**Курс: Машинное обучение: фундаментальные инструменты и практики (AML-44)**

Автор: Михаил Лепихин

**Дипломная работа**

Использование методов машинного обучения для автоматизации процессов мебельной мастерской.

Оглавление

[Дипломная работа 0](#_Toc189160541)

[Введение и постановка задачи 2](#_Toc189160542)

[Подготовка датасета 3](#_Toc189160543)

[Обучение модели 6](#_Toc189160544)

[Автоматизация 3D-моделирования 9](#_Toc189160545)

[Реализация системы взаимодействия 10](#_Toc189160546)

[Заключение 12](#_Toc189160547)

[Список литературы 13](#_Toc189160548)

## Введение и постановка задачи

**Аннотация**

Данная работа посвящена разработке системы автоматизации процессов общения с клиентами для мебельной мастерской.

В проекте рассматриваются задачи взаимодействия с клиентами через различные платформы, включая мессенджеры, веб-сайт и платформу «Авито».

Целью работы является создание системы, способной анализировать текстовые запросы клиентов и генерировать на их основе 3D-проекты для дальнейшего утверждения дизайна.

**Основная идея автоматизации**

Основная идея автоматизации заключается в создании интеллектуальной системы, которая способна извлекать ключевые параметры мебели (такие как ширина, высота и глубина) из текстовых запросов клиентов и на их основе автоматически генерировать 3D-модели.

Реализация такой системы поможет значительно ускорить процесс обработки заказов и улучшить качество взаимодействия с клиентами.

**Этапы реализации проекта:**

1. Создание датасета:  
   Поскольку готового набора данных нет, требуется сформировать собственный датасет. Он должен включать как тексты обращений клиентов, так и данные о параметрах мебели.
2. Обучение модели нейросети:  
   Подготовка и обучение модели для классификации именованных сущностей (Named Entity Recognition, NER). Это позволит выделять ключевые параметры мебели (например, размеры) из текстов запросов.
3. Автоматизация 3D-моделирования:  
   Разработка скрипта для автоматического формирования 3D-моделей на основе выделенных параметров. Для реализации выбрана программа SketchUp в качестве инструмента для создания 3D-проектов.
4. Реализация системы взаимодействия:  
   Создание системы, похожей на чат, которая обеспечит взаимодействие клиента с системой и позволит оперативно согласовывать дизайн.

**Результатом работы** должна стать система, способная значительно упростить процесс обработки заказов и снизить трудозатраты сотрудников мастерской.

## Подготовка датасета

В рамках поставленной задачи требовалось подготовить датасет, предназначенный для обучения и тестирования модели. Однако, на начальном этапе доступного набора данных не существовало. В распоряжении находилось лишь ограниченное количество пользовательских запросов, полученных посредством платформы «Авито».

Исходя из этого, было решено сгенерировать дополнительный набор данных с применением одной из современных моделей модели LLM, а именно GigaChat.

Далее описан процесс подготовки датасета.

**Создание исходного набора запросов**

На первом этапе был подготовлен набор из 10–15 запросов реальных запросов клиентов, и набор из 10-15 вымышленных запросов, но написанных человеком. Вымышленные запросы были составлены в том же стиле, что и реальные, чтобы максимально соответствовать их структуре и содержанию.

**Предобработка текстовых данных**

Для улучшения качества данных была осуществлена предварительная обработка текстов:

* Устранены орфографические и грамматические ошибки.
* Проведена токенизация текста (разделение на отдельные элементы — токены).

**Ручная разметка данных**

После предобработки текста была выполнена ручная разметка данных. Для каждого токена вручную были назначены соответствующие NER-теги (Named Entity Recognition), которые отражают принадлежность токена к определённой категории сущностей: ширина, высота илиглубина.

Такой подход является стандартным для задач извлечения из текста классифицированных сущностей. Это позволяет в дальнейшем использовать набор данных для обучения моделей.

Пример:

*Текст запроса клиента:* Здравствуйте! Интересует банкетки такого рода. Длину хотим 120см А глубина стандартная 35см. А по высоте можно 45 сделать?

*Токены:* ['Здравствуйте', '!', 'Интересует', 'банкетки', 'такого', 'рода', '.', 'Длину', 'хотим', '120', 'см', 'А', 'глубина', 'стандартная', '35', 'см.', 'А', 'по', 'высоте', 'можно', '45', 'сделать', '?']

