

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
АМУРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
(ФГБОУ ВО «АмГУ»)

Факультет математики и информатики
Кафедра информационных и управляющих систем
Направление подготовки 09.03.02 – Информационные системы и технологии
Направленность (профиль) образовательной программы: Безопасность
информационных систем

ДОПУСТИТЬ К ЗАЩИТЕ
Зав. кафедрой
_____ А.В. Бушманов
« _____ » _____ 2021 г.

БАКАЛАВРСКАЯ РАБОТА

на тему: Разработка информационной системы для бинарной классификации сообщений к теме ЧС методами машинного обучения

Исполнитель
студент группы 755-об

(подпись, дата)

Г.П. Плахонин

Руководитель
доцент, к.ф-м.н

(подпись, дата)

В.В. Еремина

Консультант по части
безопасности и
экологичности, доцент,
к.т.н

(подпись, дата)

А.Б. Булгаков

Нормоконтроль
доцент, к.т.н

(подпись, дата)

О.В. Жилиндина

Благовещенск 2021

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
АМУРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
(ФГБОУ ВО «АмГУ»)

Факультет математики и информатики
Кафедра информационных и управляющих систем

УТВЕРЖДАЮ

Зав. кафедрой

_____ А.В. Бушманов
« _____ » _____ 2021 г.

З А Д А Н И Е

К выпускной квалификационной работе студента Плахонина Геннадия Павловича

1. Тема дипломной работы: Разработка информационной системы для бинарной классификации сообщений к теме ЧС методами машинного обучения.

(утверждена приказом от 23.04.2021 №812-уч)

2. Срок сдачи студентом законченной работы: 24.06.2021 г.

3. Исходные данные к выпускной квалификационной работе: отчет о прохождении преддипломной практики, нормативная документация, специальная литература.

4. Содержание выпускной квалификационной работы (перечень подлежащих разработке вопросов): анализ предметной области и организации, обоснование необходимости разработки и определение требований, проектирование программного продукта, оценка надежности и качества функционирования объекта проектирования, описание способов защиты информации для программы, обоснование безопасности и экологичности продукта.

5. Консультанты по выпускной квалификационной работе:
по безопасности и экологичности – Булгаков А.Б., доцент, кандидат технических наук.

6. Дата выдачи задания: 20.02.2021 г.

Руководитель выпускной квалификационной работы: Еремина В.В. доцент, кандидат физико-математических наук.

Задание принял к исполнению:

РЕФЕРАТ

Выпускная квалификационная работа содержит 109 с., 46 рисунков, 9 таблиц, 5 приложений, 22 источника.

РАЗРАБОТКА, БАЗА ДАННЫХ, АНАЛИЗ УЧРЕЖДЕНИЯ, ПРОЕКТИРОВАНИЕ, МАШИННОЕ ОБУЧЕНИЕ, ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА, ИНФОРМАЦИОННАЯ СИСТЕМА.

В работе реализовано проектирование и разработка информационной системы для бинарной классификации сообщений к теме ЧС методами машинного обучения.

Целью работы является создание ИС позволяющей повысить скорость реагирования на ЧС, обеспечить своевременное оповещение и информирование персонала учреждения и людей, оказавшихся в зоне действия ЧС, распределение нагрузки между несколькими операторами ИС, снижение временных затрат на мониторинг и классификацию сообщений из разных источников информации организации.

В ходе работы было необходимо:

- Проанализировать предметную область и деятельность изучаемого учреждения;
- Определить цели и функции приложения, выбрать модель жизненного цикла ПО, спроектировать структуру приложения;
- Выбрать средства разработки, произвести обработку и анализ данных для обучения классификатора, разработать ИС;
- Рассмотреть угрозы ИБ приложения и предложить решения по их устранению.

СОДЕРЖАНИЕ

Введение	7
1 Анализ предметной области	9
1.1 Автоматизированная система раннего обнаружения чрезвычайных ситуаций и оповещения	9
1.2 Машинное обучение	10
1.2.1 Основные понятие и определения машинного обучения	10
1.2.2 Обучение с учителем и задача классификации	12
1.3 Natural Language Processing (NLP)	13
1.3.1 Определение и основные понятия NLP	14
1.3.2 NLP-конвейер	15
2 Анализ деятельности учреждения	19
2.1 Цели и задачи учреждения	19
2.2 Организационная структура учреждения	19
2.3 Анализ документооборота	23
2.3.1 Внешний документооборот	24
2.3.2 Внутренний документооборот	24
2.4 Анализ использования программно-технических средств	24
2.5 Анализ существующей ИС	25
3 Проектирование информационной системы	26
3.1 Цель и назначение создания системы	26
3.2 Разработка проекта автоматизации	27
3.2.1 Выбор модели жизненного цикла ПО	27
3.2.2 Ожидаемые риски на этапах ЖЦ и их описание	31
3.3 Структура ИС	32
3.4 Характеристика функциональных подсистем программы	33
3.4.1 Описание модуля авторизации	35
3.5 Проектирование БД	38
3.5.1 Инфологическое проектирование	38
3.5.2 Логическое проектирование	41
3.5.3 Физическое проектирование	42

3.6	Задачи автоматизации, решаемые разрабатываемой ИС	43
4	Разработка ИС	45
4.1	Описание разрабатываемой ИС	45
4.2	Выбор средств разработки	45
4.2.1	Выбор СУБД	45
4.2.2	Выбор языка программирования	46
4.3	Работа с данными и обучение классификатора	49
4.3.1	Токенизация и лемматизация	52
4.3.2	Очистка данных	52
4.3.3	Векторизация текста	57
4.3.4	Обучение классификатора	57
4.4	Разработка экранных форм приложения	61
5	Угрозы информационной безопасности приложения	68
5.1	Предложения по обеспечению ИБ системы	68
5.2	Требования к надежности	68
5.2.1	Состав показателей надежности системы в целом.	68
5.2.2	Перечень аварийных ситуаций, по которым регламентируются требования к надежности.	69
5.2.3	Требования к надежности технических средств и программного обеспечения	69
5.3	Требования по сохранности информации при авариях	70
6	Безопасность и экологичность	71
6.1.	Безопасность	72
6.1.1.	Определение безопасных и вредных факторов на рабочем месте пользователя ПЭВМ.	72
6.1.2.	Освещение	74
6.1.3.	Механические колебания	75
6.1.4.	Электромагнитное и ионизирующее излучение	76
6.1.5.	Микроклимат	76
6.1.6.	Анализ помещения с установленными ПЭВМ	78
6.2.	Экологичность	79
6.3.	Чрезвычайные ситуации	80

6.3.1 Меры пожарной безопасности на рабочих местах	81
6.3.2 Роль разрабатываемой системы в предупреждении ЧС	83
Заключение	84
Библиографический список	85
Приложение А	87
Приложение Б	88
Приложение В	89
Приложение Г	91
Приложение Д	92

ВВЕДЕНИЕ

Обработка естественного языка сейчас используется во многих видах деятельности. Поисковые системы для нахождения релевантных и похожих результатов, машинный перевод, спам-фильтры, автокоррекция, распознавание речи и многое другое. Но язык – это сложная совокупность различных уровней, таких как синтаксис, морфология, семантика, дискурс. Для каждого уровня выделяются свои специфические задачи. Но на практике задействуются несколько или все уровни языка. Например, задачи лемматизации, то есть приведение слова к начальной форме или токенизации – деление текста на его составляющие. Есть и задачи синтаксического анализа текста или парсинга. А машинный перевод задействует все уровни языка. И то, что человек может легко понять из контекста, у машины вызывает определенные трудности. Но автоматизация анализа и обработки текста и речи является актуальной темой в наше время, поскольку это позволяет автоматизировать выполнение широкого спектра задач, будь то распознавание речи, машинный перевод, выделение основного смысла текста, и сократить время реагирования на возникающие инциденты в различных сферах.

Многие сферы бизнеса заинтересованы в решении этих задач. В частности, в обработке естественного языка заинтересованы контактные центры, которым требуется оперировать большим потоком входящих запросов. Так же, обработка естественного языка может использоваться для уведомления соответствующих служб о чрезвычайных ситуациях, ведь в данном случае важна скорость реагирования на возникающую опасность, которой можно добиться с помощью методов машинного обучения.

