МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ УНИВЕРСИТЕТ «МИФИ»

Направление подготовки:
01.04.02 «Прикладная математика и информатика»

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Проектный практикум на тему:

Прогнозирование погоды с помощью нейросетевых моделей

Магистр группы М24-525 Груданов Николай Алексеевич
(индекс) (фамилия, имя, отчество)

МОСКВА 2025

АННОТАЦИЯ

Прогнозирование погоды является одной из важнейших задач, влияющих на множество сфер, включая сельское хозяйство, транспорт, энергетику и повседневную жизнь. Современные технологии искусственного интеллекта, такие как нейросетевые модели и платформы обработки данных, открывают новые возможности для повышения точности и эффективности метеорологических прогнозов.

В рамках данной работы был разработан телеграм-бот, который интегрирует современные технологии, включая LangChain и GigaChat, а также несколько нейросетевых моделей, специально обученных для прогнозирования погодных условий. В качестве моделей использованы BERT для анализа текстовых данных, Forecast для обработки временных рядов и LSTM для предсказания ключевых метеорологических параметров, таких как температура, влажность и осадки.

Кроме того, работа содержит сравнительный анализ существующих решений в области прогнозирования погоды с использованием искусственного интеллекта. Проведённый анализ позволил выявить преимущества и ограничения различных подходов, что стало основой для выбора оптимальной архитектуры системы.

Настоящая работа состоит из введения, трех глав и заключения. Общий объем данной работы – 20 страниц, 5 рисуноков, 2 таблица.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Прогнозирование погоды является важной задачей, которая находит применение в различных сферах, таких как сельское хозяйство, транспорт, энергетика и повседневная жизнь людей. Повышение точности и доступности прогнозов погоды способствует улучшению планирования и снижению рисков, связанных с неблагоприятными погодными условиями. Современные технологии искусственного интеллекта и машинного обучения предоставляют новые возможности для создания интеллектуальных систем прогнозирования.

Актуальность

Использование методов искусственного интеллекта для прогнозирования погоды является актуальной задачей, поскольку традиционные подходы имеют ограничения в обработке больших объемов данных и учете сложных нелинейных зависимостей. Интеграция современных моделей машинного обучения с инструментами обработки естественного языка и пользовательскими интерфейсами позволяет создавать более эффективные и удобные решения для пользователей.

Цель работы

Разработка интеллектуальной системы прогнозирования погоды с использованием современных технологий искусственного интеллекта и машинного обучения.

Задачи

1. Провести анализ современных моделей для прогнозирования погоды.
2. Сформировать набор данных о погоде для обучения моделей.
3. Обучить модели прогнозирования (BERT, Forecast, LSTM).
4. Интегрировать модели через LangChain.
5. Подключить GigaChat для обработки запросов на естественном языке.
6. Создать Telegram-бота для взаимодействия с пользователями.

Теоретическая и практическая значимость

Результаты работы могут быть использованы для повышения точности прогнозов погоды, а также в качестве аналитического инструмента для метеорологических исследований. Практическая реализация системы в виде Telegram-бота обеспечивает удобный доступ к прогнозам для широкого круга пользователей.

Объект исследования

Модели, методы и средства сбора, анализа и визуализации данных для мониторинга и прогнозирования погоды.

Предмет исследования

Разработка положений использования новых информационных и коммуникационных технологий с целью повышения эффективности управления процессами прогнозирования погоды.

1. ОБЗОР И АНАЛИЗ РЕШЕНИЙ ДЛЯ ПРОГНОЗА ПОГОДЫ

В данной главе представлен обзор и анализ моделей и методов, используемых для прогнозирования погодных условий. Рассмотрены как классические подходы, так и современные нейросетевые архитектуры, применяемые в рамках проекта.

Модель **BERT** (Bidirectional Encoder Representations from Transformers), разработанная для обработки текстовых данных, была адаптирована для задач анализа временных рядов и прогнозирования погоды. Её способность учитывать контекст данных в обоих направлениях делает её полезной для анализа сложных зависимостей между погодными параметрами. Однако её архитектура позволяет адаптировать модель для анализа временных рядов, включая данные о погоде.

Входные данные: временные ряды или последовательности текстовых описаний (например, погодные явления).

