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, применяемых для обработки последовательностей данных с учетом долгосрочных зависимостей.

**BERT** – Bidirectional Encoder Representations from Transformers, модель обработки естественного языка, которая учитывает контекст с двух сторон предложения.

**FinBERT** – Финансовая версия модели BERT, дообученная на финансовых текстах, позволяющая точно анализировать тональность новостей в контексте финансов.

Helsinki-NLP/opus-mt-ru-en – модель для перевода текста с русского на   
английский.

**MAE** – Средняя абсолютная ошибка (Mean Absolute Error), оценивающая среднее значение абсолютных ошибок в прогнозе.

**MSE** – Среднеквадратичная ошибка (Mean Squared Error), вычисляемая как среднее квадратов разности между предсказанными и фактическими значениями.

**R²** – Коэффициент детерминации, характеризующий качество модели и точность предсказаний.

# **Введение**

В условиях современного финансового мира, где рынки подвержены постоянным изменениям, информация играет ключевую роль в принятии торговых решений. Поток новостей, поступающих в реальном времени, оказывает значительное влияние на настроения инвесторов и, соответственно, на динамику фондовых рынков. Исследования и разработки, направленные на анализ новостного фона с использованием методов машинного обучения, приобретают все большее значение, так как они позволяют автоматизировать оценку новостей и прогнозировать поведение активов.

Целью данного исследования является создание аналитического торгового робота, способного на основе интеллектуального анализа новостного контента формировать рекомендации по покупке и продаже финансовых активов. Разрабатываемая система выступает в роли аналитического инструмента, способного глубоко анализировать данные о котировках и новостях, предоставляя трейдерам обоснованные рекомендации. Торговый робот не осуществляет самостоятельных сделок, а лишь выполняет функции поддержки принятия решений, оценивая новостной фон и рыночные данные.

Для достижения поставленной цели в работе решаются следующие задачи:

1. Собрать данные новостей и котировок для проведения анализа;
2. Предобработать исходные данные для последующего анализа;
3. Построить архитектуру нейронной сети для прогнозирования на основе новостей;
4. Обучить нейронной сеть и протестировать модель;
5. Провести анализ результатов с использованием метрик качества для оценки точности модели.

Актуальность данного исследования обусловлена постоянными изменениями финансовых рынков и их высокой чувствительностью к новостным событиям. Применение методов машинного обучения в анализе и оценке новостей открывает перспективы создания интеллектуальных торговых систем, способных оперативно реагировать на изменения информационного фона. В рамках работы предлагается анализ тональности новостей и разработка моделей глубокого обучения для прогнозирования движения активов на основе эмоциональной окраски новостных сообщений.

Настоящая работа направлена на формирование подхода к использованию новостного фона для анализа и прогнозирования, что будет полезно для трейдеров и финансовых аналитиков в условиях высокой волатильности рынков.  
 **Обзор источников литературы**

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

В последние годы появилось множество исследований, фокусирующихся на изучении связи между новостями и финансовыми рынками.

1. В работе Bollen, J., Mao, H., & Zeng, X. (2010) исследуется взаимосвязь между общественным настроением, оцененным по сообщениям в Twitter, и индексом Dow Jones Industrial Average (DJIA). Авторы анализируют возможность предсказания рыночных изменений на основе изменений общественных настроений, используя методы, такие как Granger-каузальность и нелинейные модели, в частности, Self-Organizing Fuzzy Neural Network (SOFNN). Результаты исследования демонстрируют, что настроение, определяемое с помощью Twitter, может служить индикатором для прогнозирования изменений на фондовом рынке, причем 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) рассматривается влияние тематической категоризации новостей на точность прогнозов цен акций. Исследователи разделили новости на 20 тематических групп, используя модель BERT, и определяли тональность новостей с помощью модели FinBERT. Выяснилось, что учет настроений в рамках одной тематической группы улучшает точность предсказания, по сравнению с анализом всего новостного потока или его полного отсутствия. Эмоциональная насыщенность новостей играет важную роль, показывая, что как положительные, так и отрицательные новости могут существенно влиять на точность прогнозов. Таким образом, тематическая классификация и учет настроений оказываются полезными инструментами для более глубокого понимания рыночных тенденций и поведенческих факторов, что подтверждает возможность повышения точности прогнозирования при учете специфических тематических категорий.

Эти исследования подтверждают, что анализ настроений в сочетании с тематической категоризацией новостей может существенно повысить точность моделей прогнозирования цен акций и предоставить более глубокое понимание рыночного поведения.