Машинное обучение — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. При этом, машина

намного быстрее анализирует и обрабатывает поступающие данные, что является очень важным качеством в современном мире.

В свою очередь, социальные сети глубоко укоренились в нашем обществе, и мы пользуемся ими постоянно. Например, Twitter насчитывает около 330 миллионов активных пользователей в месяц (по состоянию на четвертый квартал 2020 года)¹ и около 500 миллионов твитов² (сообщений) отправляются каждый день. Согласно официальному документу Telegram Open Network ICO Whitepaper³, число активных российских пользователей составляет 9.8 млн. человек, а прирост новых пользователей со всего мира составляет по меньшей мере 500 тысяч человек ежедневно. По состоянию на начало 2021 года Telegram входит в тройку самых популярных мессенджеров в России, где его опережают только Whatsapp и Viber и количество сообщений, отправляемых пользователями ежедневно, достигает 70 млрд. Это огромное количество информации и с помощью алгоритмов машинного обучения для анализа и обработки текста и речи можно извлечь много полезных данных.

Целью работы является создание автоматизированной системы раннего обнаружения чрезвычайных ситуаций и оповещения, которая позволит повысить скорость реагирования на ЧС, обеспечить своевременное оповещение и информирование персонала учреждения и людей, оказавшихся в зоне ЧС, распределить нагрузку между несколькими операторами ИС, снизить временные затраты на мониторинг и классификацию сообщений из разных источников информации организации.

¹ Источник https://s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Selected-Company-Metrics-and-Financials.pdf

² Источник <https://www.internetlivestats.com/twitter-statistics/>

³ Telegram Open Network ICO Whitepaper, 5-6 стр., URL: <https://relayto.com/relayto/telegram-open-network-ton-ico-whitepaper-6kf4rycn/pdf>

1 АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

1.1 Автоматизированная система раннего обнаружения чрезвычайных ситуаций и оповещения

Автоматизированная система раннего обнаружения чрезвычайных ситуаций и оповещения предназначена для автоматического непрерывного наблюдения и предварительной обработки информации относительно текущего состояния потенциально опасных объектов и зданий, инженерных сооружений и сетей, расположенных на территориях с риском проявления опасных природных явлений и процессов, оперативного предоставления пользователям фактической и прогнозируемой информации, а также оповещения, в случае необходимости, населения, работников и руководителей потенциально опасных объектов, ответственных за состояние техногенной безопасности, должностных лиц органов исполнительной власти и местного самоуправления при непосредственном участии человека-оператора.

Существуют уровни таких систем: федеральный, межрегиональный, региональный, муниципальный и объектовый. Для оповещения населения используют региональный, муниципальный и объектовый уровни. Системой оповещения населения в России ведает ГОЧС, в том числе Общероссийской комплексной системой информирования и оповещения населения в местах массового пребывания людей («ОКСИОН»), а также интегрированную в неё региональной автоматизированной системой централизованного оповещения гражданской обороны (РАСЦО ГО) и Системой защиты от чрезвычайных ситуаций природного и техногенного характера, информирования и оповещения населения на транспорте («СЗИОНТ»)

Сам процесс автоматизированной системы раннего обнаружения чрезвычайных ситуаций и оповещения похож во многих системах. Сначала осуществляется сбор данных с помощью аппаратных и/или программных средств, далее эти данные анализируются посредством заложенных в систему правил или методами машинного обучения и в конце, данные передаются

через средства связи пользователям или населению. Многие решения предлагают именно аппаратную реализацию сбора информации, например, пожарная сигнализация, которая необходима для обнаружения загорания, сообщения о месте его возникновения и обработки сигнала, также процесс получения, обработки, передачи и представления информации о пожаре в заданном виде потребителям с помощью данных технических средств.

Разрабатываемое решение работает именно с программной реализацией, где средством сбора информации является API сервиса, а обработкой и анализом данных выступают методы машинного обучения.

Система способна анализировать сообщения, написанные на естественном (человеческом) языке и разделять их по ЧС и не ЧС. Система обучена различать разные виды ЧС, будь то: природные (землетрясение, бури), техногенные (пожары, взрывы), биологические (эпидемии), социальные (терроризм, насилие), антропогенные (ошибочные действия людей). Это возможно благодаря тому, что система обучалась на большом объеме данных и может отличать такие сообщения от тех, которые не относятся к чрезвычайным ситуациям.

Если при анализе сообщения чрезвычайная ситуация подтверждена, то об этом уведомляется оператор ИС, который даёт окончательный ответ. Оператору ИС необходимо предпринять все необходимые меры для предотвращения ЧС или полной минимизации ущерба.

1.2 Машинное обучение

1.2.1 Основные понятие и определения машинного обучения

Машины так же могут обучаться и развиваться в ходе анализа и обработки данных. Машинное обучение реализует класс методов искусственного интеллекта, когда накапливаются данные, модель обучается на этих данных и решается не конкретная задача, а множества сходных задач. Это позволяет не менять конкретный алгоритм для каждой задачи и абстрагироваться от внутренней реализации. Тогда данные подаются на вход

алгоритму и выводится некоторый результат, зависящий от специфики задачи. Это работает по принципу «черного» ящика, где важна не внутренняя реализация алгоритма, а только данные, которые можно подать на вход алгоритму. В свою очередь, для построения методов машинного обучения используются средства математической статистики, численных методов, математического анализа, методов оптимизации, теории вероятностей, теории графов, различные техники работы с данными в цифровой форме.

Выделяют 2 типа машинного обучения:

- Индуктивное или по прецедентам, которое основано на нахождении зависимости во входных (экспериментальных) данных и позволяет получить приблизительный результат, который в типичных ситуациях будет близок к точному.

- Дедуктивное, где знания эксперта формализуются и переводятся в цифровую форму, где хранятся в базе знаний.

Термин “машинное обучение” употребляют в контексте индуктивного типа, так как дедуктивный тип принято относить к области экспертных систем. Прецедентами или обучающей выборкой выступают наборы входных объектов и соответствующие им результаты. Но стоит учесть, что нет какой-то четкой формулы, которая смогла бы полностью описать зависимость между результатами и входными данными. И на результат может влиять множество факторов, которые зависят от конкретной области исследования. При этом, необходимо построить алгоритм, который будет выдавать достаточно точный результат для любых возможных входных данных. Так же, система должны быть способна к адекватному отклику на данные, выходящие за пределы имеющейся обучающей выборки. На практике, входные данные могут быть неполными, неточными или разнородными, из-за чего с ними сложно работать без предварительной их обработки. Поэтому и существует множество методов машинного обучения. Можно сказать, что с помощью методов машинного обучения, строятся предположения на основе подобных случаев

(прецедентов). Чем больше таких прецедентов, тем качественнее обучается модель.

А почему именно машинное обучение в настоящее время постоянно развивается и используется всё чаще в различных сферах? Человечество смогло накопить большой объём данных, на которых модель можно обучить и эффективно решать сходные задачи. Разработка и обучение модели занимают также много времени и поэтому машинное обучение нужно не всегда. Если имеется, условно, 1 тысяча записей и их нужно рассортировать по категориям, при чём, это единичный случай, то построение модели машинного обучения не будет эффективно. Значит, реализация данной технологии должна проверяться на адекватность в каждом конкретном случае. В общем случае, если имеется огромный массив данных, исчисляемый сотнями тысяч или миллионами записей, или некоторый объём данных постоянно пополняется и их нужно обрабатывать, то построение модели машинного обучения более обосновано. В данной работе исследуется набор данных на более 11 тысяч записей, где 8 тысяч записей необходимы для обучения модели, так называемая обучающая выборка и 3 тысячи записи используются для анализа эффективности построенной модели. Проверка модели на тестовых данных также является важной задачей, так как модель, которая будет построена, должна быть не привязана или подогнана только к тем данным, на которых она обучается, то есть не должна возникать проблемы переобучение модели или утечки данных. Построение модели машинного обучения в данной работе является более эффективным решением проблемы, чем обработка информации самостоятельно, поскольку данные собираются из мессенджера Telegram и база данных может постоянно пополняться новыми данными быстрее, чем человек будет самостоятельно их анализировать.

