисунке 2.1, где показано сравнение фактических и предсказанных цен акций на каждом временном шаге. Видно, что предсказания модели близко следуют за реальными значениями, демонстрируя способность модели улавливать тенденции рынка с учетом новостного фона.**

****

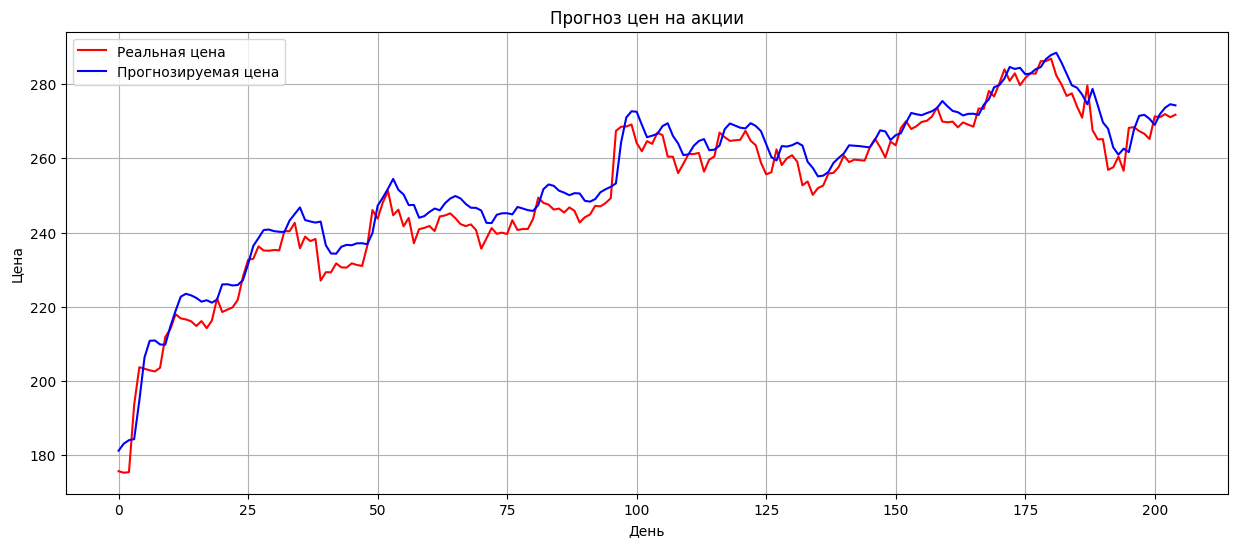
**Рисунок 2.1.** – Визуализация результатов прогнозирования модели с учетом тональности новостей

**Для сравнения была обучена аналогичная модель, но без использования признака тональности новостей. Результаты обучения данной модели следующие:**

* **Эпоха 0, Потери: 5.1408**
* **Эпоха 10, Потери: 0.5888**
* **Эпоха 20, Потери: 0.5146**
* **Эпоха 30, Потери: 0.3997**
* **Эпоха 40, Потери: 0.4187**

**На тестовой выборке эта модель показала значение RMSE равное 5.13 рубля, что составляет примерно 2% от цены актива на момент прогноза. Хотя разница в процентах незначительна, абсолютное значение RMSE выше по сравнению с моделью, учитывающей тональность новостей.**

**Визуализация результатов модели без учета тональности представлена на рисунке 2.2. Из графика видно, что предсказания модели несколько менее точно повторяют фактические значения, особенно в периодах резких изменений цен.**

****

**Рисунок 2.2** – Визуализация результатов прогнозирования модели без учета тональности новостей

****

Рисунок 2.3 – Сравнение реальных цен акций и прогнозов моделей на тестовой выборке

Из графика видно, что модель с учетом тональности новостей (синяя линия) наиболее точно повторяет динамику реальных цен акций. Модель без учета тональности (зеленая линия) показывает несколько менее точные предсказания, особенно в периодах резких изменений цен.

**Сравнивая результаты двух моделей, можно отметить следующее:**

* **Учет тональности новостей улучшает точность прогнозирования: модель с признаком тональности показала более низкое значение RMSE (4.73 рубля) по сравнению с моделью без этого признака (5.13 рубля).**
* **Быстрота сходимости: модель с учетом тональности новостей показала более значительное снижение потерь к 40-й эпохе обучения, что свидетельствует о лучшей обучаемости модели при наличии дополнительной информации.**
* **Абсолютное улучшение RMSE составляет 0.4 рубля, что, хотя и может показаться небольшим, в контексте финансовых рынков и при больших объемах торгов может иметь существенное значение.**

Для проверки статистической значимости различий между моделями был проведен непараметрический тест Манна-Уитни. Результаты показали, что различия между RMSE моделей с учетом тональности и без нее статистически значимы (p-значение < 0.05). Это подтверждает, что включение тональности новостей в модель улучшает качество прогнозирования.