*Теги NER:* ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-WIDTH', 'O', 'O', 'O', 'O', 'B-DEPTH', 'O', 'O', 'O', 'O', 'O', 'B-HIGHT', 'O', 'O']

Из примера видно, что каждому токену сопоставлен NER-тег. В частности, 10-й тег (B-WIDTH) соответствует сущности «ШИРИНА» и относится к токену №10 - «120».

В качестве инструментов разметки использовались программы LabelStudio и MS Excel.

**Формирование базового датасета**

На основе размеченных данных был сформирован базовый датасет в формате json, который впоследствии использовался в качестве примера для генерации новых данных с помощью LLM.

Данный датасет включал 14 размеченных текстов.

**Создание запроса для модели LLM**

Следующим этапом стало создание системы, позволяющей с помощью модели LLM (GigaChat) сгенерировать дополнительный набор размеченных текстов объемом от 500 до 1000 записей.

На вход LLM передавались следующие данные:

* общие настройки LLM,
* параметры, обеспечивающие формирование ответов в необходимом формате,
* примеры ранее созданного датасета, что одновременно решало две задачи:
  + снижало вероятность ошибок LLM в форматировании данных,
  + уменьшало число дубликатов — повторяющихся текстов.

**Разработка автоматизированной системы генерации данных**

Для упрощения процесса генерации большого объёма данных была реализована автоматизированная система взаимодействия с моделью LLM. Система позволила в автоматическом режиме передавать запросы к модели и получать результаты в виде новых размеченных текстов.

В ходе работы были:

* протестированы несколько принципиально различных подходов к структуре информации;
* изучены и применены разные методы работы с памятью LLM;
* протестированы различные версии GigaChat (базовая версия, версия PRO и самая новая версия – MAX).

Таким образом, на основе ограниченного количества исходных данных и с использованием возможностей LLM была создана полноценная структура для генерации датасета. Это позволило не только расширить объём доступных данных, но и повысить их качество за счёт автоматизации и стандартизации процесса подготовки. стандартизации процесса подготовки.

Для оптимизации процесса была разработана автоматизированная система взаимодействия с моделью LLM. Эта система позволяла автоматически отправлять запросы к модели и получать новые размеченные тексты в ответ. В ходе работы над системой были исследованы различные подходы:

* Опробованы разные структуры представления информации.
* Тестировались различные методы управления памятью LLM.
* Были использованы три версии GigaChat: базовая, PRO и новейшая MAX.

**Итог:**

1. Поучилось сгенерировать датасен на 308 уникальных записи и всего 540 записей.
2. Генерация производилась пачками от 2 до 10 записей.
3. Более 10 записей модель сгенерировать не может, т.к. из-за длины текста модель теряет фокус и генерирует несвязанный текст.
4. В процессе генерации более 50% - дубли.

Было сгенерировано 540 записей, из которых 308 уникальных. Дальнейшаяя генерация создавала всё большее количество дублей, что сделало процесс нецелесообразным

## Обучение модели

В качестве базы для создания нейросети была выбрана модель **BERT** (Bidirectional Encoder Representations from Transformers) от **DeepPavlov**. Данная модель была выбрана в силу её высокой эффективности для обработки русскоязычных текстов решении задач обработки естественного языка (NLP). Основные характеристики:

* 768 скрытых слоёв;
* 180 миллионов параметров
* обучен на основе русской части Википедии и новостных данных.

Процесс обучения и применения модели был организован с использованием подходов, представленных в библиотеке **HuggingFace**, что позволило обеспечить удобный интерфейс для работы с моделями и их адаптации к конкретным задачам.

Подготовка данных и обучение модели состояли из нескольких этапов:

**Подготовка токенизатора:**

1. Разделение данных на тренировочные, тестовые и валидационные (555, 20 и 14 записей)
2. Привести к единому формату, который будет понятен модели классификации сущностей (NER):
   1. каждая запись состоит из двух массивов: массив токенов и массив тегов-NER
   2. i-й элемент массива тегов-NER соответствует i-й элемент массива токенов.
3. Токенизация (методология HuggingFace подразумевает, что у каждой модели свой токенизатор).

При этом, нужно учитывать алгоритм работы модели. Он подразумевает, что если одной сущности соответстует несколько токенов, то для первого токена проставляется в соответствие тег NER с меткой B- (begin), а последующие меткой I- (inner).