1.2.2 Обучение с учителем и задача классификации

Обучение с учителем — наиболее распространённый случай. Имеется множество объектов (ситуаций) и множество ответов. Находится зависимость

между объектами и ответами и на основе этой зависимости строится алгоритм, принимающий на входе описание объекта и выдающий на выходе ответ. Чтобы проверить качество построенной модели, обычно определяется средняя ошибка ответов, выданных алгоритмом.

Задача классификации множество допустимых ответов конечно. Их называют метками классов. Класс — это множество всех объектов с данным значением метки. Данная задача относится к виду обучения с учителем.

В рамках данной работы будет использоваться тип по прецедентам. Данные представлены в формате “объект-ответ”, поэтому, это обучение с учителем. На тестовой выборке будет происходить проверка качества построенной модели, то есть её эффективности. В данном случае, решается задача бинарной классификации, то есть ИС предсказывает, к какой из двух групп (реальное ЧС или нет) относится данное сообщение. Бинарной классификации достаточно, так как целью данной работы является обнаружение самого факта ЧС. Для анализа сообщений, написанных на естественном языке, необходимо использовать методы машинного обучения для обработки естественного языка – NLP.

1.3 Natural Language Processing (NLP)

Обработка естественного языка – сложная и нетривиальная задача. Естественный язык постоянно меняется, появляются новые слова, а старые обретают новый смысл. Многие слова и фразы имеют как прямое, так и переносное значение и неоднозначно идентифицировать сказанное не всегда получается даже у человека.

Основная проблема при анализе естественных языков является их сложность и многозначность, что отличает их от искусственных языков, со строгой структурой и относительной простотой. Сложность строения естественных языков связано с их эволюционной природой. Лексика и грамматика естественного языка формируется на протяжении длительного времени, под влиянием разных эпох, культур и народов. Поэтому, при

построении модели и проверки её на реальных данных, добиться 100% результата практически невозможно, так как одни и те же слова могут сильно менять свой основной смысл от текущего контекста.

1.3.1 Определение и основные понятия NLP

Natural Language Processing (далее – NLP) – обработка естественного языка – подраздел информатики и AI (Artificial Intelligence), посвященный тому, как компьютеры анализируют естественные (человеческие) языки. NLP позволяет применять алгоритмы машинного обучения для текста и речи. Оно изучает проблемы компьютерного анализа и синтеза текстов на естественных языках. Применительно к искусственному интеллекту анализ означает понимание языка, а синтез — генерацию грамотного текста. Решение этих проблем будет означать создание более удобной формы взаимодействия компьютера и человека.

NLP решает большой набор задач, который можно разбить по уровням (в скобках). Среди этих задач, можно выделить следующие:

- Распознавание текста, речи, синтез речи (сигнал);
- Морфологический анализ, канонизация (слово);
- POS-тэгирование, распознавание именованных сущностей, выделение слов (словосочетание);
- Синтаксический разбор, токенизация предложений (предложение);
- Извлечение отношений, определение языка, анализ эмоциональной окраски (абзац);
- Аннотация документа, перевод, анализ тематики (документ);
- Дедубликация, информационный поиск (корпус).

Анализ текстов реализуется в трёх основных форматах: классификации, отражении содержания и анализе тональности.

Все задачи по классификации текстов (text classification) можно разделить на два типа:

- бинарная классификация позволяет отнести документ к одному из двух определенных классов;

- мультиклассовая классификация позволяет определить тематику документа и отнести его к одному из сотни тематических классов.

Отражение содержания текста (text summarization) работает по следующему принципу: на вход NLP-система принимает текст большого размера, а на выходе отдаёт текст меньшего размера, отражающий содержание большого.

Анализ тональности текста (sentiment analysis) позволяет находить в тексте мнения и выявлять их свойства. В зависимости от поставленной задачи, будут исследоваться разные свойства. Например, сам автор может стать целью анализа, то есть анализ тональности определяет типичный для него стиль, эмоциональную окраску текста и т.д.

1.3.2 NLP-конвейер

Реализация какой-либо сложной комплексной задачи в машинном обучении обычно означает построение пайплайна (с англ. трубопровод) или конвейера. Так мы можем разбить (декомпозировать) одну проблему на несколько маленьких частей и решить каждую часть отдельно. При этом, при решении одной маленькой задачи, результат её работы направляется на вход следующей и так продолжается до тех пор, пока не выполнится последняя подзадача. Это позволяет разделить обязанности внутри команды разработчиков, абстрагироваться от реализации каждого этапа и решать одну конкретную задачу в данный момент времени.

Этапы пайплайна:

- 1) Токенизация – необходим для предварительной обработки текста. На данном этапе длинные участки текста заменяются на более мелкие. Например, текст можно поделить на абзацы, абзацы на предложения, а предложения на отдельные слова.

2) Нормализация – процесс унификации текста. Этот этап необходим для качественной обработки текста, так как он позволяет применять методы обработки текста в унифицированном виде. Обычно, он включает в себя приведение слов к единому регистру, удаление знаков пунктуации, расшифрование сокращений, словесное написание чисел и т.д.

3) Стеммизация (Stemming) – лингвистическая основа слова. В старославянских языках близко к понятию «корень слова». Это операция приведения слова к его корню с помощью удаления приставки, суффиксов и окончания.

4) Лемматизация – приведение слова к его начальной форме. Например, приведение существительных и прилагательных к единственному числу и именительному падежу, приведение глаголов к инфинитиву и т.д.

5) Удаление стоп-слов – очистка текста от слов, которые не несут смысловой нагрузки (артикли, междометья, союзы, предлоги и т.д.).

По завершении всех этих операций текст становится пригодным для его перевода в числовую форму, чтобы дальше продолжить извлечение признаков.

6) Синтаксический анализ (парсинг) зависимостей – процесс сопоставления токенов языка с его формальной грамматикой, то есть установление связи между словами предложения. Исходный текст преобразуется в дерево, где каждый токен имеет единственного родителя и таким образом устанавливается связь между двумя словами.

7) Распознавание именованных сущностей (Named Entity Recognition, NER). Целью данного этапа является обнаружение существительных и связыванием их с реальными концепциями. Например, слово «Лондон» относится к географической сущности. NER-системы не просто просматривают словари. Они анализируют контекст токена в предложении и используют статистические модели, чтобы угадать какой

объект он представляет. Хорошие NER-системы способны отличить актрису Brooklyn Decker от города Brooklyn.

8) Разрешение кореференции. Кореференция – это попытка связать несколько разных отсылок в тексте к одному реальному объекту. Например, в английском языке очень много местоимений. Они являются сокращениями, которые мы заменяем на письме настоящими именами и названиями. Человек может проследить взаимосвязь этих слов от предложения к предложению, основываясь на контексте. Но NLP-модель не знает о том, что означают местоимения, ведь она рассматривает всего одно предложение за раз.

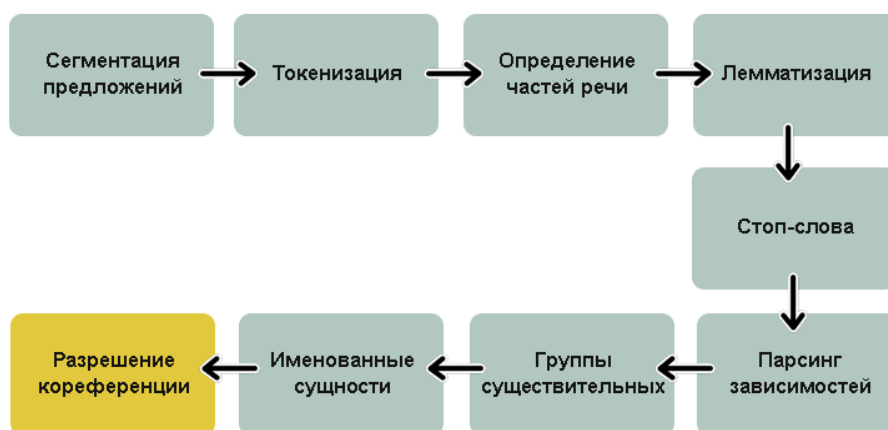


Рисунок 1 – NLP-конвейер

Это стандартные этапы обычного NLP-конвейера, но в зависимости от конечной цели проекта и особенностей реализации модели, некоторые из них можно пропускать или менять местами.