Преимущества:

- Отлично работает с последовательными данными благодаря двунаправленному анализу.
- Может учитывать сложные взаимосвязи между параметрами.

Недостатки:

- Высокая вычислительная сложность.
- Требует больших объемов данных для обучения.

LSTM (Long Short-Term Memory)— это разновидность рекуррентных нейронных сетей, специально разработанная для работы с временными рядами. Её архитектура позволяет эффективно моделировать долгосрочные зависимости в данных, что особенно важно для прогнозирования погодных условий.

Входные данные: временные ряды (температура, влажность, осадки и другие параметры).

Преимущества:

- Эффективно моделирует долгосрочные зависимости.
- Хорошо справляется с сезонными и временными вариациями.

Недостатки:

- Склонна к переобучению на небольших наборах данных.
- Обучение может быть медленным при увеличении объема данных.

Random Forest — это ансамблевый метод машинного обучения, основанный на построении множества решающих деревьев. Этот подход обеспечивает высокую устойчивость к шуму в данных и позволяет получать точные прогнозы даже при наличии выбросов или пропусков в данных.

Входные данные: числовые или категориальные признаки (например, температура, давление, влажность).

Преимущества:

- Высокая устойчивость к шуму в данных.
- Простота реализации и интерпретации результатов.

Недостатки:

- Ограниченная способность работать с временными зависимостями.
- Менее эффективен для обработки больших объемов данных по сравнению с нейросетевыми моделями.

GraphCast представляет собой инновационный подход к прогнозированию погоды, основанный на графовых нейронных сетях (GNN). Эта модель учитывает пространственно-временные зависимости между различными метеорологическими станциями и регионами, что позволяет улучшить точность прогнозов на больших территориях. Она использует сложные наборы данных, такие как ERA5, содержащие многомерные климатические параметры.

Входные данные: пространственно-временные данные (например, метеорологические измерения с различных станций).

Преимущества:

- Учитывает пространственные связи между регионами.
- Эффективна при анализе сложных многомерных данных.

Недостатки:

- Высокая вычислительная сложность.
- Требует качественных пространственных данных для корректной работы.

Autoformer — это трансформерная архитектура, специально разработанная для анализа временных рядов. Её ключевая особенность заключается в способности эффективно обрабатывать длинные временные последовательности благодаря механизму автоматического выделения важных признаков (Auto-Correlation).

Входные данные: временные ряды (температура, осадки и другие параметры).

Преимущества:

- Эффективен при долгосрочном прогнозировании.
- Устойчив к шуму в данных.

Недостатки:

- Требует значительных вычислительных ресурсов.
- Может быть избыточным для простых задач.

PatchTST — это модель на основе трансформеров, которая разделяет временные ряды на небольшие «патчи» (фрагменты) для более детального анализа локальных зависимостей. Такой подход позволяет повысить точность краткосрочных прогнозов за счёт фокусировки на локальных особенностях данных.

Входные данные: временные ряды (например, температура или влажность за короткие промежутки времени).

Преимущества:

- Отлично справляется с локальными изменениями в данных.
- Высокая точность краткосрочных прогнозов.

Недостатки:

- Ограниченная способность к долгосрочному прогнозированию.
- Требует тщательной настройки параметров модели.

Ниже представлена матрица сравнительного анализа для всех моделей (см). Для проведения анализа были выбраны следующие ключевые характеристики: архитектура, применение, тип входных данных, аппаратные требования, адаптивность, преимущества и недостатки. Эти параметры позволяют всесторонне оценить функциональные и технические особенности каждой модели, что способствует более объективному выбору оптимального решения в зависимости от конкретных потребностей и условий эксплуатации.