*Пример:* Предложение "ширина 1058 мм" соответствует тегам NER ["O", "B-WIDTH", "O"]. Но после токенизации, мы получим массив токенов ['[CLS]', 'ширина', '105', '##8', 'мм', '[SEP]'], т.е. размер 1058 был разбит на два токена (105 и 8) + добавились служебные токены CLS и SEP, которые тоже нужно обработать. В итоге, для данного массива токенов нужно будет сформировать следующий массив тегов NER: [-100, "O", "B-WIDTH", "I-WIDTH", "O", -100]

**Подготовка к обучению модели:**

- Подготовка функции объединения строк датасета. Поскольку модель обучается на данных, передаваемых в виде пакетов (батчи), а длина строк в батчах может различаться, требуется подготовить объект DataCollator, который будет формировать батчи одинаковой длины. выравнивая записи в батчах (batch alignment);

- Подготовка функции-загрузчика данных в модель.

- Определить функцию потерь. В данном случае была выбрана AdamW, т.к. она наиболее подходит для данного вида задачи (классификация токенов).

- Подготовить функцию оценки качества модели (показатели accuracy, F1-score и другие).

**Обучение модели**

Обучение модели проводилось на 10 эпохах, в результате была получена точность F1-score от 73% до 100% на тестовой выборке, в зависимости от NER-тега:

* + WIDTH 73%
  + HEIGHT 100%
  + DEPTH 73%

**Результат.**

Проверка модели на произвольном тексте.

*Текст:* **“***Здравствуйте! Я бы заказал у вас банкетку.*

*Только не знаю какие размеры подобрать. У меня ширина пройма 120 см. А глубина 40 или 35. А вот по высоте не очень понимаю как лучше сделать.*

*Либо 45, либо 50см. Что посоветуете? Может быть 120х35х45см?*”

*Результат классификации:*

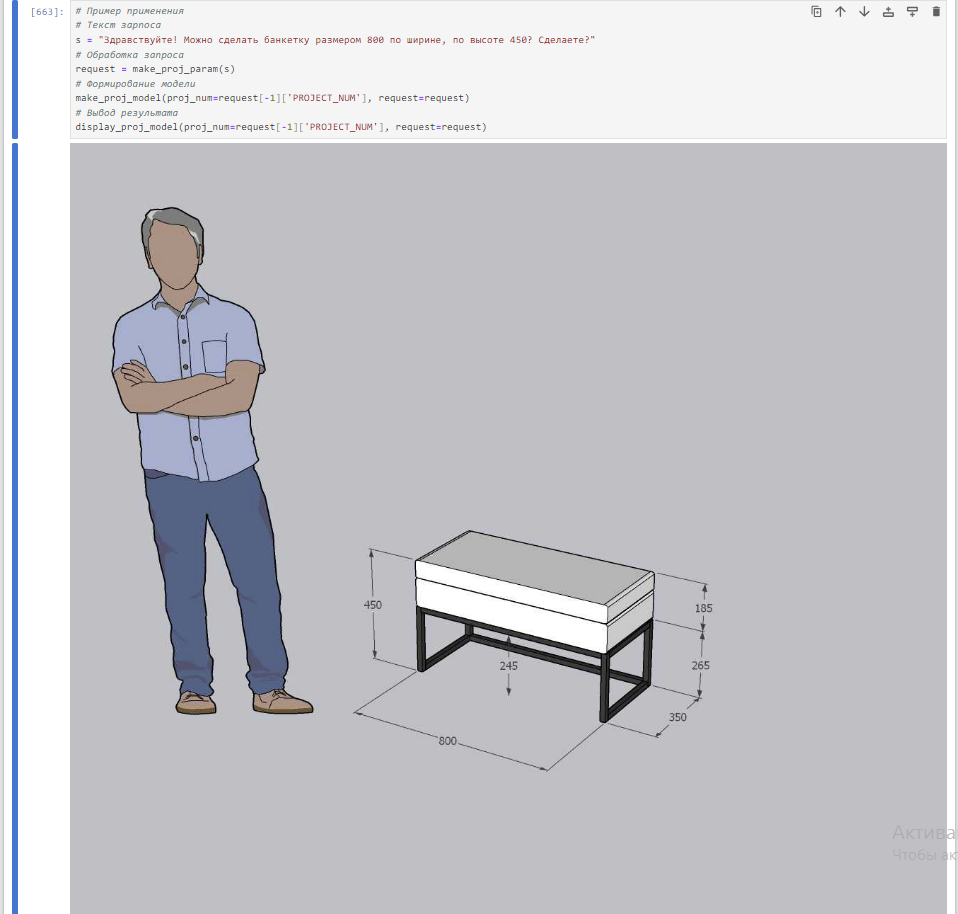
|  |  |
| --- | --- |
| **word** | **ner** |
| Здравствуйте | O |
| ! | O |
| Я | O |
| бы | O |
| заказал | O |
| у | O |
| вас | O |
| банкетку | O |
| . | O |
| Только | O |
| не | O |
| знаю | O |
| какие | O |
| размеры | O |
| подобрать | O |
| . | O |
| У | O |
| меня | O |
| ширина | O |
| пройма | O |
| 120 | WIDTH |
| см. | O |
| А | O |
| глубина | O |
| 40 | DEPTH |
| или | O |
| 35 | DEPTH |
| . | O |
| А | O |
| вот | O |
| по | O |
| высоте | O |
| не | O |
| очень | O |
| понимаю | O |
| как | O |
| лучше | O |
| сделать | O |
| . | O |
| Либо | O |
| 45. | HEIGHT |
| либо | O |
| 50 | HEIGHT |
| см. | O |
| Что | O |
| посоветуете | O |
| ? | O |
| Может | O |
| быть | O |
| 120 | WIDTH |
| х | O |
| 35 | DEPTH |
| х | O |
| 45 | HEIGHT |
| см | O |
| ? | O |