1.4 Сравнительный анализ существующих решений

Основной проблемой при анализе существующих решений является то, что разрабатываемая ИС объединяет систему автоматизированного раннего обнаружения чрезвычайных ситуаций и анализ социальных медиа, при этом, используются и методы машинного обучения для анализа текстов. Как говорилось в пункте 1.1 данной главы, многие решения основаны именно на аппаратной части сбора данных и объектовые системы данного типа редко

освещаются, из-за чего нахождение систем с подобными параметрами становится очень сложной задачей.

Плюсом же разрабатываемой системы, помимо вышеназванных, является то, что она не привязана к конкретной области. Организация может внедрить эту систему и изменить лишь источники сообщений, которые будут анализироваться на ЧС.

2 АНАЛИЗ ДЕЯТЕЛЬНОСТИ УЧРЕЖДЕНИЯ

2.1 Цели и задачи учреждения

Центр развития современных компетенций детей «АмурТехноЦентр» «Дом научной коллаборации имени академика РАН М.Т.Луценко» (далее – ДНК) – структурное подразделение ФГБОУ ВО «Амурский государственный университет» (далее – ФГБОУ ВО «АмГУ»), осуществляет образовательную деятельность по образовательным программам основного, дополнительного образования.

ДНК призван решать следующие задачи:

- развитие современных компетенций у широких слоев населения, в первую очередь у обучающихся и преподавателей образовательных организаций общего, профессионального и дополнительного образования детей посредством оказания им образовательных услуг с использованием современных методов и технологий развития современных компетенций;
- разработка и сопровождение перспективных методов, технологий и образовательных программ развития компетенций, в том числе при участии (в сотрудничестве) международных и российских компаний, в том числе участвующих в создании научных и научно-образовательных центров мирового уровня или обеспечивающих деятельность центров компетенций национальной технологической инициативы.

2.2 Организационная структура учреждения

На упрощенной организационной структуре ФГБОУ ВО «АмГУ» видно, что Учреждение находится под прямым подчинением Проректора по информатизации и новым образовательным технологиям, который в свою очередь, подчиняется ректору ФГБОУ ВО «АмГУ». Всё управление ДНК осуществляет в соответствии с законодательством РФ, приказами ректора и положением о ДНК. Курирует ДНК также Проректор по информатизации и новым образовательным технологиям, но создание, реорганизация,

дополнения и ликвидация осуществляются только ректором ФГБОУ ВО «АмГУ».

В общем случае, ректорат осуществляет организацию работы по исполнению приказов, приведение работ в соответствии с положениями.

Управлением персоналом исполняет приказы, составляет планы и отчеты, согласовывает проекты, приказы и договоры, касательно деятельности рассматриваемого структурного подразделения.

Международный отдел осуществляет координацию и контроль за международными проектами.

Общеобразовательный лицей интегрирует образовательную деятельность.

Управление финансового учета координирует финансовую деятельность.

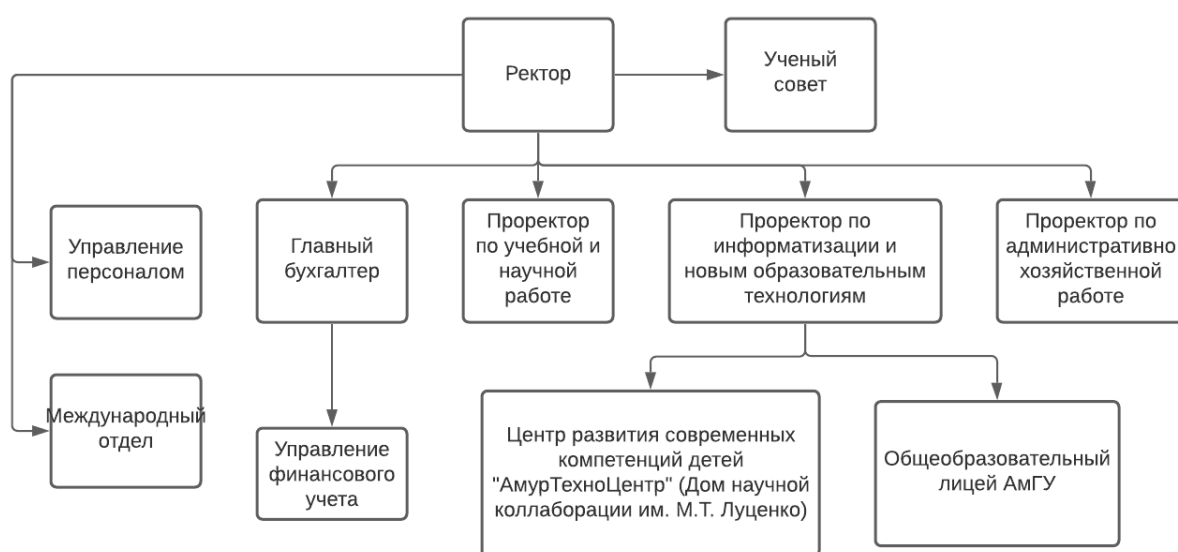


Рисунок 2 – Упрощенная организационная структура ФГБОУ ВО «АмГУ»

На рисунке 3 представлена организационная структура ДНК.

В ДНК работают 22 сотрудника:

- Директор;
- 4 заместителя;
- 12 преподавателей;

- 2 лаборанта;
- 2 администратора;
- Технический специалист.

Лидирующее место в компании занимает Директор, который управляет работой всех сотрудников и практически все документы, полученные в ходе работы ДНК подписываются или проверяются лично им. Директор является полностью ответственным за результаты работы всего ДНК и его обязанности определяются должностными инструкциями и условиями трудового договора. В обязанности директора входят:

1) Осуществления и контроль учебного процесса по реализуемым направлениям центра в соответствии с учебным планом, графиком учебного процесса и расписанием занятий;

2) Организация работы по приему, отчислению обучающихся центра в соответствии с законодательством РФ, уставом ФГБОУ ВО «АмГУ» и другими локальными нормативными актами;

3) Совершенствование работ по обеспечению качества образовательного процесса;

4) Качественное и своевременное выполнение возложенных на ДНК обязанностей, задач и функций;

На одном уровне с директором ДНК находятся его заместители. Заместители не имеют права принятия решений за руководителя. Главная их задача – оказывать помощь в выполнении отдельных функций управления, например, функциях учебного, методического или воспитательного характера.

В прямом подчинении директора находятся: преподаватели, технический специалист, лаборанты и администраторы.

Преподаватели осуществляют образовательную деятельность ДНК, в числе которых могут находиться ведущие преподаватели, магистранты, аспиранты, лаборанты ФГБОУ ВО «АмГУ», педагоги образовательных организаций общего и дополнительного образования, сотрудники

организаций-партнеров, сотрудники частных поставщиков образовательных услуг. Все привлекаемые специалисты должны иметь профессиональное образование и обладать соответствующей квалификацией.

Технический специалист должен своевременно и качественно выполнять порученные ему задачи по техническому и информационному обеспечению, выполнять разработку и поддержку несложных проектов и простых схем, обеспечивая их соответствие техническим заданиям, осуществлять наладку, настройку, регулировку и проверку оборудования на объекте, а также осуществлять сбор, обработку и накопление исходных материалов, данных отчетности, научно-технической информации.

Основным назначением должности лаборанта является оказание помощи преподавателям в организации и проведении учебных занятий, обслуживании и поддержании в рабочем состоянии оборудования учебных кабинетов. Обычно, каждый лаборант закрепляется за определенным кабинетом.

В обязанности администратора ДНК входят:

- обеспечение работы по эффективному и культурному обслуживанию посетителей, созданию для них комфортных условий;
- осуществление контроля над сохранностью материальных ценностей;
- консультация посетителей по вопросам наличия имеющихся услуг;
- принятие мер к предотвращению и ликвидации конфликтных ситуаций;
- рассмотрение претензий, связанные с неудовлетворительным обслуживанием посетителей, и проводит соответствующие организационно-технические мероприятия;
- осуществление контроля чистоты и порядка в помещениях;

- контроль соблюдения работниками организации трудовой и производственной дисциплины, правил и норм охраны труда, техники безопасности, требований производственной санитарии и гигиены;
- информирование руководства организации об имеющихся недостатках в обслуживании посетителей, принимает меры к их ликвидации;
- контроль над исполнением работниками указаний руководства организации;
- выполняет отдельные служебные поручения своего директора.