Таблица 1. «Сравнение моделей»

Модель	Архитектура	Применение	Тип входных данных	Аппаратные требования	Адаптивность	Преимущества	Недостатки
BERT	Трансформер с двунаправленным анализом	Анализ временных рядов и сложных зависимостей	Временные ряды, текстовые данные	Высокие: требует мощного GPU для обучения	Высокая: может адаптироваться для различных задач обработки последовательностей	Учитывает сложные зависимости в данных благодаря двунаправленному анализу. Универсальность и возможность адаптации для различных задач.	Высокая вычислительная сложность. Требует больших объемов данных для обучения.
LSTM	Рекуррентная нейронная сеть с механизмом долгосрочной памяти	Прогнозирование временных рядов	Временные ряды	Средние: подходит для работы на CPU и GPU	Средняя: требует настройки гиперпараметров для конкретных задач	Эффективно моделирует долгосрочные зависимости. Хорошо справляется с сезонными и временными вариациями.	Склонна к переобучению на небольших наборах данных. Медленное обучение при увеличении объема данных.
Random Forest	Ансамблевый метод, основанный на деревьях решений	Робастные прогнозы, обработка шумных данных	Числовые или категориальные признаки	Низкие: может работать на стандартных процессорах	Низкая: ограничена задачами, не требующими учёта временных зависимостей	Высокая устойчивость к шуму в данных. Простота реализации и интерпретации.	Ограниченная способность работать с временными зависимостями. Менее эффективен для обработки больших объемов данных по сравнению с нейросетевыми моделями.

Модель	Архитектура	Применение	Тип входных данных	Аппаратные требования	Адаптивность	Преимущества	Недостатки
GraphCast	Графовые нейронные сети (GNN)	Пространственно-временной анализ, работа с многомерными данными	Пространственно-временные данные (например, ERA5)	Высокие: требует мощного GPU и большого объема памяти	Высокая: хорошо масштабируется для сложных пространственно-временных задач	Учитывает пространственные связи между регионами. Эффективен при анализе сложных многомерных данных.	Высокая вычислительная сложность. Требует качественных пространственных данных для корректной работы.
Autoformer	Трансформер с механизмом автоматической корреляции (Auto-Correlation)	Долгосрочное прогнозирование временных рядов	Временные ряды	Высокие: требует мощного GPU	Средняя: адаптация возможна, но требует тщательной настройки	- Эффективен при долгосрочном прогнозировании	Требует значительных вычислительных ресурсов. Может быть избыточным для простых задач.
PatchTST	Трансформер, работающий с локальными фрагментами (патчами) временных рядов	Краткосрочное прогнозирование временных рядов	Временные ряды	Средние: подходит для работы на GPU	- Средняя: требует настройки параметров модели	Отлично справляется с локальными изменениями в данных	Ограниченная способность к долгосрочному прогнозированию

Выводы первой главы

Для прогнозирования погоды были выбраны модели BERT, LSTM и Random Forest. Выбор обусловлен их способностью эффективно обрабатывать временные ряды, выявлять сложные зависимости между параметрами и устойчивостью к выбросам и пропускам в данных.

Эти модели менее требовательны к вычислительным ресурсам по сравнению с более сложными архитектурами, что делает их подходящими для задач, где доступ к мощным аппаратным средствам ограничен. Кроме того, они способны работать с различными типами входных данных и адаптироваться к изменениям в исходной информации, что позволило создать гибкую и точную систему прогнозирования, адаптируемую к различным условиям.

2. АНАЛИЗ ДАННЫХ

Для реализации проекта использовались метеорологические данные, предоставленные платформой RP5.ru. Этот источник данных обеспечивает доступ к информации с 2004 года и включает данные от множества метеостанций по всему миру.

Описание данных:

- Источник данных: RP5.ru, разработан компанией "Расписание Погоды" (Санкт-Петербург).
- Период сбора данных: 2005–2024 гг.
- Метеостанция: Москва ВДНХ.
- Параметры данных:
- Температура воздуха .
- Давление на уровне моря и на станции .
- Относительная влажность .
- Направление и скорость ветра .
- Облачность тип облаков , количество облаков .
- Погодные явления .
- Минимальная и максимальная температура .
- Количество осадков и длительность осадков .
- Точка росы , температура почвы .

Пример данных приведен в таблице ниже (см.)

Таблица 2. Пример исходных данных

datetime	T	Po	P	U	DD	Ff	N	WW
2024-01-01 00:00	-1,3	721,8	735,4	0,95	Юго-западный ветер	1,0	1	Снег умеренный

datetime	T	Po	P	U	DD	Ff	N	WW
2024-01-01 03:00	-2,0	723,2	736,9	0,93	Юго-юго-восточный ветер	3,0	1	Снег сильныйй

Перед использованием данных для обучения моделей была проведена их предварительная обработка.

