Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ Робототехника и комплексная автоматизация (РК)

КАФЕДРА Системы автоматизированного проектирования (РК6)

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

«Обзор архитектуры и инструментов для разработки чат-бота банковского приложения»

Студент РК6-71Б _____ **Гунько Н.М.**
(Подпись, дата) И.О. Фамилия

Руководитель _____ **Витюков Ф.А.**
(Подпись, дата) И.О. Фамилия

2022 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ
Заведующий кафедрой РК6
А.П. Карпенко

«_____» _____ 2022 г.

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме: Обзор архитектуры и инструментов для разработки чат-бота банковского приложения

Студент группы РК6-71Б

Гунько Никита Макарович
(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.) учебная
Источник тематики (кафедра, предприятие, НИР) предприятие

График выполнения НИР: 25% к 5 нед., 50% к 11 нед., 75% к 14 нед., 100% к 16 нед.

Техническое задание: Исследовать архитектуру и алгоритм работы чат-ботов, определить основные этапы разработки чат-ботов, выявить возможные варианты использования чат-ботов в банковской сфере и выполнить сравнение существующих решений в банковской сфере.

Оформление научно-исследовательской работы:
Расчетно-пояснительная записка на 37 листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.):
_____
_____
_____
_____

Дата выдачи задания «15» сентября 2022 г.

Руководитель НИР _____ **Витюков Ф.А.**
(Подпись, дата) И.О. Фамилия

Студент _____ **Гунько Н.М.**
(Подпись, дата) И.О. Фамилия

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В наше время, эпоху всеобщего удаленного доступа, большинство людей при использовании банковских услуг сталкиваются с рядом проблем, в частности:

- Опасность заражения в периоды повышения уровня заболеваемости каким-либо вирусом;
- Удаленность банковского отделения от места жительства в некоторых регионах;
- Длительное время ожидания;
- Несоответствие: разные агенты по обслуживанию клиентов дают разные ответы.

Многие из приведенных проблем может решить дистанционное обслуживание клиентов. Банковский виртуальный собеседник или другими словами чат-бот – одно из таких решений, которое автоматизирует процесс взаимодействия с клиентами.

Актуальность темы обусловлена текущей эпидемиологической ситуацией в мире и потребностью предприятия ООО «ЮБС» для автоматизации процесса обработки обращений клиентов в чаты, для увеличения эффективности работы сотрудников и сокращения ресурсных затрат, а также для повышения клиентского сервиса предприятия.

Цель работы: исследовать архитектуру и алгоритм работы чат-ботов, определить основные этапы разработки чат-ботов, выявить возможные варианты использования чат-ботов в банковской сфере и выполнить сравнение существующих решений в банковской сфере.

Для выполнения поставленной цели необходимо решение следующих задач:

- изучить понятия чат-бота и его функций;
- определить роль машинного обучения в разработке чат-бота;
- рассмотреть возможные варианты использования чат-бота в банковском

приложении;

- проанализировать существующие решения чат-ботов в банковском приложении;

1. Архитектура и алгоритм работы чат-ботов

1.1 Обзор концепций компьютерного бота

Бот – это виртуальный робот или искусственный интеллект, работающий по набору алгоритмов, который описан в виде компьютерной программы. Он автоматически выполняет определенные задачи, заложенные в него разработчиком.

Существует большое количество разновидностей ботов, которые отличаются наборами выполняемых задач: чат-боты, которые имитируют разговор с человеком, боты для совершения покупок, которые осуществляют отслеживание цен и выполняют поиск лучшей цены на продукты, интересные пользователю, боты-поисковики, боты-загрузчики и т.д.

Можно выделить следующие плюсы использования компьютерных и интернет-ботов:

- Быстрее людей выполняют повторяющиеся задачи;
- Доступны круглосуточно (24/7);
- Приложения для обмена сообщениями позволяют компаниям общаться с большим количеством людей;

Есть и ряд минусов использования компьютерных и интернет-ботов:

- Ботов нельзя настроить для выполнения определенных задач, в которых есть риск неправильно понять пользователей и вызвать у них разочарование в процессе;
- Для управления ботами по-прежнему требуются люди. Также участие человека необходимо в случае возникновения непонимания;
- Боты могут быть запрограммированы на совершение вредоносных действий;
- Ботов можно использовать для рассылки спама.

В данной научной работе будут рассматриваться конкретно чат-боты.

1.2 Основные определения и классификация

Чат-бот – разновидность ботов, программа, которая имитирует реальный разговор с пользователем. Они позволяют общаться с помощью текстовых или аудио сообщений на сайтах, в мессенджерах, мобильных приложениях или по телефону.

Чат-боты – это специальные аккаунты, за которыми не закреплен какой-либо человек, а сообщения, отправленные с них или на них, обрабатываются внешней системой. Кроме того, для пользователя общение с ботом выглядит как обычная переписка с реальным человеком [9].

Чат-боты помогают автоматизировать некоторые задачи, работая по заданному алгоритму. Первые программы, имитирующие общение людей, появились в 1966 году. Виртуальный собеседник Elisa достаточно убедительно пародировал диалог с психотерапевтом. С ростом популярности мессенджеров в 2010-х годах чат-боты обрели новую жизнь. Большинство работает на платформах популярных мессенджеров: Facebook, Telegram, WhatsApp, “ВКонтакте” и другие [1].

Можно выделить два общих типа классификации чат-ботов: бизнес-классификация и классификация чат-бот приложений по техническому типу. Диаграмма бизнес-классификации чат-ботов приведена на рисунке 1.

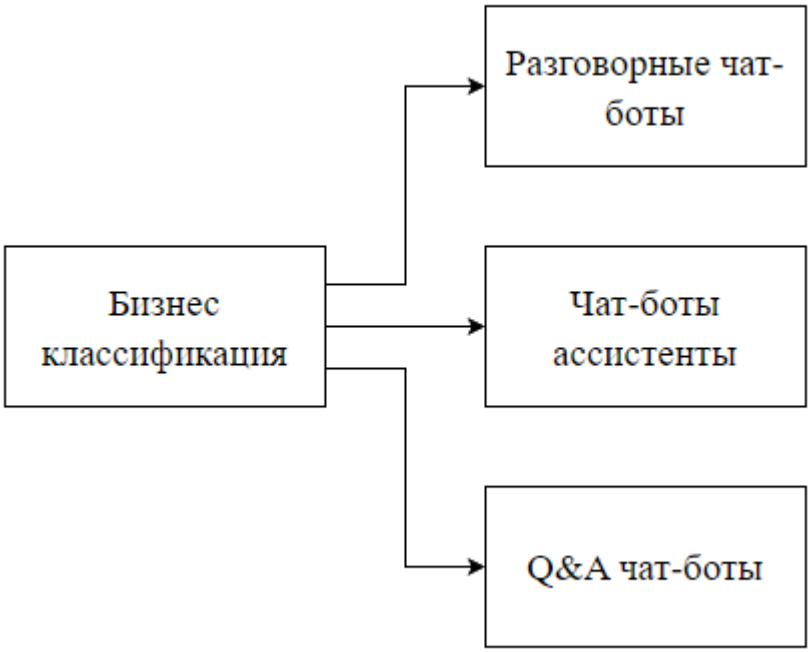


Рисунок 1 – Бизнес классификация чат-ботов

Рассмотрим каждый из типов более подробно:

1) Разговорные чат-боты созданы для общения подобно разговору с человеком. Не имеют конкретной цели;

2) Чат-боты ассистенты имеют конкретную, заранее определенную цель. Из пользовательских сообщений выделяются данные, которые используются для достижения определенных целей. Могут служить заменой или помощниками (ассистентами) в получении банковской выписки или подбора выгодного кредита on-line;

3) Q&A (question and answer) чат-боты, созданные для ответа на вопросы по принципу “1 вопрос – 1 ответ”. Могут служить заменой FAQ раздела различных сайтов.

Диаграмма с технической классификацией чат-ботов приведена на рисунке 2.

Рисунок 2 – Техническая классификация чат-ботов

Чат-ботов разделяют по алгоритму работы на *ограниченных*, *саморазвивающихся* и *гибридных*. Ограниченные (простые) чат-боты взаимодействуют с пользователем на основе запрограммированных сценариев с множественным выбором. Они имеют ограниченные возможности и обычно называются ботами, основанными на правилах. Например, опция А ведет к опции В и так далее. Таких чат-ботов легче создавать, потому что зачастую они используют простой алгоритм true-false для понимания запросов пользователей и предоставления соответствующих ответов. Недостатком таких чат-ботов является то, что они не могут отвечать ни на какие вопросы, выходящие за рамки установленных правил. Также они не обучаются посредством взаимодействия с пользователями [2].