# **Датасеты**

Для анализа финансовых новостей и обучения моделей прогнозирования цен на акции важно иметь доступ к качественным и актуальным датасетам. В процессе исследования было рассмотрено несколько доступных источников данных:

1. Sentiment Analysis for Financial News [3]: Этот набор данных включает анализ тональности 4837 статей, разделенных на три класса (положительный, нейтральный, отрицательный). Однако в датасете отсутствуют даты публикации, что не позволяет установить прямую связь с динамикой цен конкретных акций.
2. Daily Financial News for 6000+ Stocks [4]: Данный датасет содержит 843062 записей с указанием дат публикации и тикеров акций. Тем не менее, он охватывает информацию только по трем компаниям и не имеет целевого показателя для определения тональности новостей (положительная или отрицательная).
3. Financial News Sentiment Dataset (FiNeS) [5]: Включает оценку тональности заголовков новостей по шкале от -1 до 1 и предоставляет данные о российских акциях. Недостатком является небольшой объем данных – всего 532 записи, что может быть недостаточно для качественного обучения моделей.
4. Twitter Financial News Sentiment [6]: Содержит 11932 записи с классификацией новостей на "медвежьи", "бычьи" и "нейтральные". Однако в нем отсутствуют данные о дате публикации и связи с конкретными акциями, что затрудняет его использование для временного анализа.

Каждый из рассмотренных датасетов имеет свои ограничения, такие как отсутствие информации о дате публикации, недостаточное количество данных или отсутствие целевых меток для классификации тональности новостей. Это подчеркиет необходимость создания собственного парсера для сбора более детализированных и релевантных данных.

# **Теоретическая часть**

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

Разработанный в данной работе торговый робот представляет собой полуавтоматическую систему, предназначенную для анализа финансовых данных и выдачи рекомендаций по сделкам на фондовом рынке. Робот не принимает торговых решений самостоятельно, а выполняет аналитическую функцию, помогая пользователю оценить текущую рыночную ситуацию и принять взвешенные решения.

В данном разделе описана концепция многоуровневой архитектуры сети, которая включает в себя комбинацию нескольких моделей, таких как:

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

## **Общая концепция многоуровневой архитектуры**

Многоуровневая модель, реализованная в этом исследовании (см. рисунок 1.1), представляет собой последовательность этапов обработки данных, где каждая модель выполняет определенную задачу. Такой подход позволяет получать более точные прогнозы благодаря многоуровневой обработке информации.

Архитектура модели включает несколько ключевых уровней:

* 1. Перевод текстов новостей. На первом этапе происходит перевод новостных текстов с русского на английский язык. Это необходимо для использования англоязычных моделей анализа текста, таких как BERT, которые более точно интерпретируют тональность на английском языке.
  2. Анализ тональности текста. Следующий уровень предполагает применение англоязычной модели BERT для анализа тональности новостного контента. На этом этапе определяется, является ли текст позитивным, негативным или нейтральным, что позволяет оценить эмоциональный фон новостей.
  3. Прогнозирование ценовых трендов на основе временных рядов. Последний этап включает обработку временных данных с использованием модели LSTM, которая прогнозирует изменения в ценах активов на основе исторических данных.

Каждый этап модели представляет собой отдельный уровень обработки данных, при этом вывод одной модели подается на вход следующей. Такой поэтапный процесс позволяет учитывать различные аспекты информации из новостных сообщений и временных рядов, что способствует улучшению прогноза и повышению обоснованности торговых рекомендаций.

На рисунке 1.1 исходные данные, такие как новости и котировки, отображены голубыми трапециевидными блоками. Этапы предобработки данных отмечены красными прямоугольниками, нейросетевые модели представлены оранжевыми блоками, а прогноз, полученный на выходе модели, обозначен зеленым трапециевидным блоком.

РИСУНОК

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 .

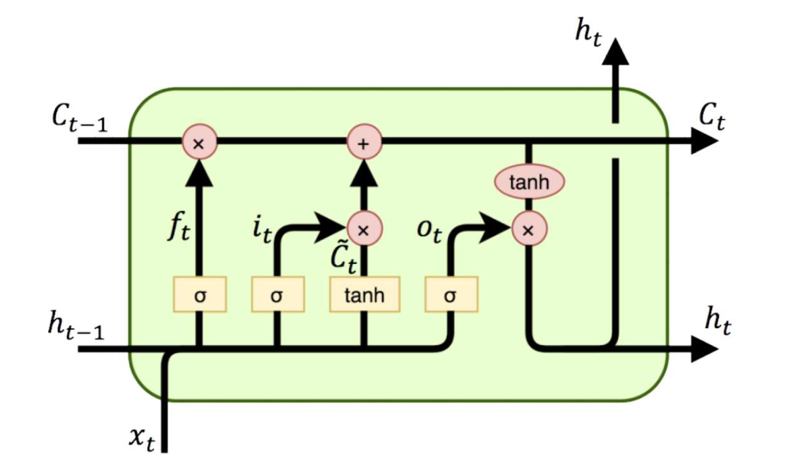


Рисунок 1.2. – Структура LSTM сети.