В процессе анализа данных были выявлены пропуски и выбросы, особенно в ключевых параметрах, таких как температура и влажность. Для их устранения применялись методы заполнения пропусков средними или медианными значениями, а аномальные значения корректировались. Категориальные параметры кодировались для дальнейшего использования в моделях машинного обучения.

После очистки данных был проведен статистический анализ, выявивший сильные корреляции между температурой и точкой росы, а также отрицательную корреляцию между температурой и влажностью. Сезонные зависимости показали, что летом температура более вариативна и зависит от времени суток, тогда как зимой распределение температур более стабильно. Влияние времени суток было подтверждено с помощью ANOVA, а сезонные изменения — корреляционным анализом, который выявил высокую корреляцию температуры с временем суток летом и слабую корреляцию зимой.

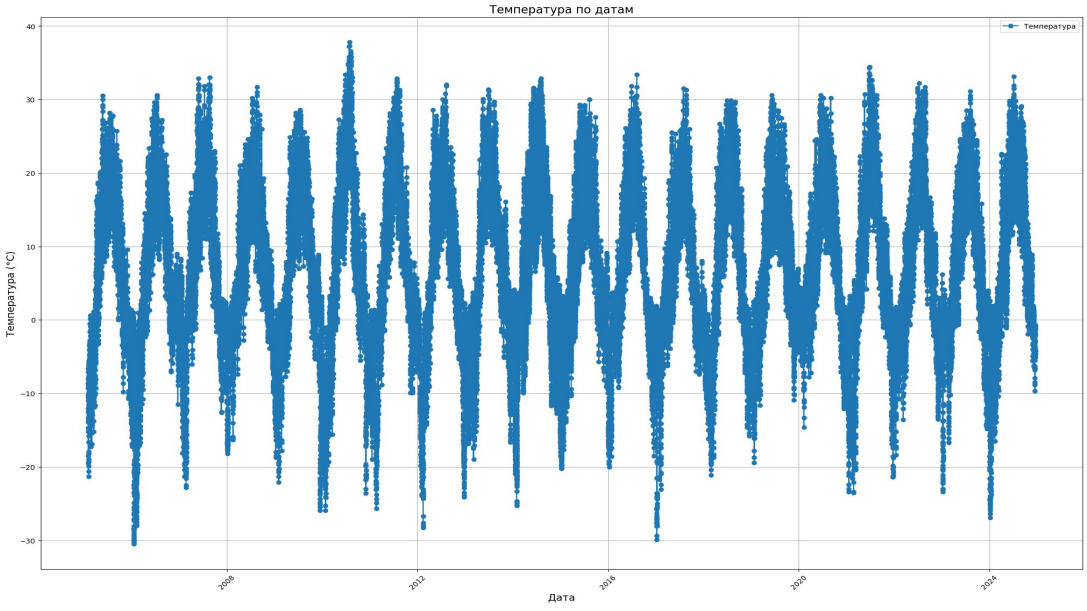

Рисунок 1 Температура по датам

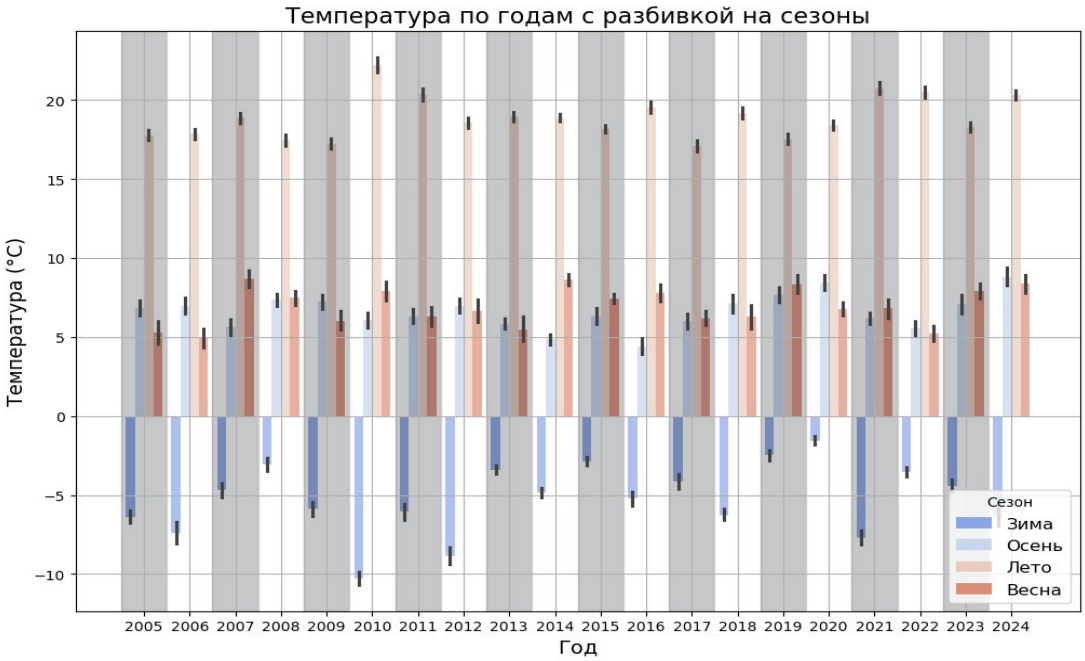