В основе саморазвивающихся чат-ботов лежит искусственный интеллект, который “понимает” контекст и цель вопроса, прежде чем формулировать ответ. Такие компьютерные ассистенты используют машинное обучение для выявления моделей общения. Благодаря постоянному

взаимодействию с людьми они учатся подражать реальным разговорам и реагируют на устные или письменные запросы, помогая найти ответы. Поскольку чат-боты используют искусственный интеллект, то понимают язык, а не просто команды. Таким образом, после каждого диалога они становятся умнее и лучше взаимодействуют с пользователями [3].

Третья группа виртуальных помощников – гибридные. Они представляют из себя комбинацию простых и умных чат-ботов. И простые, и умные чат-боты являются крайностями в спектре чат-ботов. Постоянно будет потребность в том, чтобы простые чат-боты были умнее, а умные чат-боты – проще. Гибридные чат-боты соответствуют этой золотой середине. У гибридных чат-ботов есть некоторые задачи, основанные на правилах, и они могут понимать намерения и контекст. Это делает их сбалансированным инструментом взаимодействия бизнеса с клиентами [4].

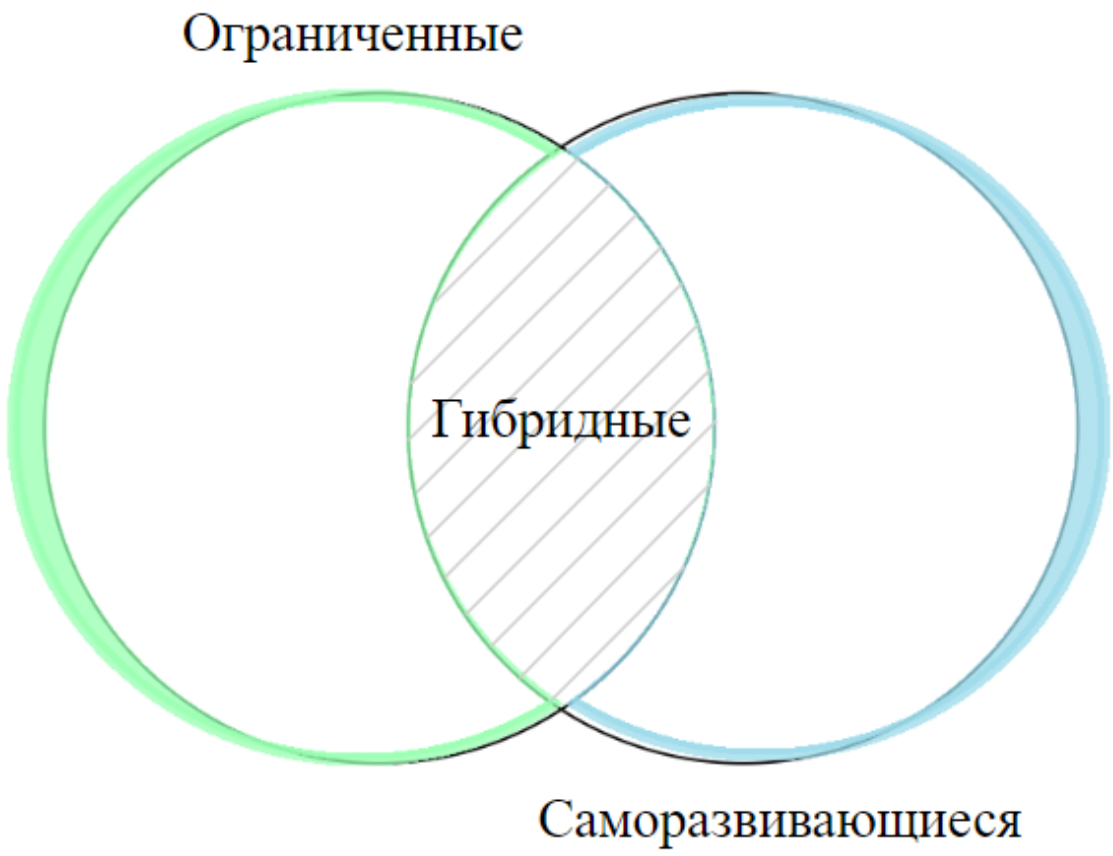


Рисунок 3 – Гибридные чат-боты

Умные, а также гибридные чат-боты используют в своей логике алгоритмы обработки и понимания естественного языка, а также алгоритмы генерации ответа, каждый из которых использует методы машинного обучения.

1.3 Обработка естественного языка

Обработка естественного языка (Natural Language Processing) – возникла из компьютерной лингвистики, использует методы из различных дисциплин, таких как информатика, искусственный интеллект, лингвистика и наука о данных, чтобы позволить компьютерам понимать человеческий язык как в письменной, так и в устной форме. В то время как компьютерная лингвистика больше сосредоточена на аспектах языка, обработка естественного языка делает упор на использование машинного обучения и методов глубокого обучения для выполнения таких задач, как языковой перевод или ответы на вопросы. Обработка естественного языка работает, беря неструктурированные данные и преобразовывая их в формат структурированных данных. Это достигается за счет идентификации именованных сущностей (процесс, называемый распознаванием именованных сущностей) и выявления шаблонов слов с использованием таких методов, как токенизация, выделение корней и лемматизация, которые исследуют корневые формы слов.

1.4 Понимание естественного языка

Понимание естественного языка (Natural Language Understanding) – это часть обработки естественного языка, которая использует синтаксический и семантический анализ текста и речи для определения значения предложения. Синтаксис относится к грамматической структуре предложения, а семантика указывает на его предполагаемое значение. NLU также устанавливает соответствующую онтологию: структуру данных, которая определяет отношения между словами и фразами. В то время как люди, естественно, делают это во время разговора, комбинация этих анализов требуется для того, чтобы машина понимала предполагаемое значение различных текстов. Наша

способность различать омонимы и омофоны хорошо иллюстрирует нюансы языка.

Например, возьмем следующие два предложения:

1. Алиса плывет против течения;
2. Текущая версия отчета находится в папке.

В первом предложении слово течение является существительным. Глагол, который предшествует ему, плавать, предоставляет читателю дополнительный контекст, позволяя нам сделать вывод о том, что мы имеем в виду течение воды в водоеме. Во втором предложении слово текущая используется, но как прилагательное. Описываемое им существительное, версия, обозначает несколько итераций отчета, что позволяет нам определить, что мы имеем в виду наиболее актуальный статус файла.

Эти подходы также широко используются в интеллектуальном анализе данных, чтобы понять отношение потребителей. В частности, анализ настроений позволяет брендам более внимательно отслеживать отзывы своих клиентов, позволяя им группировать положительные и отрицательные комментарии в социальных сетях и отслеживать чистые оценки промоутеров. Просматривая негативные комментарии, компании могут быстрее выявлять и устранять потенциальные проблемные области в своих продуктах или услугах.

1.5 Генерация естественного языка

Генерация естественного языка (Natural Language Generation) – еще одно подмножество обработки естественного языка. В то время как понимание естественного языка сосредоточено на понимании компьютерного чтения, генерация естественного языка позволяет компьютерам писать. NLG — это процесс создания текстового ответа на человеческом языке на основе некоторых входных данных. Этот текст также можно преобразовать в речевой формат с помощью служб преобразования текста в речь. NLG также включает в себя возможности суммирования текста, которые генерируют сводки из входящих документов, сохраняя при этом целостность информации.

Как и в случае с NLU, приложения NLG должны учитывать языковые правила, основанные на морфологии, лексике, синтаксисе и семантике, чтобы сделать выбор в отношении того, как правильно формулировать ответы. Они решают эту задачу в три этапа:

- Планирование текста: на этом этапе формулируется и логически упорядочивается общее содержание;
- Планирование предложений: на этом этапе учитываются пунктуация и поток текста, разбивка содержания на абзацы и предложения и включение местоимений или союзов, где это уместно;
- Реализация: этот этап учитывает грамматическую точность, гарантируя соблюдение правил пунктуации и спряжения.

1.6 Машинное обучение

1.6.1 Основные принципы

Машинное обучение (Machine Learning – ML) – это использование математических моделей данных, которые помогают компьютеру обучаться без непосредственных инструкций. Оно считается одной из форм искусственного интеллекта. При машинном обучении с помощью алгоритмов выявляются закономерности в данных. На основе этих закономерностей создается модель данных для прогнозирования новых, не встречавшихся ранее случаев, исход которых неизвестен. Чем больше данных обрабатывает такая модель и чем дольше она используется, тем точнее становятся результаты. Это очень похоже на то, как человек оттачивает навыки на практике [7].