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, или BERT [8], является моделью, разработанной Google для решения задач обработки естественного языка. Она базируется на архитектуре трансформера и отличается от традиционных рекуррентных и сверточных нейронных сетей рядом особенностей, которые позволяют ей глубже анализировать контекст текста (см. Рисунок 2.1).

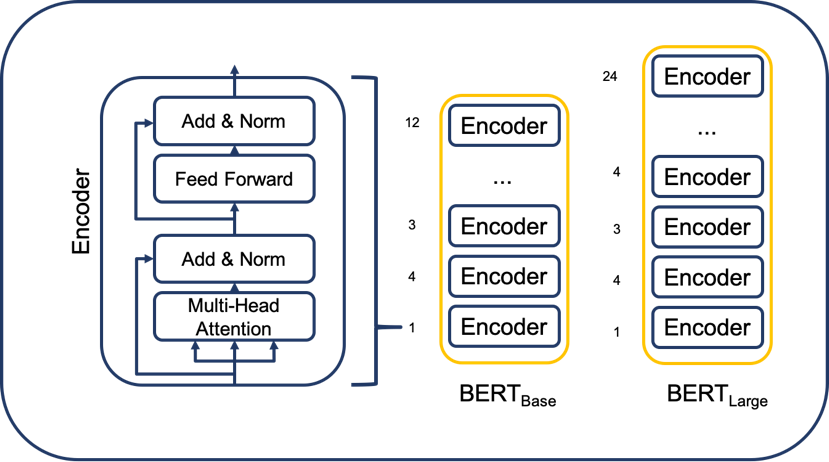


Рисунок 2.1. – Архитектура Bert

Основные черты BERT включают:

* Учет контекста с обеих сторон: В отличие от моделей, которые учитывают контекст лишь слева направо или справа налево, BERT рассматривает слова как с левой, так и с правой стороны. Это позволяет модели более точно понимать смысл слов, зависящий от контекста.
* Архитектура трансформера: BERT использует трансформер, включающий энкодеры и механизмы внимания (attention layers), что позволяет модели фокусироваться на ключевых словах и фразах в предложении. Слои самообучения и многоуровневая структура трансформера обеспечивают глубокое изучение контекста.
* Маскирование слов при обучении: При обучении BERT случайно скрывает некоторые слова и учится предсказывать их на основе остальных слов в предложении. Этот подход помогает модели лучше понимать контекст и выявлять зависимости между словами.
* Предсказание связи предложений: В процессе обучения модель также анализирует пары предложений и учится определять, следует ли одно предложение за другим, что помогает BERT лучше работать с последовательностями предложений.

Эти особенности сделали BERT одной из ведущих моделей в области обработки естественного языка, так как она способна эффективно обучаться на больших объемах данных и обобщать свои знания для анализа текста.

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

FinBERT представляет собой адаптацию BERT, специально разработанную для обработки текстов финансовой тематики. Эта модель была дообучена на текстах, относящихся к финансовым рынкам, включая новости, отчеты, прогнозы и инвестиционные обзоры.

Главной особенностью FinBERT является его способность эффективно обрабатывать финансовую лексику и контексты, характерные для данной отрасли. Это достигается путем дополнительного обучения на специализированных данных, что позволяет модели точнее определять тональность и смысл текста, исходя из специфики финансового языка. FinBERT успешно справляется с задачами анализа текстов из финансовой сферы, что делает её более точной для задач, связанных с анализом новостей, оценкой отчетов и других финансовых документов.

## **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:

Интерпретация результатов

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 2016 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 3396 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 | 1957 |
| Российский рынок | 757 |
| Среднесрочные тренды | 247 |
| Новости | 159 |
| Торговые рекомендации | 62 |
| Мнения | 53 |
| Дивиденды | 26 |
| Обучение | 22 |
| Около рынка | 15 |
| Мировые рынки | 9 |
| Инвестидеи | 3 |
| Реклама | 1 |

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

ТАБЛИЦА

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**

## **Описание использования LSTM**

## **Подбор гиперпараметров модели**

## **Описание полученных результатов**

# **Заключение**

**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).