Рисунок 2 Температура по годам с разбивкой на сезоны

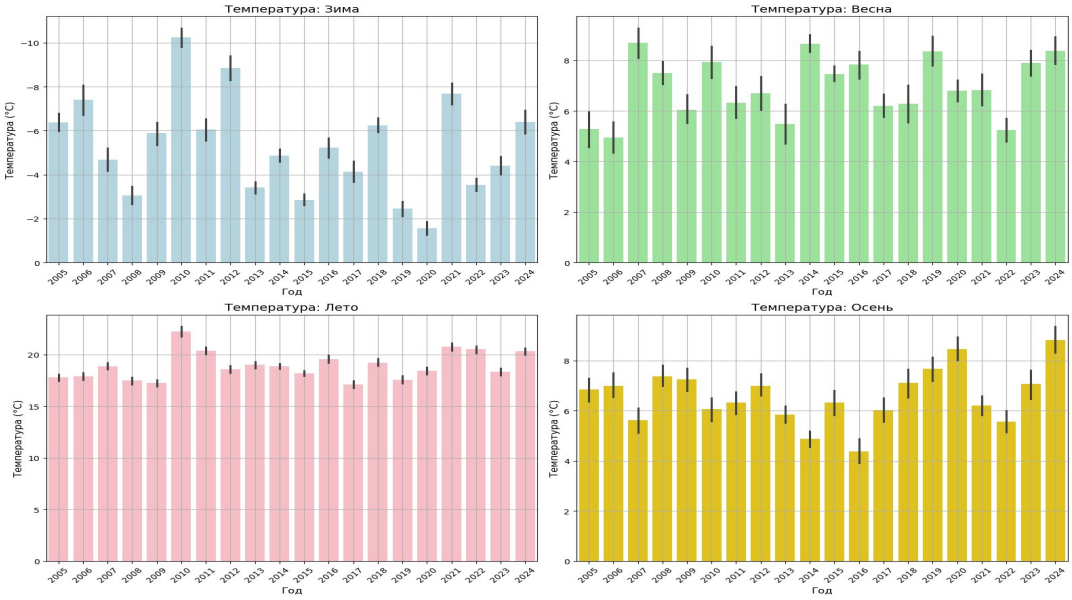

Рисунок 3 Температура по годам для каждого отдельного сезона

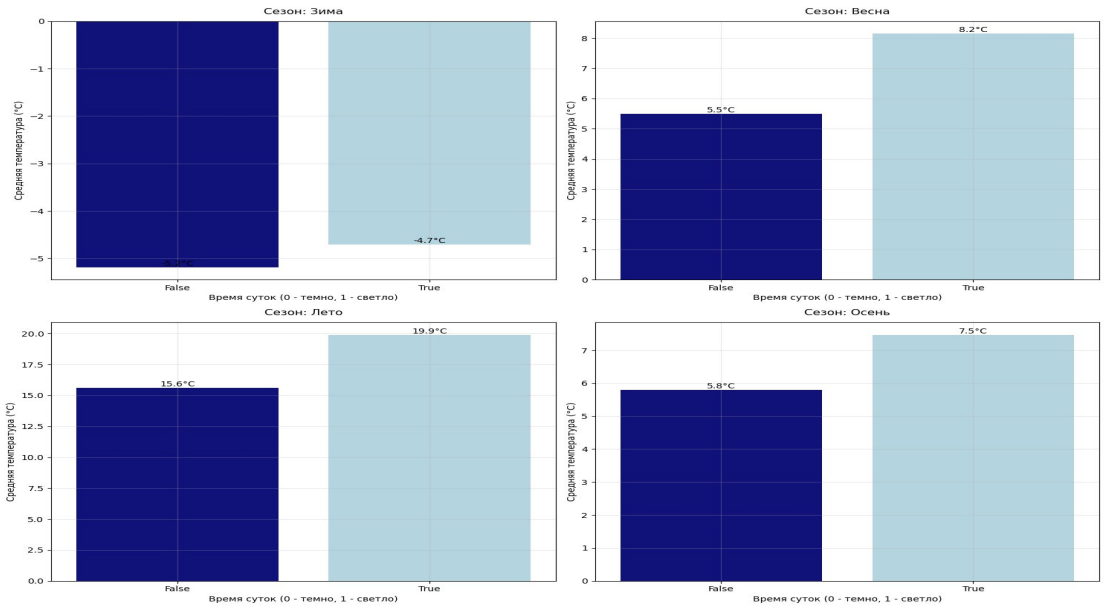

Рисунок 4 Зависимость светового дня для температуры в каждый сезон

Выводы второй главы

1. Анализ показал, что данные содержат значимые зависимости между погодными параметрами, которые могут быть использованы для построения прогнозных моделей:
2. Температура сильно зависит от времени суток и сезона.
3. Влажность и точка росы являются важными факторами для предсказания температуры.
4. Данные требуют тщательной обработки из-за наличия пропусков и выбросов.

Эти результаты стали основой для выбора подходящих моделей машинного обучения и настройки их параметров в дальнейшей реализации проекта.

3. РЕАЛИЗАЦИЯ ПРОЕКТА

Система прогнозирования погоды была разработана с использованием модульной архитектуры, что обеспечивает её гибкость и масштабируемость. Основные компоненты системы включают:

1. Модуль прогнозирования:
    a. Используемые модели:
        i. BERT: трансформер для анализа временных рядов.
        ii. LSTM: рекуррентная нейронная сеть с долгосрочной памятью.
        iii. Random Forest: ансамблевый метод для робастных прогнозов.
    b. Задача модуля — обработка входных данных и генерация прогноза погодных условий.
2. Модуль обработки естественного языка:
    a. Интеграция с GigaChat, который обрабатывает текстовые запросы пользователей.
    b. Возможность задавать вопросы о погоде в естественной форме (например, «Какая будет погода завтра?»).
3. Telegram-бот:
    a. Пользовательский интерфейс для взаимодействия с системой.
    b. Функционал включает запросы прогноза погоды, получение рекомендаций и обратной связи.
4. Вспомогательные модули:
    a. Обработка данных: функции для очистки, нормализации и подготовки данных.
    b. API для взаимодействия: интерфейс для связи между компонентами системы.

Ниже приведено схематическое представления решения (см.)