1.6.2 Взаимосвязь в рамках технологий ИИ

Как мы уже выяснили, машинное обучение считается одной из форм искусственного интеллекта. В дискуссиях об искусственном интеллекте вообще и о машинном обучении в частности обычно смешиваются нейросети, машинное и глубокое обучение.

Рисунок 4 – Иерархия терминов из области искусственного интеллекта

- Нейросети – один из видов машинного обучения.
- Глубокое обучение – это один из видов архитектуры нейросетей.

Рисунок 5 – Иерархия машинного обучения в рамках технологий искусственного интеллекта

1.6.3 Основные виды

1.6.3.1 Классическое обучение

Это простейшие алгоритмы, которые являются прямыми наследниками вычислительных машин 1950-х годов. Они изначально решали формальные задачи – такие, как поиск закономерностей в расчетах и вычисление траектории объектов. Сегодня алгоритмы на базе классического обучения – самые распространенные. Именно они формируют блок рекомендаций на многих платформах.

Классическое обучение подразделяется на следующие типы:

Обучение с учителем – когда у машины есть некий учитель, который знает, какой ответ правильный. Это значит, что исходные данные уже размечены (отсортированы) нужным образом, и машине остается лишь определить объект с нужным признаком или вычислить результат.

Такие модели используют в спам-фильтрах, распознавании языков и рукописного текста, выявлении мошеннических операций, расчете финансовых показателей, скоринге при выдаче кредита. В медицинской диагностике классификация помогает выявлять аномалии – то есть возможные признаки заболеваний на снимках пациентов.

Обучение без учителя – когда машина сама должна найти среди хаотичных данных верное решение и отсортировать объекты по неизвестным признакам. Например, определить, где на фото собака.

Эта модель возникла в 1990-х годах и на практике используется гораздо реже. Ее применяют для данных, которые просто невозможно разметить из-за их колоссального объема. Такие алгоритмы применяют для риск-менеджмента, сжатия изображений, объединения близких точек на карте, сегментации рынка, прогноза акций и распродаж в ретейле, мерчендайзинга. По такому принципу работает алгоритм iPhoto, который находит на фотографиях лица (не зная, чьи они) и объединяет их в альбомы.

1.6.3.2 Обучение с подкреплением

Это более сложный вид обучения, где ИИ нужно не просто анализировать данные, а действовать самостоятельно в реальной среде – будь то улица, дом или видеоигра. Задача робота – свести ошибки к минимуму, за что он получает возможность продолжать работу без препятствий и сбоев. Обучение с подкреплением инженеры используют для беспилотников, роботов-пылесосов, торговли на фондовом рынке, управления ресурсами компании. Именно так алгоритму AlphaGo удалось обыграть чемпиона по игре Го: просчитать все возможные комбинации, как в шахматах, здесь было невозможно.

1.6.3.3 Ансамбли

Это группы алгоритмов, которые используют сразу несколько методов машинного обучения и исправляют ошибки друг друга. Их получают тремя способами:

- Стекинг – когда разные алгоритмы обучают по отдельности, а потом передают их результаты на вход последнему, который и принимает решение;
- Беггинг – когда один алгоритм многократно обучают на случайных выборках, а потом усредняют ответы;
- Бустинг – когда алгоритмы обучают последовательно, при этом каждый обращает особое внимание на ошибки предыдущего.

Ансамбли работают в поисковых системах, компьютерном зрении, распознавании лиц и других объектов.

1.6.3.4 Нейронные сети и глубокое обучение

Самый сложный уровень обучения ИИ. Нейросети моделируют работу человеческого мозга, который состоит из нейронов, постоянно формирующих между собой новые связи. Очень условно можно определить их как сеть со множеством входов и одним выходом.

Нейроны образуют слои, через которые последовательно проходит сигнал. Все это соединено нейронными связями – каналами, по которым

передаются данные. У каждого канала свой «вес» – параметр, который влияет на данные, которые он передает.

ИИ собирает данные со всех входов, оценивая их вес по заданным параметрами, затем выполняет нужное действие и выдает результат. Сначала он получается случайным, но затем через множество циклов становится все более точным.

Хорошо обученная нейросеть работает, как обычный алгоритм или точнее. Настоящим прорывом в этой области стало *глубокое обучение*, которое обучает нейросети на нескольких уровнях абстракций.

Здесь используют две главных архитектуры:

- *Сверточные нейросети* первыми научились распознавать неразмеченные изображения – самые сложные объекты для ИИ. Для этого они разбивают их на блоки, определяют в каждом доминирующие линии и сравнивают с другими изображениями нужного объекта;
- *Рекуррентные нейросети* отвечают за распознавание текста и речи. Они выявляют в них последовательности и связывают каждую единицу – букву или звук – с остальными.

Нейросети с глубоким обучением требуют огромных массивов данных и технических ресурсов. Именно они лежат в основе машинного перевода, чат-ботов и голосовых помощников, создают музыку и дипфейки, обрабатывают фото и видео [10].

1.6.4 Классы задач

Регрессия – это прогнозирование числового значения на основе выборки объектов с различными признаками. Например, оценка платёжеспособности заёмщика, ожидаемого дохода компании или цены квартиры на рынке недвижимости.

Классификация – отнесение объектов на основе имеющихся параметров к одному из предопределённых классов. В рамках работы “Центра изучения и сетевого мониторинга молодёжи” именно качественная классификация помогает выявить деструктивный контент среди текстовых или визуальных

объектов. Ежедневно благодаря машинному обучению анализируется более миллиона изображений и текстов.

Кластеризация – объединение похожих данных в группы (кластеры). Например, поиск сообществ, похожих по контенту, или объединение схожих по смыслу постов в социальной сети.

Прогнозирование временного ряда – работа с данными, полученными в определённый период времени, и предсказание на их основе значений в задаваемый исследуемый период. Решение этой задачи позволяет спрогнозировать сейсмическую активность или изменение стоимости ценных бумаг.

Также существуют вспомогательные задачи, которые можно решить с помощью машинного обучения – распознавание текста на изображениях, детекция символов, идентификация речи и так далее.

1.6.5 Применение в разработке чат-ботов

Проанализировав основные направления методов машинного обучения, можно сделать вывод, что они все могут быть использованы при разработке чат-ботов.

Обработка естественного языка (NLP) используется для того, чтобы чат-боты могли понимать, интерпретировать и генерировать человекоподобную речь. Методы NLP включают в себя моделирование языка, анализ настроений, распознавание сущностей и тегирование частей речи.

Алгоритмы контролируемого обучения используются для обучения чат-ботов понимать и реагировать на пользовательский ввод. При контролируемом обучении виртуальный ассистент тренируется на наборе данных, содержащем маркированные примеры пользовательских вводов и соответствующих ответов. Чат-бот учится определять закономерности в данных и обобщать их, чтобы делать прогнозы по новым входным данным.

Алгоритмы обучения без учителя могут использоваться для группировки похожих пользовательских данных и объединения их в

категории. Это может помочь чат-ботам определить общие темы и вопросы, которые интересуют пользователей.

Алгоритмы обучения с подкреплением могут использоваться для обучения чат-ботов на основе обратной связи. При обучении с подкреплением помощник получает вознаграждение за действия, которые приводят к желаемому результату, и наказание за действия, которые приводят к нежелательному результату. Это позволяет чат-боту учиться методом проб и ошибок и со временем улучшать свою работу.

В целом, чат-боты опираются на сочетание методов машинного обучения, чтобы понимать и реагировать на входные данные пользователя подобно человеку.