## Автоматизация 3D-моделирования

Программа для 3D-проектирования SketchUp включает механизм автоматизации процесса создания 3D-моделей через скрипты на языке Ruby.

Был подготовлен скрипт, который принимает на вход JSON-файл с параметрами и генерирует 3D-модель вместе с набором изображений.

Объединение этого скрипта с моделью классификации токенов, дало следующий результат.



## Реализация системы взаимодействия

**Создание интерфейса системы «запрос-ответ»**

Для обеспечения эффективного взаимодействия клиента с системой разработан интерфейс, позволяющий оперативно согласовывать дизайн изделия. Основой интерфейса станет чат-бот, который будет выполнять следующие функции:

* Общение с клиентом для уточнения параметров изделия.
* Формирование полного набора параметров для последующего создания 3D модели.
* Генерация набора изображений для согласования с клиентом.

Чат-бот будет реализован на основе языковой модели GigaChat, что обеспечит высокую точность обработки запросов и удобство взаимодействия.

**Технические параметры чата**

1. **Модель LLM**: GigaChat.
2. **API для взаимодействия с LLM**: LangChain.
3. **Механизм памяти**: Для повышения удобства работы с клиентом используется механизм памяти, который позволяет системе запоминать контекст взаимодействия.
4. **Архитектура чата**: Чат будет построен с применением графовой структуры (LangGraph), что позволит разбить функционал системы на отдельные модули и упростить управление процессами.

**Алгоритм работы чата**

1. **Парсинг сообщений пользователя**: Чат-бот анализирует сообщения клиента для извлечения ключевых параметров.
2. **Сбор данных**: Если данных для построения модели недостаточно, система продолжает задавать уточняющие вопросы до получения всех необходимых параметров.
3. **Создание 3D модели**: После сбора полного набора данных запускается процесс генерации 3D модели изделия.
4. **Отображение результатов**: Система формирует набор изображений модели для согласования с клиентом.