Таким образом, использование признака тональности новостей в сочетании с историческими данными о ценах акций улучшает эффективность модели и может быть рекомендовано для практического применения в задачах поддержки принятия торговых решений на финансовых рынках.

# **Заключение**

**В ходе данной работы была разработана комплексная система для прогнозирования цен акций с учетом новостного фона. Основой решения стала многоуровневая архитектура, включающая этапы сбора, предобработки, анализа и прогнозирования данных. Благодаря использованию моделей машинного перевода, анализа тональности и рекуррентной нейронной сети LSTM, удалось создать эффективную систему, способную учитывать влияние новостей на динамику цен.**

**Проведенные эксперименты показали, что добавление новостной информации в виде признака тональности позволяет улучшить точность модели. Модель с учетом тональности показала RMSE 4.73 рубля, в то время как модель без тональности достигла RMSE 5.13 рубля. Таким образом, снижение RMSE на 0.4 рубля указывает на более точное предсказание цен акций, что подтверждается графической визуализацией, на которой модель с тональностью лучше повторяет фактическую динамику цен.**

**Для подтверждения значимости различий был проведен тест Манна-Уитни, показавший статистически значимые отличия между RMSE двух моделей. Это свидетельствует о том, что учет новостного фона улучшает качество прогнозов, делая модель более чувствительной к информационному фону рынка.**

**Разработанная система продемонстрировала высокую точность и эффективность в анализе и прогнозировании цен акций. Она обладает потенциалом для практического применения в задачах поддержки принятия торговых решений, а также может быть доработана и масштабирована для прогнозирования других финансовых активов.**

**Список литературы**

1. Bollen, J., Mao, H., & Zeng, X. (2010). Twitter mood predicts the stock market. arXiv. Получено 14 ноября 2024, из <https://doi.org/10.48550/arXiv.1010.3003>
2. Куликова, Т. Д., Ковтун, Е. Ю., & Буденный, С. А. (2023). Получаем ли мы пользу от категоризации потока новостей в задаче прогнозирования цен акций? Доклады Российской академии наук. Математика, информатика, процессы управления, 514(2), 385–394.
3. Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 65(4), 782-796.
4. Daily Financial News for 6000+ Stocks // Kaggle URL: https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests/data (дата обращения: 07.11.2024).
5. Financial News Sentiment Dataset (FiNeS) // GitHub URL: <https://github.com/WebOfRussia/financial-news-sentiment> (дата обращения: 07.11.2023).
6. Twitter Financial News Sentiment // HuggingFace URL: <https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment> (дата обращения: 07.11.2024).
7. Elsworth, Steven, and Stefan Güttel. "Time series forecasting using LSTM   
   networks: A symbolic approach." arXiv preprint arXiv:2003.05672 (2020).
8. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019.   
   Bert: Pre-training of deep bidirectional transformers for language   
   understanding. In Proceedings of NAACL, pages 4171–4186.
9. Математические методы обработки данных: учебно-методическое   
   пособие для студентов бакалавров физико-математического факультета   
   / Н.Б. Ивирсина, М.В. Танзы, Е.К. Бичи-оол, А.М. Хомушку ; Тувинский   
   государственный [университет. – Кызыл : Издательство ТувГУ, 2021. –   
   129 с.
10. Горлач, Б. А. Теория вероятностей и математическая статистика :   
    учебно-методическое пособие / Б. А. Горлач. – Санкт-Петербург : Лань,   
    2022. – 320 с. – ISBN 978-5-8114-1429-1. – Текст : электронный //   
    Лань : электронно-библиотечная система. – URL:   
    https://e.lanbook.com/book/211082 (дата обращения: 27.11.2024). –   
    Режим доступа: для авториз. пользователей.
11. Буре, В. М. Теория вероятностей и математическая статистика : учебник   
    / В. М. Буре, Е. М. Парилина. – Санкт-Петербург : Лань, 2022. – 416 с.   
    – ISBN 978-5-8114-1508-3. – Текст : электронный // Лань : электронно-  
    библиотечная система. – URL: https://e.lanbook.com/book/211250 (дата   
    обращения: 27.11.2024). – Режим доступа: для авториз. пользователей.
12. Good P. Permutation, Parametric and Bootstrap Tests of Hypotheses. - 3rd ed.   
    NY: Springer Science+Business Media, 2004. - 331 p.
13. Helsinki-NLP opus-mt-en-ru / Helsinki-NLP [Электронный ресурс] //   
    huggingface.co – URL: https://huggingface.co/Helsinki-NLP/opus-mt-en-ru   
    (дата обращения: 25.10.2024).
14. Helsinki-NLP opus-mt-en-ru / Helsinki-NLP [Электронный ресурс] //   
    github.com – URL: https://github.com/Helsinki-NLP/Opus-MT (дата   
    обращения: 25.10.2024).
15. ProsusAI/finbert // huggingface.co URL:   
    https://huggingface.co/ProsusAI/finbert (дата обращения: 25.10.2024).