1.6.6 Примеры методов для создания разговорного чат-бота

Для создания разговорного чат-бота, можно использовать следующие методы машинного обучения:

1. Рекуррентные нейронные сети (RNN) – это класс нейронных сетей, которые могут обрабатывать последовательности данных, такие как текст. Примеры моделей на основе RNN, которые могут использоваться для создания чат-ботов, включают в себя:
   - LSTM (Long Short-Term Memory)
   - GRU (Gated Recurrent Unit)
   - BiLSTM (Bidirectional LSTM)
2. Преобразовательные нейронные сети (Transformer) – это класс нейронных сетей, который был создан для обработки последовательностей данных. Примеры моделей на основе Transformer, которые могут использоваться для создания чат-ботов, включают в себя:
   - GPT (Generative Pretrained Transformer)
   - BERT (Bidirectional Encoder Representations from Transformers)
   - T5 (Text-to-Text Transfer Transformer)
3. Seq2Seq модели – это класс моделей, которые используются для пре-

образования одной последовательности в другую последовательность. Примеры моделей Seq2Seq, которые могут использоваться для создания чат-ботов, включают в себя:

- Seq2Seq модель с использованием RNN
- Seq2Seq модель на основе Transformer
- Pointer-generator network

4. Генеративные модели – это класс моделей, которые могут генерировать текст на основе заданного контекста. Примеры моделей генеративных моделей, которые могут использоваться для создания чат-ботов, включают в себя:

- Variational Autoencoder (VAE)
- GAN (Generative Adversarial Network)
- Языковые модели на основе марковских цепей

Это только некоторые из методов машинного обучения, которые могут использоваться для создания чат-ботов. Для каждого конкретного проекта нужно выбрать наиболее подходящие методы в зависимости от требований и целей проекта.

1.6.7 Нейронные сети LSTM и Transformer

Нейронные сети LSTM и Transformer – это два из самых популярных типов нейронных сетей, которые используются для обработки последовательных данных. LSTM была разработана для решения проблемы затухания и взрыва градиента в рекуррентных нейронных сетях, а Transformer – для обработки последовательностей с использованием механизма внимания.

LSTM (Long Short-Term Memory) – это архитектура рекуррентной нейронной сети, которая позволяет сохранять долгосрочные зависимости в данных. Она была разработана в 1997 году Хохрайтером и Шмидхубером.

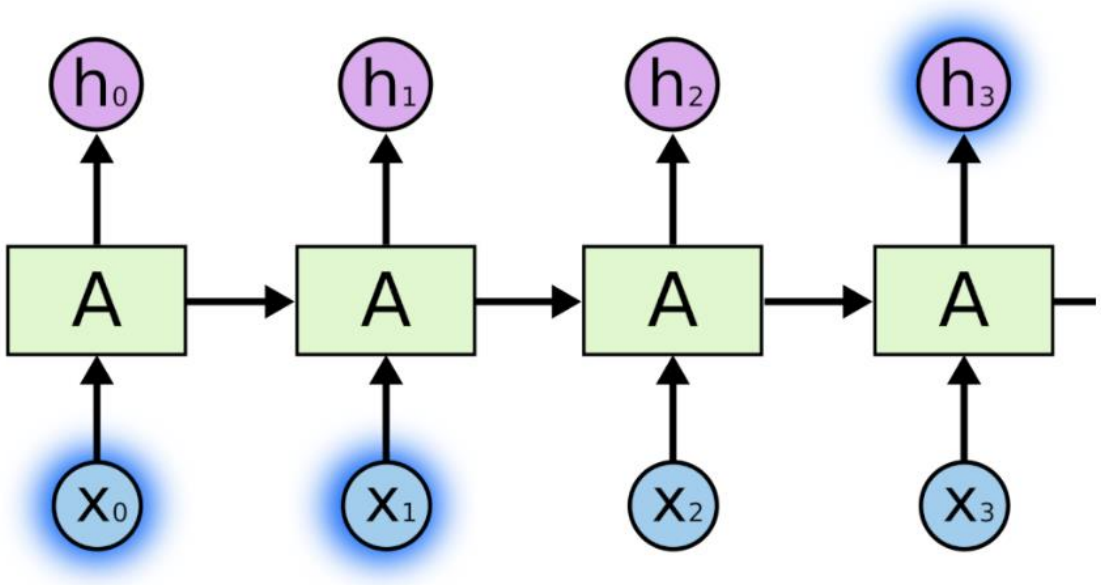


Рисунок 6 – Развертка цикла в рекуррентных нейронных сетях

Все рекуррентные нейронные сети имеют форму цепочки повторяющихся модулей нейронной сети. В стандартных РНС этот повторяющийся модуль имеет простую структуру, например, один слой **tanh**.

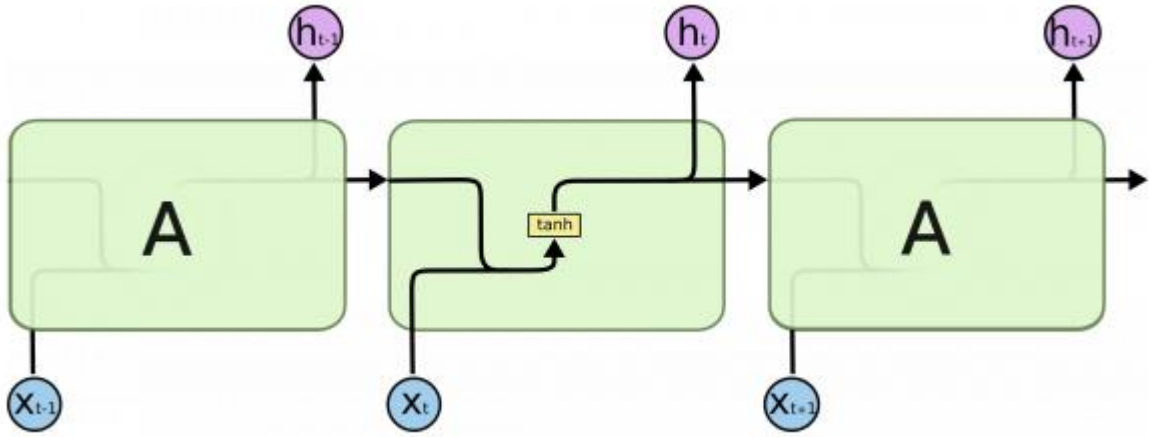


Рисунок 7 – Повторяющийся модуль стандартной РНС, состоящий из одного слоя

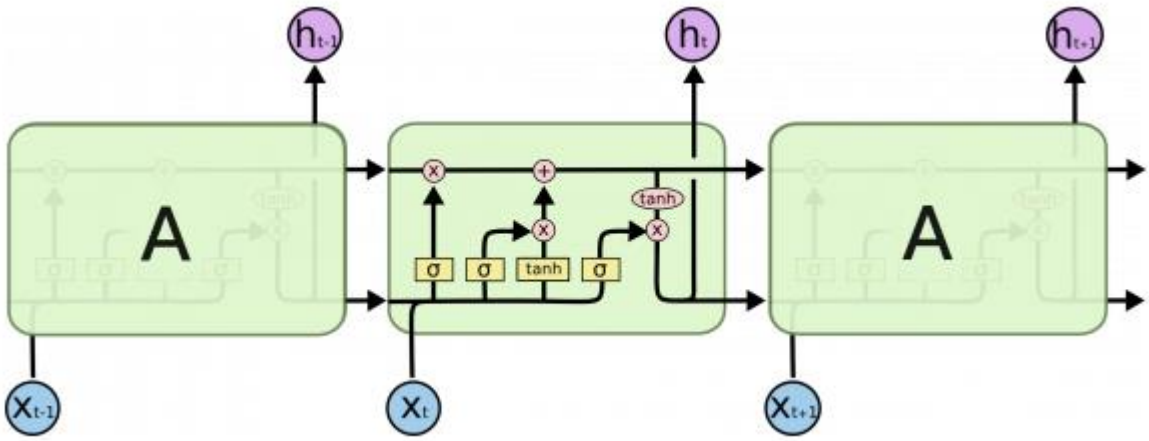


Рисунок 8 – Повторяющийся модуль LSTM, состоящий из четырех слоев

Основная идея LSTM заключается в использовании специальных блоков памяти, которые могут добавлять или удалять информацию в зависимости от ситуации. Каждый блок памяти состоит из трех компонентов: забывания, входа и выхода.

Алгоритм работы LSTM начинается с ввода входных данных в сеть. Далее данные поступают на входной уровень, где они проходят через ряд слоев, в каждом из которых происходит обработка. В случае LSTM, наиболее важным является блок памяти, который определяет, какую информацию следует сохранить и какую следует забыть.

Входной слой сети принимает информацию от предыдущего временного шага и текущего входа. Затем эта информация проходит через четыре уровня, каждый из которых выполняет определенные функции: забывание, добавление новой информации, обновление состояния памяти и вывод.

Забывание осуществляется за счет использования сигмоидальной функции, которая решает, какую информацию нужно забыть. Затем выполняется процесс добавления новой информации, который определяет, какую информацию нужно сохранить. Для этого используется гиперболический тангенс, который возвращает новую информацию, которую необходимо добавить в память.