**Демонстрация работы:**

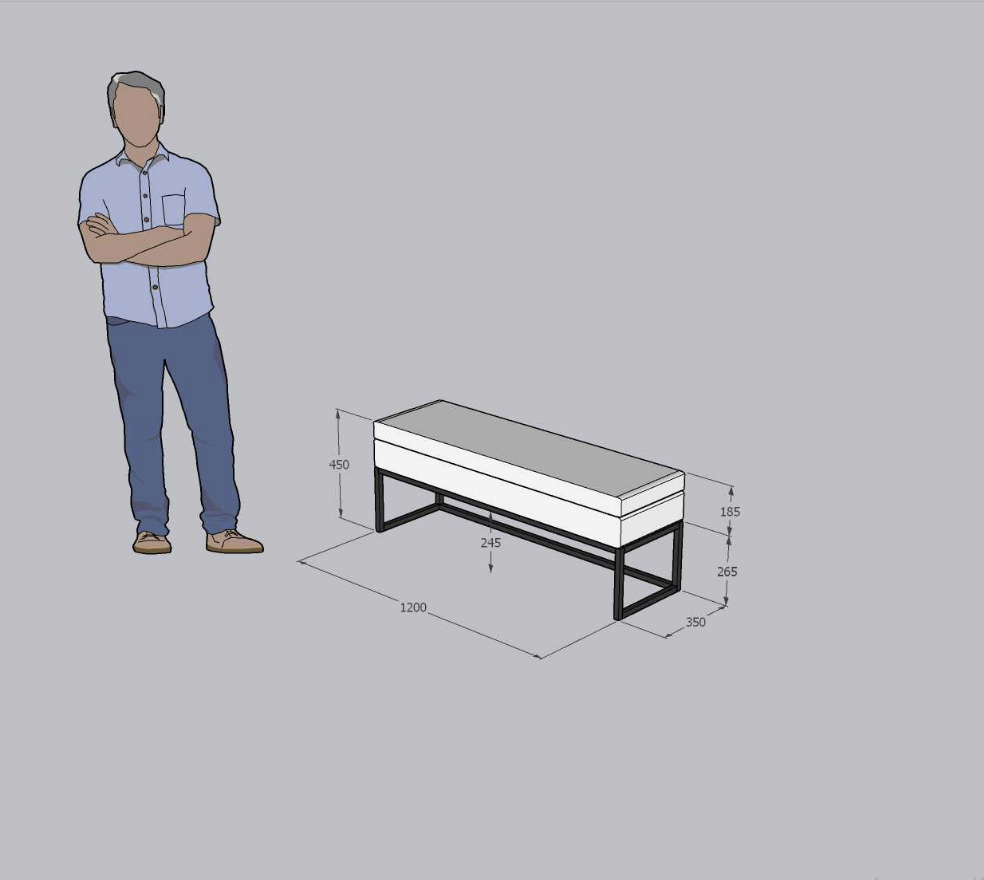
*Сообщение:* Здравствуйте!

*Ответ*: Добрый день! Чем я могу помочь вам?

*Сообщение:* Я бы хотел заказать банкетку

*Ответ:* Отлично! Для оформления заказа мне необходимо уточнить несколько деталей. Какой размер банкетки вас интересует? Укажите, пожалуйста, желаемую ширину, глубину и высоту.

*Сообщение:* Ширина 1200, высота 450, глубина 350



*Ответ:* Спасибо за информацию! Все параметры соответствуют стандартным размерам материалов, поэтому никаких проблем с изготовлением не возникнет. Могу я оформить ваш заказ?

*Сообщение*: Завершить

## Заключение

**Основные результаты работы**

1. **Автоматизация обработки запросов**: В рамках дипломной работы была разработана система, которая позволяет автоматически извлекать ключевые параметры мебели из текстовых запросов клиентов.
2. **Создание датасета**: Для обучения модели был подготовлен уникальный датасет, включающий реальные и сгенерированные запросы клиентов. Датасет прошёл этапы предобработки, ручной разметки и дополнения с использованием языковой модели GigaChat.
3. **Обучение модели NER**: Разработанная модель для классификации именованных сущностей (NER) успешно справляется с выделением ключевых параметров мебели (ширина, высота, глубина) из текстов запросов.
4. **Автоматизация 3D-моделирования**: Был реализован скрипт для автоматического создания 3D-моделей на основе выделенных параметров. В качестве инструмента для 3D-проектирования использовалась программа SketchUp.
5. **Реализация системы взаимодействия**: Разработан интерфейс системы «запрос-ответ», основанный на интеграции чат-бота с языковой моделью GigaChat. Это обеспечивает удобное взаимодействие клиента с системой и оперативное согласование дизайна.

**Практическая значимость**

Реализованная система автоматизации предоставляет следующие преимущества:

* Сокращение времени на обработку клиентских запросов.
* Повышение точности извлечения данных из текстов.
* Упрощение процесса согласования дизайна мебели.
* Снижение трудозатрат сотрудников мастерской.

**Перспективы развития**

1. **Расширение функционала системы**: Добавление новых параметров для извлечения (например, моделей, материалов, стиля или дополнительных деталей мебели).
2. **Улучшение качества 3D-моделирования**: Интеграция с более сложными инструментами для создания фотореалистичных моделей.
3. **Масштабируемость**: Адаптация системы для работы с другими сущностями или бизнес-направлениями.

Разработанная система демонстрирует успешное применение методов машинного обучения и обработки естественного языка для автоматизации бизнес-процессов.

## Список литературы

1. HuggingFace. Hugging Face Inc. Доступно на: https://huggingface.co

2. LangChain. Доступно на: https://www.langchain.com

3. DeepPavlov. Доступно на: <http://deeppavlov.ai>

4. GigaChat (Сбер). Доступно на: https://sber.ru