После внедрения ИС, предполагается, что администраторы будут операторами ИС и тем самым, их круг обязанностей расширится. А в обязанности технического специалиста будут включены поддержка и модификация ИС, настройка рабочих мест и настроек сервера.

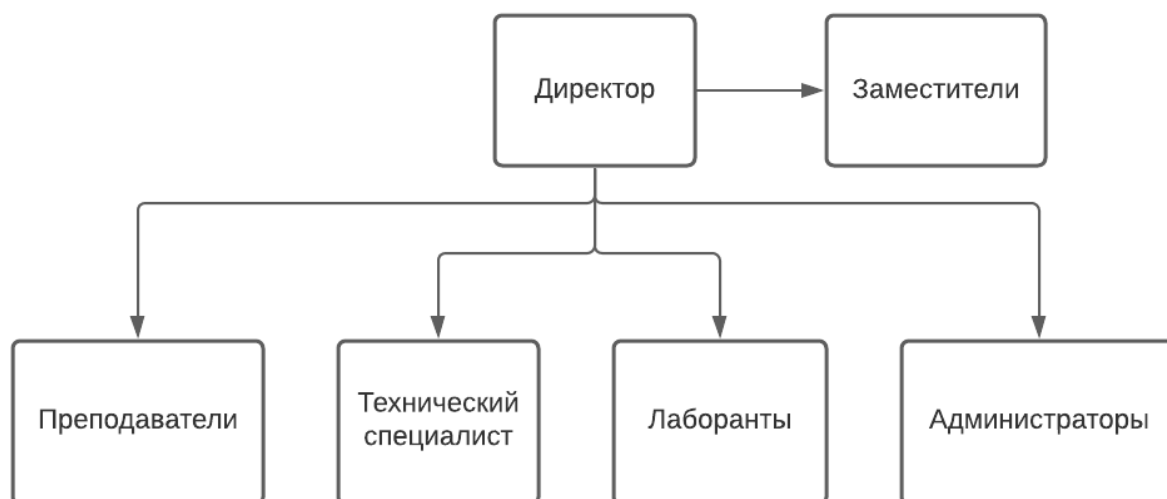


Рисунок 3 – Организационная структура учреждения

2.3 Анализ документооборота

DFD – это методология графического структурного анализа, описывающая внешние по отношению к системе, источники и адресаты данных, логические функции, потоки данных и хранилища данных, к которым осуществляется доступ.

Диаграмма потоков данных это один из основных инструментов структурного анализа и проектирования информационных систем.

2.3.1 Внешний документооборот

Внешний документооборот – это движение документов в правовом пространстве, в котором действуют и реализуют правоотношения различные субъекты права – физические и юридические лица, граждане, предприятия и организации, органы местного самоуправления, органы государственной власти как между однородными по виду субъектами, так и с другими их видами.

Диаграмма внешнего документооборота представляет собой контекстную диаграмму, построенную в нотации DFD. Данная диаграмма представлена в (рисунок А.1 Приложение А).

2.3.2 Внутренний документооборот

Внутренний документооборот – это движение документов внутри предприятия или организации, которые регулируются ведомственными или корпоративными нормативными правовыми актами. Диаграмма также строится в нотации DFD (рисунок А.2 Приложение А). При описании внутреннего документооборота представить основные функции каждого подразделения и рабочего места, охарактеризовать хранилища данных и основные документы, циркулирующие внутри предприятия.

Для построения диаграммы внутреннего документооборота необходимо декомпозировать контекстную диаграмму, т.е. диаграмму внешнего документооборота.

2.4 Анализ использования программно-технических средств

Предприятие оснащено современными средствами хранения, передачи и обработки информации. Локальная сеть, соединяющая все компьютеры и обеспечивающая доступ в сеть Интернет. Программные продукты, используемые предприятием представлены в таблице 1.

Таблица 1 – ПО, используемое на предприятии

Наименование программно-технического средства	Место установки	Основные функции
Kaspersky internet security	Все подразделения	Комплексная защита на всех каналах поступления и передачи информации
ОС Windows 10	Все подразделения	Операционная система
Браузер Google Chrome	Все подразделения	Браузер для просмотра интернет-ресурсов
WinRAR	Все подразделения	Архиватор файлов
Microsoft Office	Все подразделения	Офисный пакет приложений
Discord	Все подразделения	Используется для организации дистанционной работы

2.5 Анализ существующей ИС

В учреждении используется сайт и ИС для учета складских запасов. Они являются отдельными системами, которые используют определенные участники организационной структуры. ИС по автоматической обработке сообщений и идентификации ЧС в учреждении нет. При проектировании и внедрении системы не будет объединений БД или объединений отдельных компонентов систем, так как это 3 независимых ИС.

3 ПРОЕКТИРОВАНИЕ ИНФОРМАЦИОННОЙ СИСТЕМЫ

3.1 Цель и назначение создания системы

Целью создания системы является повышение скорости реагирования на ЧС, обеспечение своевременного оповещения и информирования персонала учреждения, распределение нагрузки между несколькими операторами ИС, снижение временных затрат на мониторинг и классификацию сообщений из разных источников информации организации.

Проектируемая система должна быть ориентирована на решение следующих задач:

- автоматизированное получение сообщений;
- автоматизированная обработка и бинарная классификация сообщений по факту ЧС, выявление угрозы;
- сортировка сообщений по факту ЧС, группировка данных, вероятности ответа системы о принадлежности сообщения к ЧС и по дате получения этих сообщений;
- автоматическое оповещение ответственных должностных лиц с последующим информированием работников, находящихся в зоне действия ЧС;
- распределение нагрузки между операторами ИС и повышение скорости реагирования на ЧС;
- хранение и визуализация данных для удобной работы с ними оператора ИС;
- автоматическое резервное копирование по расписанию;
- ведение логов об ошибках и действиях в системе;
- управление пользователями ИС с учетной записи администратора;
- предоставление отчётов на основе данных о ЧС за выбранный период.

3.2 Разработка проекта автоматизации

3.2.1 Выбор модели жизненного цикла ПО

Жизненный цикл (ЖЦ) ИС – это период разработки и эксплуатации ИС, который начинается с момента возникновения потребности в ИС и заканчивается на этапе её выхода из эксплуатации.

Отношения между системой и ее системными элементами обычно рассматривается как иерархия до простейших ее составляющих. Для сложных систем, системные элементы на каждом отдельном этапе можно рассматривать как систему, чтобы упростить процесс разработки. Таким же способом рекурсивно применяются к рассматриваемой системе и соответствующие процессы жизненного цикла системы, хоть и не всегда система подразумевает иерархические отношения, например, как в сетях или иных распределенных системах.

Под моделью ЖЦ понимается структура, которая определяет взаимосвязи задач, действий и процессов на протяжении всего ЖЦ. Основными моделями ЖЦ, которые используются в настоящее время, являются: каскадная, инкрементная и спиральная.

Каскадная и инкрементная модели ЖЦ совпадают и включают в себя следующие этапы:

- анализ требований;
- проектирование;
- разработка;
- тестирование и отладка;
- ввод в действие;
- эксплуатация и сопровождение.

Каскадная модель подразумевает разработку ПО, когда все этапы разработки происходят последовательно, при этом, каждый последующий шаг начинается только после полного завершения выполнения предыдущего шага. Результатам каждого шага является промежуточный продукт, который не

может изменяться на последующих шагах. Данная модель удобна для разработки с четкими и неизменяемыми в течении ЖЦ требованиям. Также, данная модель пригодна при разработке продукта подразумевающего выпуск новой версии этого продукта или для ПО такого типа, которые разрабатывались разработчиками ранее.