Обновление состояния памяти осуществляется путем использования ранее полученных результатов и новых данных, которые были приняты на

входном уровне. Наконец, вывод позволяет выбрать, какую информацию нужно передать на следующий временной шаг.

Нейросеть Transformer – это одна из наиболее популярных архитектур глубокого обучения, используемых для обработки последовательностей данных, таких как тексты или звуковые сигналы. Общий алгоритм работы Transformer состоит из двух частей: энкодера и декодера.

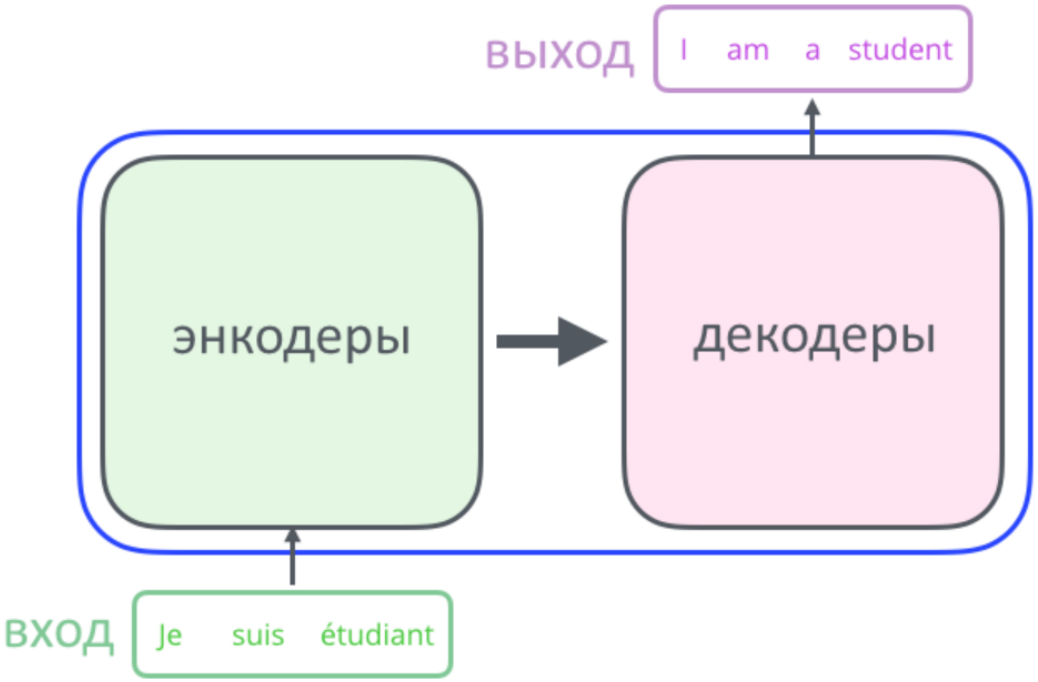


Рисунок 9 – Общий алгоритм работы трансформера

Энкодер в нейросети Transformer работает следующим образом. На вход энкодеру подается последовательность данных, которую необходимо обработать. Сначала каждый элемент последовательности преобразуется в вектор фиксированной размерности, называемый эмбеддингом. Затем эти эмбеддинги проходят через Positional Encoding – метод, используемый для добавления информации о позиции каждого элемента входной последовательности. Далее данные проходят несколько слоев нейросети, которые последовательно вычисляют некоторые преобразования. На каждом слое используется механизм внимания, который позволяет энкодеру фокусироваться на наиболее важных элементах последовательности. На

выходе энкодера получается набор векторов, которые содержат информацию о каждом элементе последовательности.

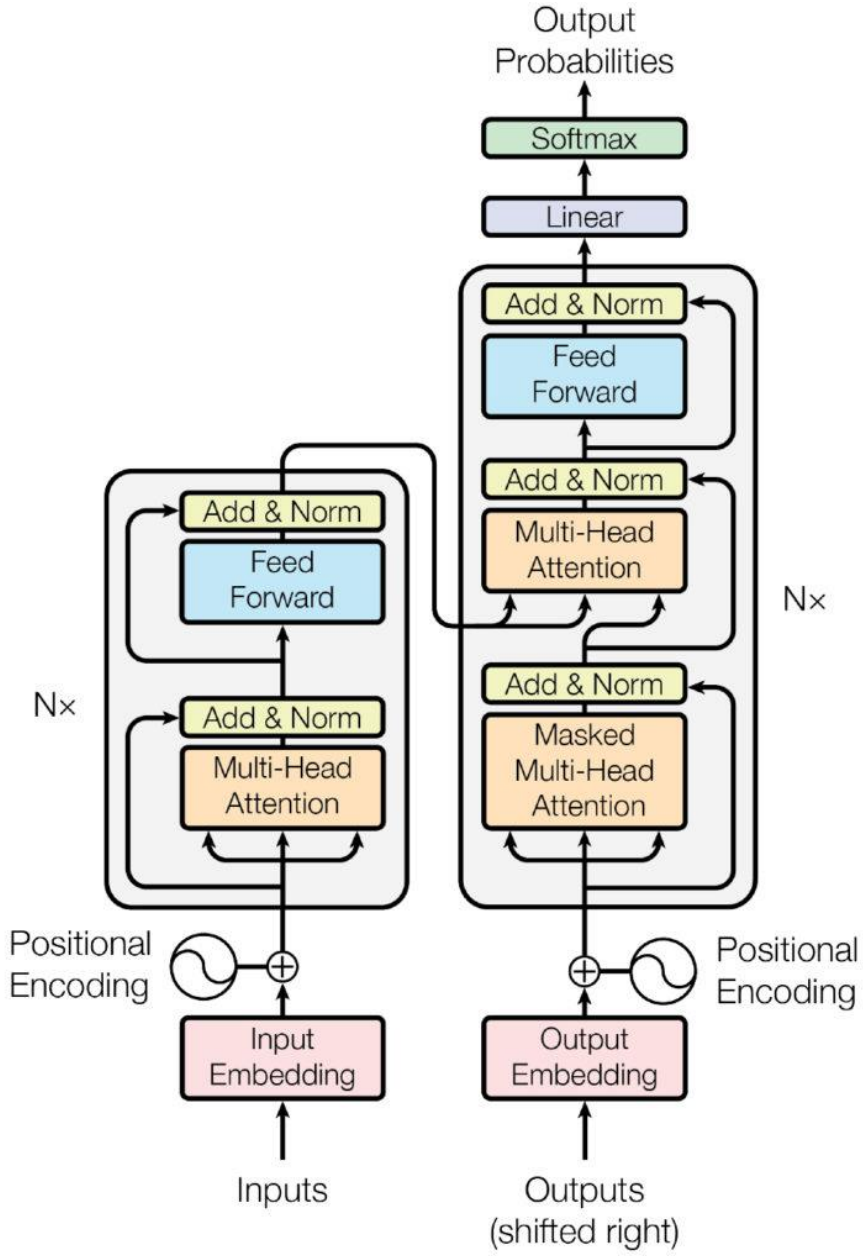


Рисунок 10 – Архитектура трансформера

Внимание в нейросети Transformer представляет собой механизм, который позволяет энкодеру или декодеру обращаться к определенным элементам последовательности, наиболее важным для решения задачи. В частности, внимание позволяет нейросети определять важность каждого элемента последовательности и вычислять взвешенные суммы этих элементов. Механизм внимания включает в себя несколько слоев, в каждом из

которых вычисляются веса, отображающие важность каждого элемента последовательности.

Декодер в нейросети Transformer работает следующим образом. На вход декодеру также подается последовательность данных, но в отличие от энкодера, декодер получает еще и выход энкодера, который содержит информацию о важности каждого элемента входной последовательности. Декодер постепенно генерирует выходную последовательность, элемент за элементом. На каждом шаге декодер использует механизм внимания, чтобы определить, на какие элементы входной последовательности следует сосредоточиться для генерации очередного элемента выходной последовательности. По мере генерации каждого элемента, декодер получает все больше информации об исходной последовательности и использует ее для генерации последующих элементов.

Основное преимущество Transformer заключается в его способности использовать параллельную обработку для ускорения вычислений, что делает его особенно полезным для больших наборов данных и высокопроизводительных приложений.

В целом, как LSTM, так и Transformer являются мощными инструментами для обработки последовательных данных. Они используются для решения широкого спектра задач в различных областях, включая естественный язык, компьютерное зрение и анализ данных.