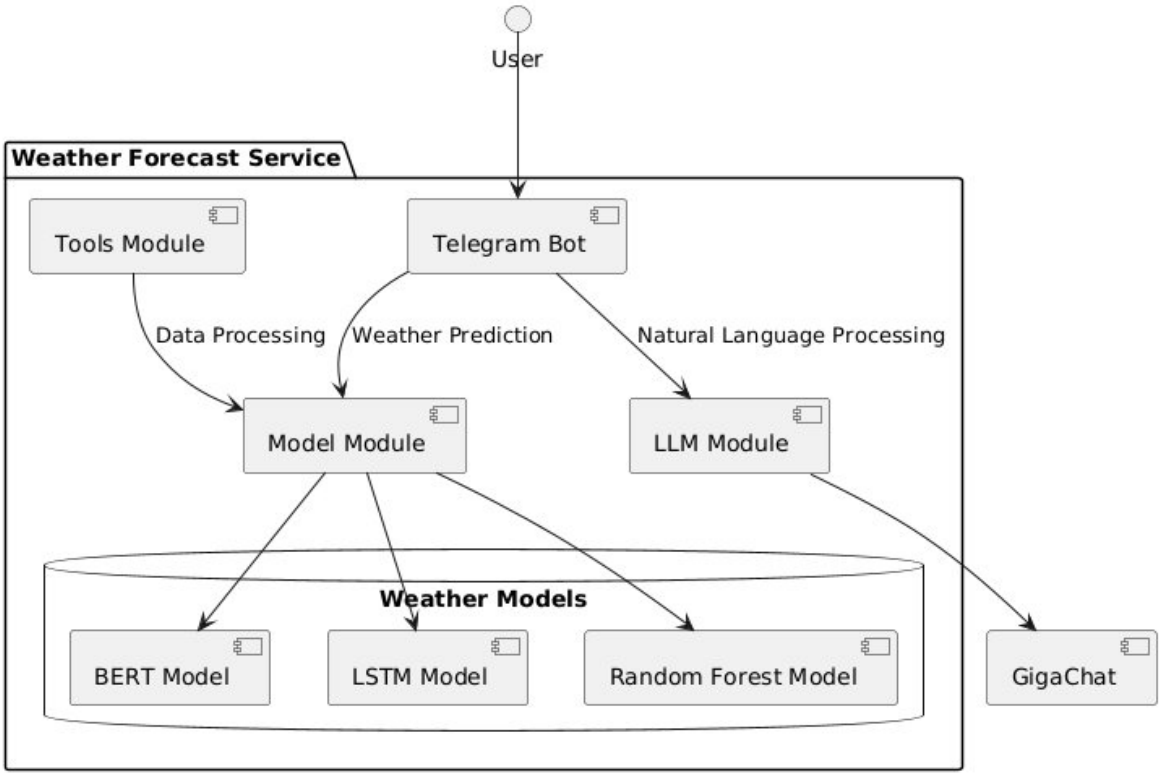

Рисунок 5 Архитектура системы

Для обучения каждой модели проводились эксперименты с настройкой гиперпараметров и оценкой их качества на тестовых данных. В частности, для модели LSTM был выбран размер входного слоя в 18 признаков, включая температуру, влажность и сезонные параметры. Количество скрытых нейронов составляло 64, а модель состояла из двух слоёв. В качестве оптимизатора использовался Adam с learning rate 0.001, а функцией потерь была выбрана MSE (Mean Squared Error).

Модель BERT была обучена на предобученной модели и дообучена на временных рядах. Размер батча составлял 32, а обучение проводилось в течение 10 эпох. Для модели Random Forest были выбраны 100 деревьев с максимальной глубиной дерева 10.

Для оценки качества моделей использовались следующие метрики: MSE (Mean Squared Error) — среднеквадратичная ошибка предсказания, RMSE (Root Mean Square Error) — среднеквадратичное отклонение в градусах, MAE (Mean Absolute Error) — средняя абсолютная ошибка предсказания и R^2 (Coefficient of Determination) — коэффициент детерминации, показывающий долю объяснённой вариации.

Результаты обучения и тестирования моделей показали, что модель BERT продемонстрировала наилучшие результаты по всем метрикам. Для неё значения MSE, RMSE и MAE были наименьшими, а коэффициент детерминации — самым высоким.

Модель LSTM также показала хорошие результаты, но немного уступила BERT. Модель Random Forest показала наихудшие результаты среди всех рассмотренных моделей.

Для интеграции всех компонентов системы была использована библиотека LangChain, обеспечивающая эффективное взаимодействие между различными модулями. В рамках этого подхода был организован процесс маршрутизации запросов от Telegram-бота к моделям прогнозирования и модулю обработки естественного языка. Одним из ключевых аспектов интеграции является автоматический выбор наиболее подходящей модели прогнозирования для каждого конкретного запроса. Этот выбор основывается на анализе типа запроса и доступных данных, что позволяет оптимизировать процесс обработки и повысить точность прогнозов. Запросы пользователей поступают в модуль обработки естественного языка, в частности, в GigaChat, который выполняет их интерпретацию. После анализа GigaChat направляет запрос в модуль прогнозирования. В этом модуле используются различные модели машинного обучения, такие как BERT, LSTM или Random Forest, для формирования прогноза.