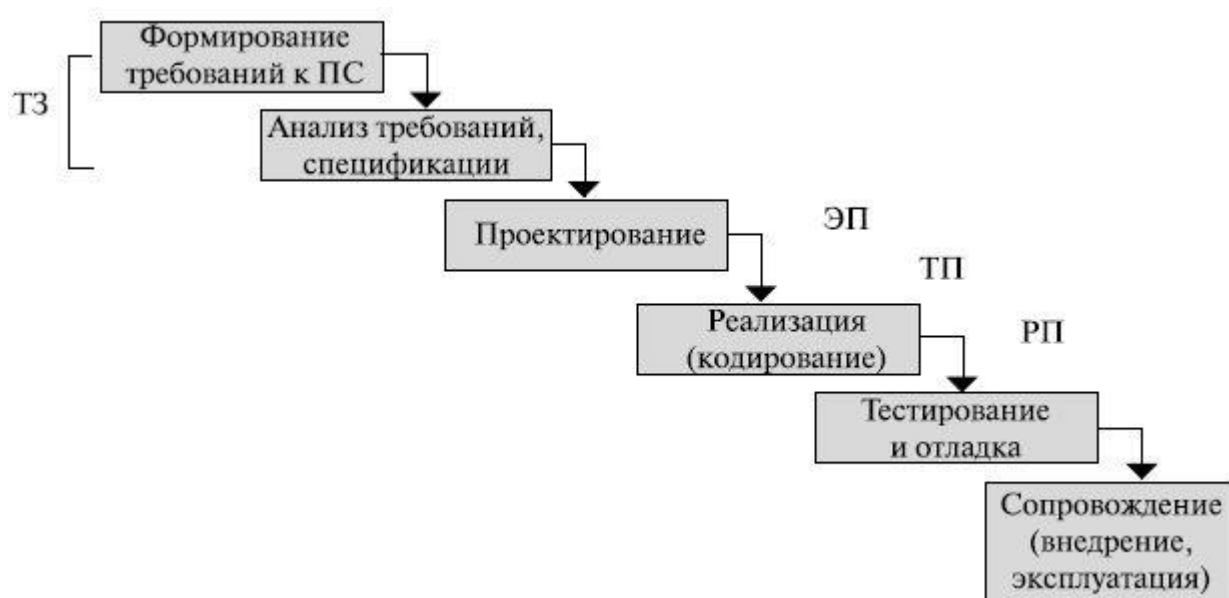


Рисунок 4 - Каскадная модель ЖЦ ПО

Инкрементная модель похожа на каскадную, но в данном случае, разработка ведётся в несколько инкрементов (версий), которые подразумевают улучшение продукта, пока ЖЦ разработки ПО не прекратится. Это означает, что на каждом этапе ЖЦ возможны обратные связи, которые являются межэтапными корректировками. При этом, процесс ЖЦ растягивается на весь период разработки. Работа над проектом начинается с определения основных требований к системе. Далее, разработчик по принципу приращений, добавляют определенную функциональность в систему, при этом, сначала разрабатывает компоненты с наивысшим приоритетом и постепенно их детализирует. В то же время, возможны уточнение требований других частей системы. При разработке определенно работающего компонента, он предоставляется клиенту, который может уточнить требования на основе использования этого компонента.



Рисунок 5 – Инкрементная модель ЖЦ ПО

Недостатки и достоинства данной модели такие же, как и у каскадной, но в данном случае, заказчик может раньше увидеть результат и скорректировать требования, при этом, корректировка требований не является настолько затратной процедурой, по сравнению с каскадной моделью. Основным недостатком является ухудшение структуры системы, когда при добавлении новых компонентов и изменение требований ухудшают структуру системы. Чтобы избежать этого, нужно дополнительное время на рефакторинг.

При ЖЦ *спиральной модели* на каждом витке спирали выполняется создание очередной версии продукта. На каждой версии происходит уточнение требований проекта и планируются работы следующего витка, если необходимое качество продукта ещё не достигнуто. Анализ и проектированию уделяются особое внимание на начальных этапах разработки, так как это позволяет легче проверять технические решения и реализуемость конкретного проекта и таким образом можно обосновать создание следующего прототипа. Данная модель позволяет уделять внимание рискам, влияющим на организацию жизненного цикла, что является недостатком для проектов, имеющих низкую степень риска, так как оценка рисков на каждом витке ведёт к большим затратам. Модель сочетает в себе возможности модели прототипирования и каскадной модели. Главная же задачи, как можно быстрее предоставить заказчику работоспособный продукт.



Рисунок 6 – Спиральная модель ЖЦ ПО

Целью дипломного проекта является разработка информационной системы для анализа сообщений и отнесение их к определенной группе с помощью методов машинного обучения. Так как на этапе кодирования возможны определенные изменения и при этом, необходимо предъявлять заказчику каждый модуль и согласовывать с ним необходимые изменения, например, для обработки естественного языка и точности построенной модели, а также для интерфейса и функционала приложения как для эксперта, так и для администратора ПО, было принято решение использовать инкрементную модель ЖЦ.

На этапе анализа необходимо собрать информацию об основных источниках текстовых сообщений в Учреждении, посредством которых преподавательский состав и ученики делятся информацией.

На основе анализа основных источников сообщений, происходит этап проектирования, где программисты формируют структуру программы и проектируют базу данных (БД).

На этапе реализации, программисты создают базу данных и разрабатывают отдельные компоненты приложения. При окончании

разработки отдельного компонента системы, она поставляется клиенту для уточнения требований следующих компонентов. Если есть изменения, то данный компонент дорабатывается, уточняются требования для следующих компонентов и продолжается их разработка. Ключевые этапы этого процесса — простая реализация подмножества требований к программе и совершенствование модели в серии последовательных релизов до тех пор, пока не будет реализовано ПО во всей полноте.

Этап тестирования и отладки подразумевает проведение предварительных испытаний, опытной эксплуатации и приемочных испытаний, которые описаны в Приложении А «Техническое задание».

На этапе ввода в действие необходимо установить серверные и клиентские приложения на рабочие станции и настроить локальную сеть, после чего провести окончательное тестирование системы на рабочих станциях и убедиться в работоспособности всех компонентов системы.

Для корректной эксплуатации системы следует разработать руководство пользователей и провести инструктаж сотрудников. В соответствии с функционалом внедряемой системы, полномочия некоторых сотрудников могут быть изменены, о чём сотрудники должны быть заранее проинформированы.

3.2.2 Ожидаемые риски на этапах ЖЦ и их описание

Хоть инкрементная модель ЖЦ не настолько критична к изменению требований заказчика, по сравнению с каскадной моделью, на этапе анализа необходимо точно определить все каналы информации, которые будут поступать в систему. Если какой-то канал информации будет не учтён, то эффективность ИС сильно снижается.

На этапе проектирования нужно учесть все риски, которые были выявлены на этапе анализа и правильно спроектировать систему. Например, сервер, который анализирует входящую информацию из источников

сообщений, должен отправлять конкретное сообщение только одному эксперту, во избежание дублирование информации в БД и конфликта записи.

На этапе реализации необходимо минимизировать возможность совершить пользователю ошибочное действие или предотвратить его, пока эта ошибка не начала влиять на работоспособность системы.

На этапе тестирования и отладки необходимо учесть все сценарии использования ИС, как со стороны эксперта, так и со стороны администратора ИС. Также, необходимо обнаружить нежелательное поведение ИС и устранить его. При тестировании программы, в идеале, её должны опробовать как можно большее количество пользователей.

На этапе внедрение необходимо проконтролировать правильную установку на рабочие станции и использовать только лицензионное ПО для предотвращения несанкционированного доступа.

На этапе эксплуатации необходимо обеспечить правильное и поэтапное обучение персонала и провести контроль знаний по эксплуатации ИС.

3.3 Структура ИС

Функциональная модель программы в нотации IDEF0 представлена на рисунке 7.