1.6.8 Готовые решения для обработки и генерации естественного языка

Существует множество моделей для обработки и генерации естественного языка, вот некоторые из них:

BERT (Bidirectional Encoder Representations from Transformers) – это модель, основанная на трансформерах, которая используется для понимания естественного языка и выполнения различных задач NLP, таких как

классификация текста, ответы на вопросы и машинный перевод. Она работает путем "питания" текста в нейронную сеть в двух направлениях – вперед и назад – и обучения ее вычленять смысловую информацию из текста. Примером готового решения, основанного на BERT, является библиотека Hugging Face Transformers, которая позволяет использовать предобученные модели BERT для различных задач NLP.

GPT (Generative Pre-trained Transformer) – это модель, основанная на трансформерах, которая используется для генерации естественного языка. Она работает путем обучения на огромном количестве текстовых данных и позволяет генерировать новые тексты, имитирующие стиль и смысловую связь обучающих данных. Примером готового решения, основанного на GPT, является GPT-3 от OpenAI, доступный для использования через их API.

Seq2Seq (Sequence to Sequence) – это модель, используемая для машинного перевода, генерации ответов на вопросы и других задач генерации естественного языка. Она работает путем преобразования входного текста в вектор фиксированной длины с помощью энкодера, а затем декодирования этого вектора в выходной текст с помощью декодера. Примером готового решения, основанного на Seq2Seq, является Google Neural Machine Translation (GNMT), который используется для машинного перевода на множество языков.

2. Чат-боты в банковской сфере

2.1 Возможные варианты использования

В сегодняшней тенденции автоматизации банковский мир постепенно ориентируется на самообслуживание, чтобы удовлетворить потребности и требования клиентов, разбирающихся в цифровых технологиях. Таким образом, включение чат-ботов в финансовую отрасль является замечательным

явлением, которое в значительной степени снижает общую банковскую задачу. С помощью чат-бота клиенты банка могут без особых хлопот совершать любые финансовые операции с помощью текстового или голосового сообщения.

Какое отношение банк имеет к чат-ботам? Ответ довольно прост: для автоматизации сервисов. Как видите, сервисы в наши дни работают довольно медленно и иногда даже неприятно, поскольку люди относительно более склонны к непониманию и ошибкам, чем компьютерные программы.

Таким образом, диалоговый чат-бот может помочь вам обеспечить исключительное обслуживание клиентов, поскольку он доступен 24/7, никогда ничего не забывает, никогда не болеет и никогда не становится непродуктивным. Виртуальный помощник для банков может быть установлен для выполнения повседневных операций и повышения качества обслуживания клиентов в секторе цифрового банкинга [5].

На рисунке 5 приведены некоторые примеры использования чат-ботов в банковской сфере:

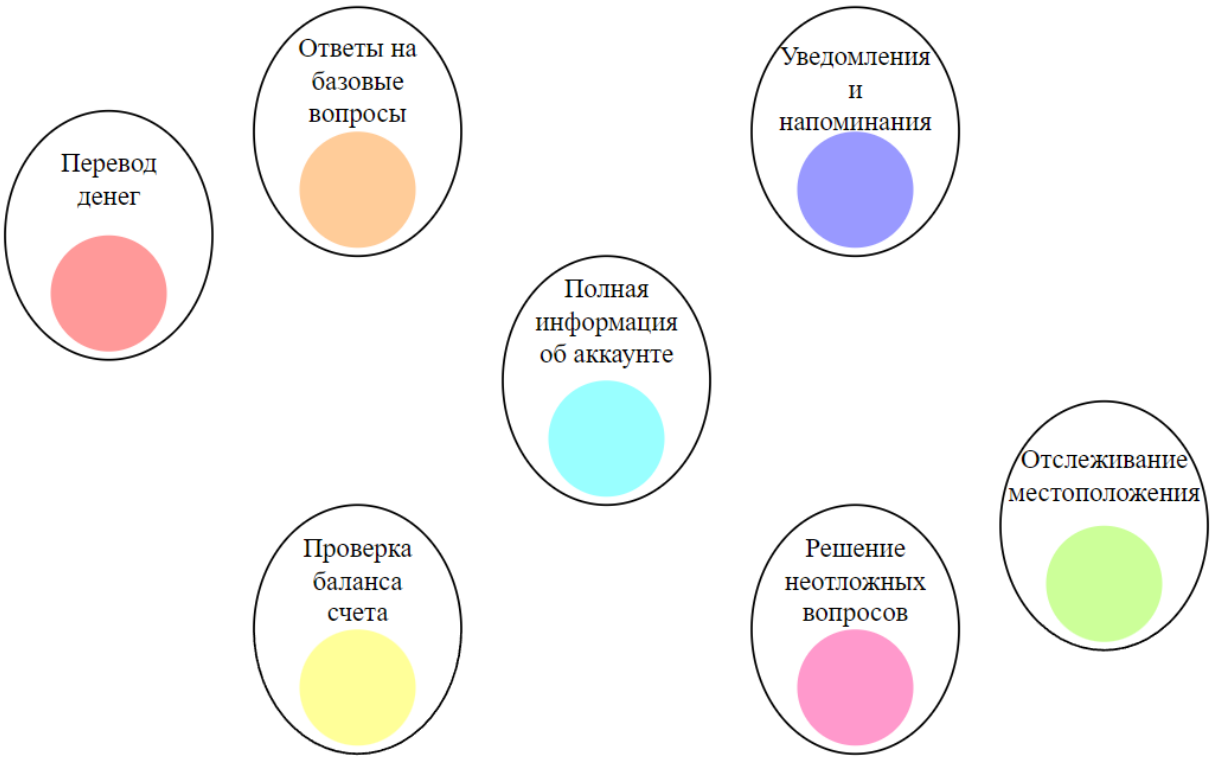


Рисунок 11 – Направления использования чат-бота в банкинге

Более подробно про каждое из направлений использования:

1) Перевод денег. Пользователи могут использовать чат-ботов для быстрого перевода денег, написав ему всего одну фразу, к примеру “переведи 2000 Петру Иванову”. Также можно просить компьютерного ассистента отменить какую-либо транзакцию и т.д.;

2) Ответы на базовые вопросы. Чат-боты могут отвечать на разные фундаментальные вопросов, касающихся счетов клиентов или банковских продуктов. Например, они могут отвечать на такие вопросы, как “Как я могу подать заявку на получение кредитной карты?”;

3) Уведомления и напоминания. Большинство банков используют чат-ботов, чтобы отправлять своим клиентам своевременные напоминания и регулярные уведомления об их банковских счетах. Некоторые из частых напоминаний, которые часто получают клиенты, касаются сроков оплаты счетов, предложения кредита в последний день и так далее. Все эти напоминания предназначены для того, чтобы информировать клиентов обо всех действиях, которые могут принести им пользу, и оставаться с ними;

4) Проверка баланса счета. Пользователи могут попросить чат-ботов предоставить им информацию о балансе счета под своим именем;

5) Предоставить полную информацию. Помимо остатка на счете, пользователи также могут запрашивать другие детали счетов, такие как регулярные платежи и расходы, бонусные баллы по карте и лимиты денежных переводов. Можно также восстановить данные своей учетной записи и внести изменения, такие как обновление текущего адреса или номера телефона;

6) Отслеживание местоположения в режиме реального времени. В зависимости от местоположения ответы на вопросы пользователей могут различаться. Например, если пользователь спросит: “Где ближайшее отделение банка?” В этом случае чат-бот будет отвечать в зависимости от местоположения пользователя. Кроме того, чат-боты могут отслеживать

местоположение с помощью мобильного GPS, тем самым каждый раз давая правильные ответы;

7) Решать неотложные вопросы в приоритете. Чат-боты в банковской сфере могут помочь клиентам с проблемами, которые могут быть несложными, но срочными. Эти проблемы включают разблокировку или блокировку карт, сброс, проверку банковских выписок и выполнение денежных переводов. Чат-бот с искусственным интеллектом позволяет клиентам завершить весь процесс, не дожидаясь ответа по телефону.

2.2 Будущее чат-ботов в банковской сфере

Доля банков, использующих ИИ-решения и, в частности, чат-ботов, постоянно растет. В качестве еще одного фактора, использование смартфонов и других интеллектуальных устройств также является быстро растущей тенденцией. Эти две движущие силы определяют ближайшее будущее помощников искусственного интеллекта в банковской сфере.

Качество чат-ботов определенно улучшится в ближайшие несколько лет. Они станут более «человечными» и научатся гораздо лучше интерпретировать просьбы. В качестве дальнейшего развития чат-боты будут более точно предсказывать поведение человека и использовать эту информацию для самообучения.