В рамках дальнейшего развития проекта предполагается интеграция компьютерного зрения для повышения эффективности классификации одежды по погодным условиям. В частности, планируется использовать алгоритмы распознавания изображений, такие как YOLO, для анализа и классификации одежды в соответствии с прогнозируемыми погодными условиями. Это позволит создать базу данных одежды, содержащую характеристики, необходимые для предоставления персонализированных рекомендаций пользователям.

Одним из ключевых направлений развития проекта является разработка системы Retrieval-Augmented Generation (RAG). RAG представляет собой метод, который объединяет технологии генерации текста с механизмами поиска информации в базах данных. Это позволит создать базу знаний о соответствии одежды различным погодным условиям, что значительно улучшит точность рекомендаций для пользователей. Интеграция RAG-системы с существующим модулем прогнозирования позволит обеспечить более персонализированные рекомендации. Система будет учитывать не только прогноз погоды, но и индивидуальные предпочтения пользователя, а также его предыдущий опыт использования приложения. Это сделает процесс подбора одежды более удобным и эффективным.

Дополнительно планируется улучшить пользовательский опыт путем добавления функции загрузки фотографий в Telegram-бот. Это позволит пользователям анализировать

свою одежду и получать рекомендации на основе текущих погодных условий. Персонализированные рекомендации будут основываться на прогнозах температуры, осадков и силы ветра, что поможет пользователям быть лучше подготовленными к изменениям погоды.

Выводы по третей главе

Разработанная система представляет собой инновационную платформу для прогнозирования погоды с использованием передовых технологий машинного обучения и обработки естественного языка, таких как BERT, LSTM и Random Forest. Интеграция с Telegram-ботом значительно увеличивает её функциональность, обеспечивая оперативное получение и использование данных о погоде. В дальнейшем планируется расширение функционала и улучшение интерфейса, что сделает систему удобной для пользователей и полезной для метеорологических служб.

ЗАКЛЮЧЕНИЕ

В рамках данного проекта была разработана интеллектуальная система прогнозирования погоды с использованием современных технологий искусственного интеллекта и машинного обучения. Все поставленные задачи были успешно решены:

1. Проведен анализ современных моделей для прогнозирования погоды, включающий их сравнение по архитектуре, входным данным, преимуществам и недостаткам. На основе анализа были выбраны модели **BERT**, **LSTM** и **Random Forest**, которые продемонстрировали высокую точность и устойчивость к недостаткам данных.

2. Выполнен сбор и обработка метеорологических данных из источника RP5.ru за период 2005–2024 гг. Проведена очистка данных от пропусков и выбросов, а также выполнен статистический анализ параметров для выявления закономерностей.

3. Обучены выбранные модели прогнозирования с настройкой гиперпараметров. Результаты обучения показали высокие значения метрик качества (MSE, RMSE, MAE, R^2), что подтверждает точность разработанных моделей.

4. Реализована интеграция моделей через библиотеку LangChain, что позволило объединить их в единую систему. Telegram-бот был разработан как пользовательский интерфейс для взаимодействия с системой, а модуль обработки естественного языка на базе GigaChat предоставил возможность задавать запросы в удобной текстовой форме.

5. Система была протестирована на реальных данных, что подтвердило её функциональность и практическую применимость.

Разработанная система представляет собой гибкий инструмент для прогнозирования погодных условий, который может быть адаптирован под различные задачи и условия. В будущем планируется дальнейшее развитие проекта, включая интеграцию компьютерного зрения для анализа изображений, создание RAG-системы для персонализированных рекомендаций по гардеробу и улучшение пользовательского опыта через расширение функционала Telegram-бота.

Таким образом, проект достиг поставленных целей и продемонстрировал потенциал использования современных технологий машинного обучения для решения задач метеорологического прогнозирования.