Рисунок 7 – Функциональная модель программы в нотации IDEF0

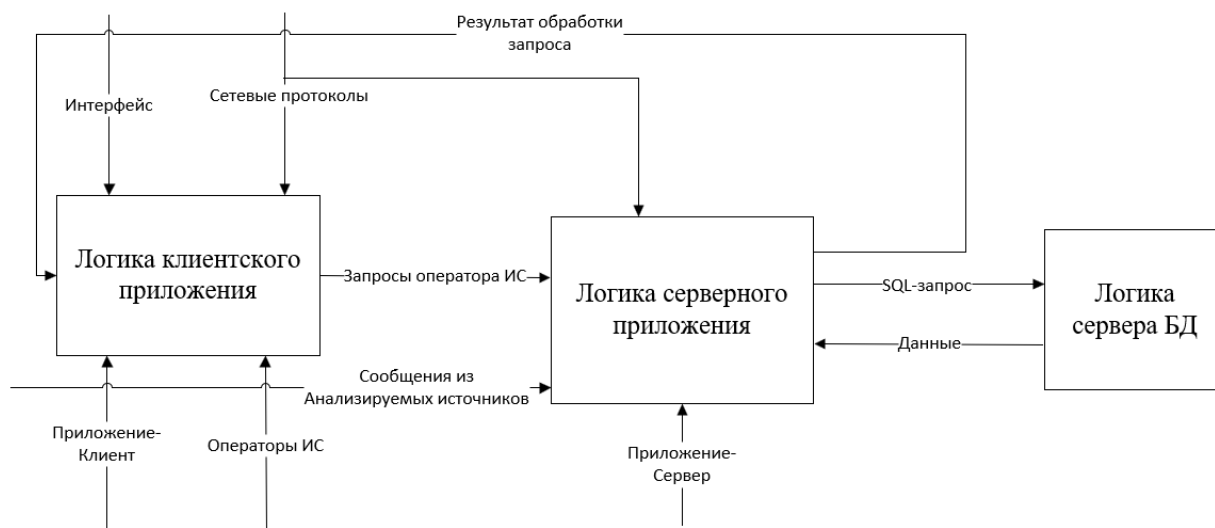


Рисунок 8 – Декомпозиция функциональной модели программы

Разрабатываемая ИС является трехуровневым клиент-серверным приложением. Все взаимодействие между клиентом и сервером строится за счёт запросов.

Декомпозиция логики клиентского и серверного приложений представлены в приложении Б (рисунок Б.1 и Б.2 соответственно).

3.4 Характеристика функциональных подсистем программы

Для начала работы пользователю необходимо авторизоваться, путём отправки данных на сервер и получение успешного ответа от него. После авторизации, в зависимости от типа учетной записи (администратор или эксперт), пользователю предоставляется соответствующий функционал приложения.

Эксперт может анализировать сообщения, которые были обработаны сервером и подтверждать или отклонять факт ЧС, получить отчёт, в соответствии с шаблоном, о конкретном сообщении, дополнив его необходимой информацией прямо из приложения, печатать отчёт, визуализировать данные по ЧС за определенный период или всё время, с выбором групп для анализа. При этом, можно отменить вывод и получение сообщений с привязкой к конкретному эксперту, но подтверждать или

отклонять факт ЧС может только тот эксперт, которому было адресовано конкретное сообщение. При необходимости, эксперт может оставить приложение включенным, но отключить режим прослушивания сервера. Это необходимо, если эксперту нужно покинуть своё рабочее место, и он не сможет своевременно отреагировать на сообщения, классифицированные системой как ЧС. Также, пользователь сможет менять некоторые настройки, такие как: количество сообщений контекста, темная или светлая тема приложения, выводить ли только сообщения, которые не были рассмотрены экспертом и т.п.

Администратор ИС может менять строгость классификации системы в соответствии с 3-мя уровнями: обычный, повышенный и строгий. Он также может вносить новых экспертов в ИС, делать резервные копии и настраивать их автоматическое сохранение, управлять логами. Ему также доступны функции визуализации данных и формирование отчетов. Администратор ИС в праве отключать некоторые источники анализируемых сообщений, в данном случае, чаты в мессенджере Telegram при необходимости.

На основании перечисленных выше функций, можно выделить следующие подсистемы, формирующие структуру ИС:

а) Сервер:

1) Подсистема авторизации. Включает в себя идентификацию пользователя и выдачу соответствующих ему прав, в зависимости от типа учетной записи.

2) Подсистема получения сообщений из анализируемых источников. Включает в себя сбор сообщений с помощью API бота, находящимся в анализируемых чатах Telegram.

3) Подсистема обработки и классификации сообщений. Включает в себя предобработку сообщений, написанных на естественном языке, перевод отдельных токенов в векторную форму, классификацию сообщений.

4) Подсистема работы с БД: запись и хранение данных в БД, чтение этих данных с помощью средств СУБД.

5) Подсистема оповещение должностных лиц на терминалы связи: отправка запросов на рассмотрение сообщений активным в текущий момент операторам ИС.

б) Клиент:

1) Подсистема интерфейса пользователя. Данная подсистема реализована на стороне клиентского приложения. Основной задачей данной подсистемы является предоставление графического интерфейса, реагирующего на действия пользователя или сообщения от сервера и уведомление пользователя о некорректных действиях или сбоях в работе приложения. Подсистема пользовательского интерфейса не генерирует никаких данных, она лишь получает данные от других подсистем.

2) Подсистема авторизации. Обеспечивает функционал по отправке запроса об авторизации на сервер. Поддерживает ожидание ответа об успехе данной операции.

3) Подсистема функционирования клиентской логики. Реализует основной функционал по работе с сообщениями и управления самим сервером и пользователями: просмотр данных о сообщении, в том числе контекста, подтверждение или отклонение факта ЧС, пользовательские настройки, настройки конфигурации сервера и классификатора, настройки резервных копий, настройки логирования и т.д.

3.4.1 Описание модуля авторизации

Авторизация – это процесс, при котором определенное лицо или группа лиц подтверждают право на совершение определенных в системе операций.

Поскольку администратор системы имеет право менять настройки сервера удалено, при этом, уменьшить строгость классификации сообщений, для предотвращения несанкционированного доступа к данным функциям необходим модуль авторизации. Для клиента он также имеет значение, так как

в БД есть поля, какой эксперт и когда подтвердил или отклонил факт ЧС. Это обосновывается в главе 2, где по итогам анализа, была подтверждена необходимость обеспечения надежной защиты ИС.

В связи с этим, принято решение использовать технологию JWT открытого стандарта RFC 7519 для подсистемы авторизации. JSON Web Token (JWT) – это токен доступа, который основан на формате файлов JSON. Он позволяет обеспечить безопасный способ передачи информации между двумя участниками. Схема взаимодействия представлена на рисунке 9.

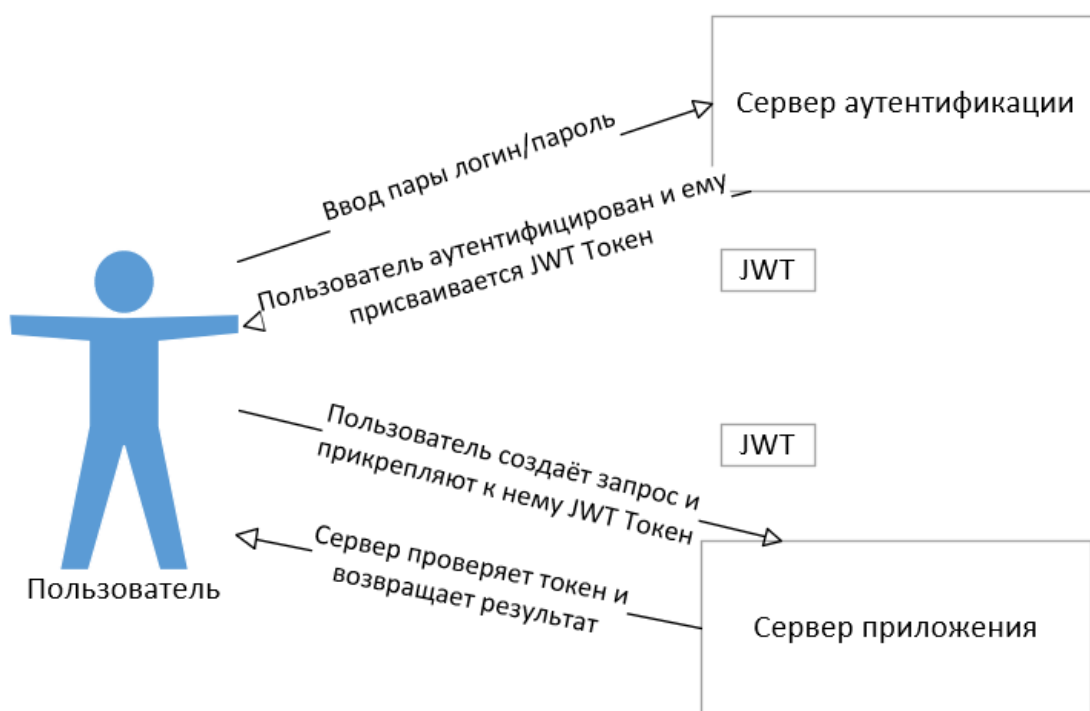


Рисунок 9 – Схема авторизации на основе JWT

Токен JWT состоит из 3-х составляющих: заголовка (header), полезной нагрузки (payload) и подписи. В заголовке указывается необходимая информация для самого токена. В данном файле содержится 1 обязательный аргумент alg, в котором определен используемый алгоритм для шифрования. В payload указывается пользовательская информация, например, имя пользователя, идентификатор пользователя, его роль в системе и т.д. Все

параметры данной секции являются необязательными. И далее необходимо сформировать подпись. Она кодирует заголовок и полезную нагрузку алгоритмом base64url, соединяя полученные строки через точку, а затем полученная строка хешируется с использованием алгоритма из заголовка.