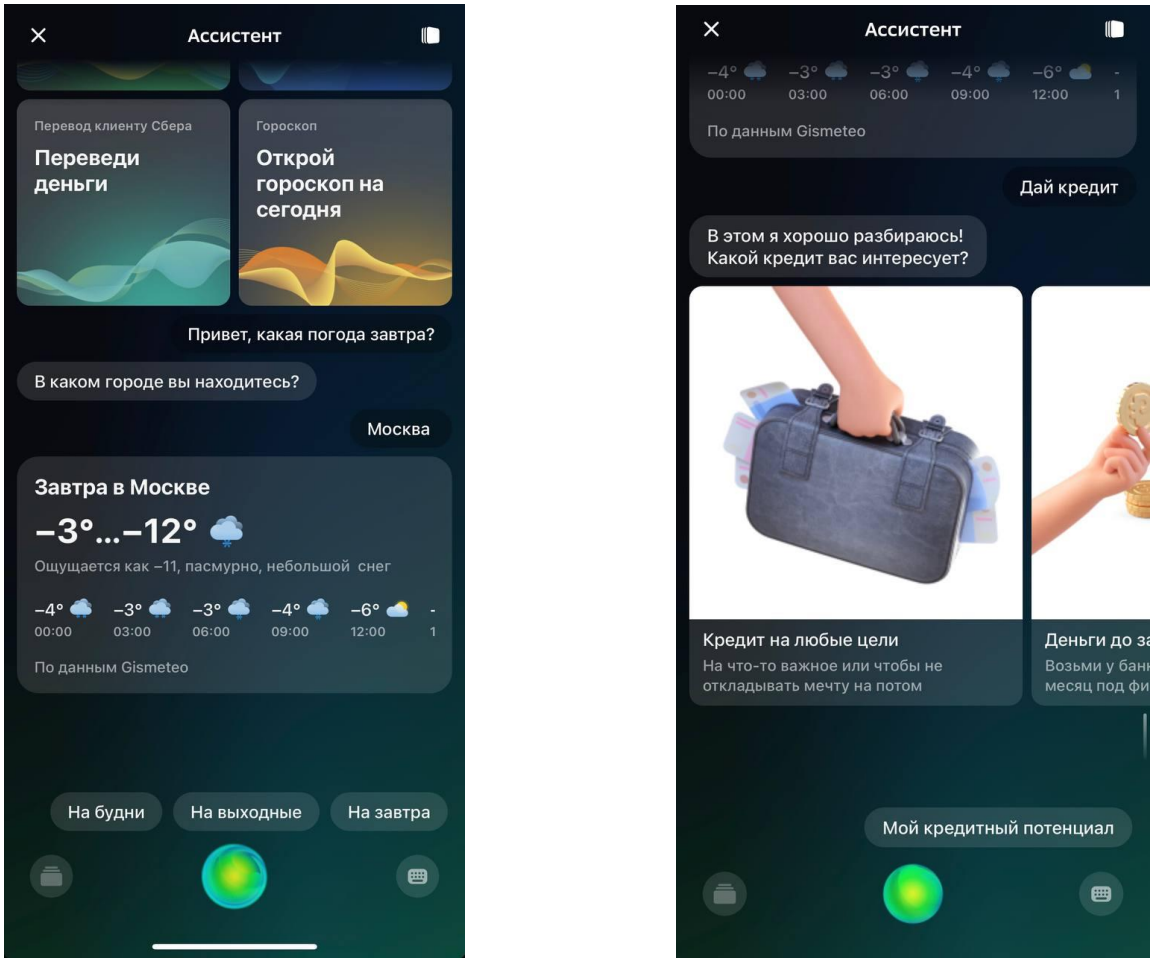
3. Обзор существующих решений

В банковской индустрии уже есть чат-бот решения. Мы рассмотрим виртуальных ассистентов, которые существуют в мобильных приложениях ведущих банков России. Для оценки чат-ботов будет руководствоваться следующими критериями:

- возможность общаться в свободной форме;
- умение переключаться между тематиками без потери контекста;
- возможность перейти на оператора по запросу или после ошибки чат-бота;
- наличие кнопок-подсказок в чате, ускоряющих консультацию.

3.1 Сбер Банк

Сбер Банк – крупнейший универсальный банк России и Восточной Европы. У него есть мобильное приложение, в котором можно найти умного чат-бота. Проанализировав компьютерного помощника, было определено, что это не узкоспециализированный банковский ассистент, который сможет помочь только с банковскими вопросами, а полноценный помощник, способный подобрать билеты в кино, поставить будильник на утро, подобрать выгодный вклад и посоветовать хороший кредит.



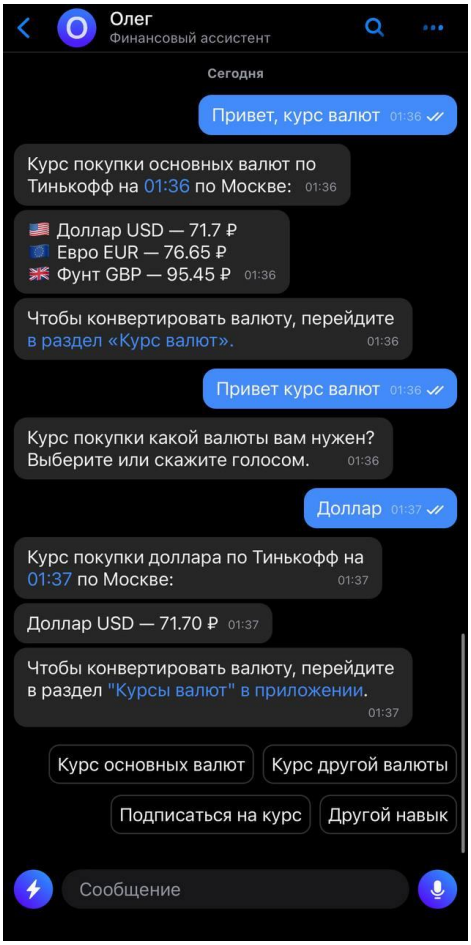
Рисунки 12-13 – Взаимодействие с виртуальным ассистентом Сбер Банка

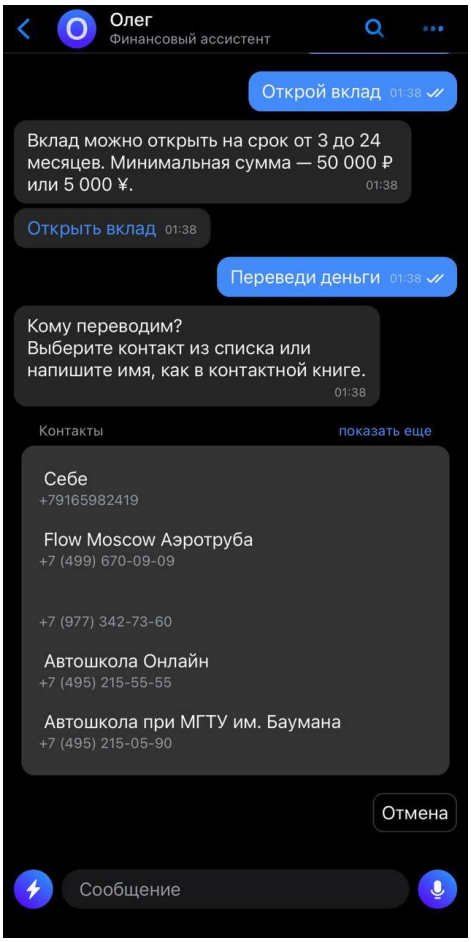
Ему можно отправить как текстовое сообщение, так и произнести команду голосом. Написав несколько команд, можно сделать вывод, что он очень хорошо определяет контекст запроса и предлагает нужные решения. Если попытаться его запутать или написать с ошибкой – он все равно поймет,

что вы имели ввиду. При попытке, к примеру, оформить вклад по рекомендации чат-бота, он открывает действие открытия вклада, как если бы вы нашли кнопку для этого действия в приложении без помощи ассистента. Это сделано для того, чтобы человек осознанно и самостоятельно производил операцию.

3.2 Тинькофф Банк

Тинькофф Банк – российский коммерческий банк, сфокусированный полностью на дистанционном обслуживании, не имеющий розничных отделений. Он считается крупнейшим в мире онлайн-банком по количеству клиентов. В его банковском приложении во вкладке “Чат” можно найти виртуального помощника Олега. Если обратиться к нему, вы получите ответы только на вопросы, связанные с банковской тематикой.





Рисунки 14-15 – Общение с чат-ботом Тинькофф Банка

Этот ассистент сможет рассказать вам о кредитах, вкладах или сообщит курс доллара к рублю. Если сравнивать его возможности с предыдущим конкурентом – ботом в Сбер Банке, то они не такие широкие. По оценочным критериям, приведенным выше, можно сделать вывод, что чат-бот Олег не очень способен общаться в свободной форме, однако имеет различные кнопки-подсказки, которые ускоряют консультацию.

3.3 Почта Банк

Почта Банк — универсальный розничный банк, созданный в 2016 году группой ВТБ и Почтой России. Ключевая цель Почта Банка — повышение доступности финансовых услуг для жителей России. Чат-бота этого банка можно найти в мобильном приложении или на сайте. Во время общения с виртуальным помощником Дмитрием, пользователь должен выбирать фразы из перечня предложенных ботом (рис. 9-10).

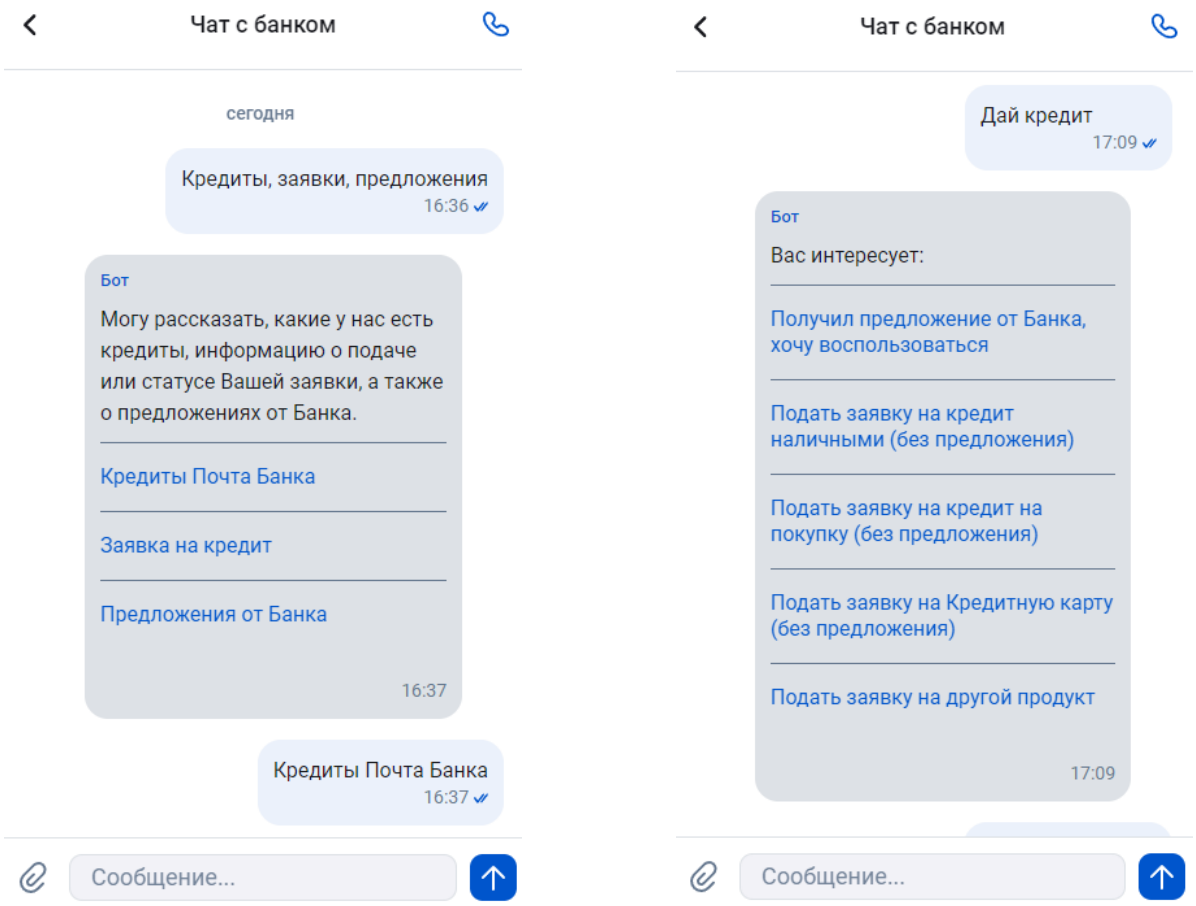


Рисунок 16-17 – Общение с виртуальным ассистентом Почта Банка

Чат-бот этого банка хорошо распознает намерение пользователя и предлагает наиболее полезные возможности. Он понимает свободную форму общения, не только при помощи кнопок. Помощник Дмитрий переключает вас к оператору, если он не понял сообщения от пользователя или клиента не устроило ничего из предложенных вариантов.

ЗАКЛЮЧЕНИЕ

В результате изучения различных источников, мы пришли к выводу, что чат-бот – программа, которая имитирует реальный разговор с пользователем. Мы узнали, что они могут быть разговорными, представлять из себя полноценного ассистента или являться ботом, который отвечает на вопросы. В данной научной работе была исследована архитектура и алгоритмы работы чат-ботов, были определены основные этапы разработки чат-ботов. Была выявлена роль машинного обучения в разработке умных виртуальных ассистентов. Основными направлениями для использования Machine Learning в чат-ботах являются:

- обработка и генерация естественного языка;
- задача классификации намерения пользователей.

В зависимости от идеи чат-бота, он может использовать и другие задачи машинного обучения.

В ходе проведения исследования получилось выявить возможные варианты использования чат-ботов и выполнить обзор существующих решений в банковской сфере.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Антонов С.* Что такое чат-боты и зачем они нужны? // Inform Бюро [Электронный ресурс]. Режим доступа: https://informburo.kz/cards/chto-takoe-chat-boty-i-zachem-oni-nuzhny.html (дата обращения 15.11.2022);
2. *Mirant Hingrajia.* How do Chatbots work? A Guide to Chatbot Architecture // maruti techlabs [Электронный ресурс]. Режим доступа: https://marutitech.com/chatbots-work-guide-chatbot-architecture/ (дата обращения 20.11.2022);
3. *Jenna Alburger*. Rule-Based Chatbots vs. AI Chatbots: Key Differences // hubtype [Электронный ресурс]. Режим доступа: https://www.hubtype.com/blog/rule-based-chatbots-vs-ai-chatbots (дата обращения 20.11.2022);
4. Types of chatbots // freshworks [Электронный ресурс]. Режим доступа: https://www.freshworks.com/live-chat-software/chatbots/three-types-of-chatbots/ (дата обращения 20.11.2022);
5. *Shambhavi Sinha.* Chatbot for Banking: Everything you Need to Know // AMEYO [Электронный ресурс]. Режим доступа: https://www.ameyo.com/blog/chatbot-for-banking-everything-you-need-to-know/#:~:text=Chatbots%20in%20banking%20industries%20can,without%20waiting%20on%20the%20phone (дата обращения 22.11.2022);
6. *Анна Юченко*. How do chatbots work? Often with a little help from AI // TechArt [Электронный ресурс]. Режим доступа: https://www.itechart.com/blog/how-do-chatbots-really-work/ (дата обращения 22.11.2022);
7. Что такое машинное обучение? // Azure [Электронный ресурс]. Режим доступа: https://azure.microsoft.com/ru-ru/resources/cloud-computing-dictionary/what-is-machine-learning-platform/#:~:text=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5%20%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%B

D%D0%B8%D0%B5%20(ML)%20%E2%80%94%20%D1%8D%D1%82%D0%BE,%D0%B8%D0%B7%20%D1%84%D0%BE%D1%80%D0%BC%20%D0%B8%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE%20%D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82%D0%B0%20(%D0%98%D0%98) (дата обращения 05.12.2022);

8. Какие задачи позволяет решать машинное обучение? // ЦИСМ [Электронный ресурс]. Режим доступа: https://www.cism-ms.ru/poleznye-materialy/kakie-zadachi-pozvolyaet-reshat-mashinnoe-obuchenie-/ (дата обращения 05.12.2022);

9. Чат-боты – кто они и что умеют? // EFSOL [Электронный ресурс]. Режим доступа: https://efsol.ru/articles/messendzhery-i-chat-boty-dlya-biznesadostavki.html (Дата обращения: 06.12.2022);

10. *Ася Зуйкова*. Что такое машинное обучение и как оно работает // РБК Тренды [Электронный ресурс]. Режим доступа: https://trends.rbc.ru/trends/industry/60c85c599a7947f5776ad409 (дата обращения 24.12.2022).