Функциональная диаграмма IDEF0 подсистемы авторизации представлена на рисунках 10 и 11.



Рисунок 10 - Функциональная модель авторизации пользователя в нотации IDEF0

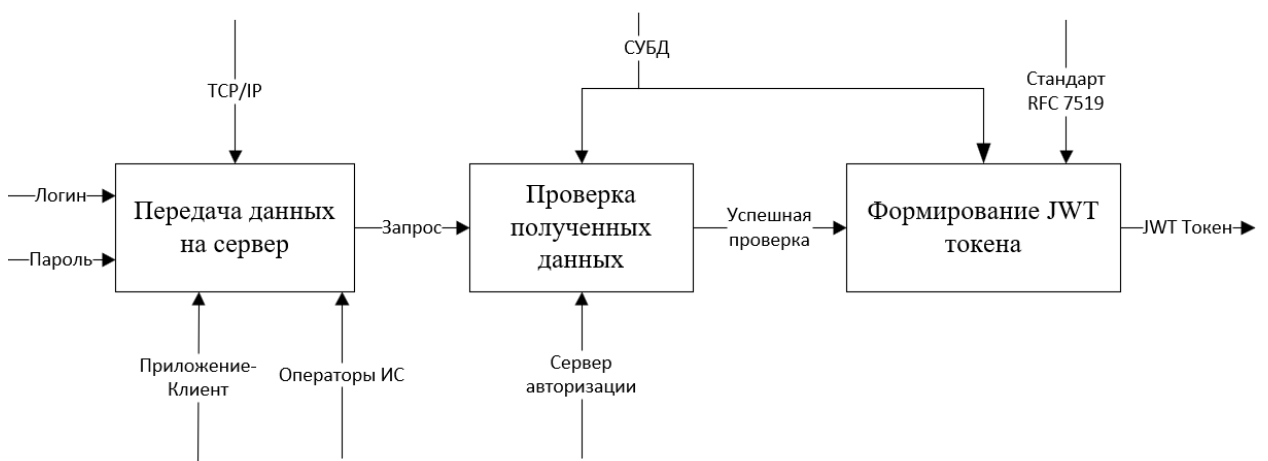


Рисунок 11 - Декомпозиция верхнего уровня подсистемы авторизации

Входными данными для подсистемы авторизации являются логин и пароль. В ходе работы подсистемы, данные попадают на сервер, анализируются и если проверка данных прошла успешно, то формируется JWT токен, который и отправляется пользователю. Регламентирующими документами, в данном процессе, являются: протоколы транспортного и сетевого уровня TCP и IP, стандарт RFC 7519, который описывает процедуру генерации JWT и СУБД.

3.5 Проектирование БД

Исходя из того, что для работы с приложением нужно пройти этап авторизации, а данные об операторах ИС и анализируемых сообщениях должны безопасно и надежно храниться, необходимо спроектировать БД.

3.5.1 Инфологическое проектирование

Инфологическое проектирование состоит из нескольких этапов. Первый этап - формирование набора сущностей. Далее приводится перечень сущностей с краткой характеристикой.

- 1) Сущность «Сотрудник» хранит данные о пользователях ИС;
- 2) Сущность «Предсказанные сообщения» хранит сообщения и информацию о сообщениях, полученных из анализируемых источников;
- 3) Сущность «Роль» хранит информацию о ролях в системе.

Второй этап – формирование спецификации атрибутов каждой сущности. Спецификация имеет вид таблицы:

Таблица 2 – Спецификация атрибутов сущности «Роль»

Название атрибута	Описание атрибута	Тип данных	Диапазон значений	Пример атрибута
ИД роли	Число, определяющее роль в системе	Числовой	>0	7
Наименование роли	Название роли	Текстовый	-	Петров П.П.

Таблица 3 – Спецификация атрибутов сущности «Сотрудник»

Название атрибута	Описание атрибута	Тип данных	Диапазон значений	Пример атрибута
ID сотрудника	Число, однозначно идентифицирующее каждого работника	Числовой	>0	7
Фамилия	ФИО сотрудника	Текстовый	-	Петров П.П.
Имя пользователя	Уникальное имя сотрудника в системе	Текстовый	-	Administrator
Хеш пароля	Хеш пароля работника	Текстовый	-	a82f4c8f29e8
ID роли	Код роли сотрудника в системе	Числовой	>0	2
Номер телефона	Телефон для связи с сотрудником	Числовой	==11	88005553535
Дата последнего входа в систему	Дата, показывающая, когда пользователь в последний раз вошел в систему	Дата	≤текущая дата	2021-06-01 00:00:00
Дата выхода из системы	Дата, показывающая, когда пользователь вышел из системы	Дата	≤текущая дата	2021-06-01 18:30:43

Таблица 4 – Спецификация атрибутов сущности «Предсказанные сообщения»

Название атрибута	Описание атрибута	Тип данных	Диапазон значений	Пример атрибута
1	2	3	4	5
ID сообщения	Число, однозначно идентифицирующее каждое сообщение	Числовой	>0	5324
Текст сообщения	Текст сообщения	Текстовый	-	Пожар на улице N

Продолжение таблицы 4

1	2	3	4	5
Название чата	Название чата, откуда пришло сообщение	Текстовый	-	Рабочая группа
Тип чата	Тип чата	Текстовый	-	Группа
Отправитель	Отправитель сообщения	Текстовый		Username
Дата отправки	Дата отправки сообщения	Дата	≤текущая дата	2021-04-01 10:35:38
Вероятность	Вероятность ЧС	Числовой	$0 \leq P \leq 1$	0.565
Предсказанное значение	Ответ системы о принадлежности сообщения к ЧС	Булевый	-	1
Действительное значение	Ответ эксперта о принадлежности сообщения к ЧС	Булевый	-	0
Дата проверки сообщения	Дата проверки сообщения экспертом	Дата	≤текущая дата	2021-04-01 10:45:21
ID сотрудника	ID эксперта, который рассматривает сообщение	Числовой	>0	7

Третий этап инфологического проектирования – выбор и обоснование первичного ключа, который однозначно идентифицирует каждую запись таблицы.

Для сущности «Сотрудник» первичным ключ будет являться «ID сотрудника», так как такой первичный ключ однозначно идентифицирует работника учреждения.

Для сущности «Предсказанные сообщения» первичным ключ будет являться «ID сообщения», так как такой первичный ключ однозначно идентифицирует каждое отдельное сообщение.

Для сущности «Роль» первичным ключ будет являться «ID роли», так как такой первичный ключ однозначно идентифицирует каждую роль в системе.

Последний этап – построение инфологической модели БД, которое позволяет обеспечить интегрированное представление о предметной области. Моделирование локального представления заканчивается графическим представлением всех выявленных сущностей, связей между ними и атрибутов с использованием любой из известных нотаций. В данной работе выбрана Нотация Чена. Итоговая диаграмма в нотации Чена представлена в приложении В1.

3.5.2 Логическое проектирование

Во время логического проектирования необходимо провести нормализацию БД. Всего нормальных форм 8, но часто достаточным является соблюдение первых трех.

Основным правилом первой нормальной формы (1НФ) является необходимость неделимости значения в каждом поле (столбце) строки – атомарность значений. В данном случае, поле ФИО в таблице Сотрудник рассматривается как неделимая сущность, так как нет необходимости работать, например, с именем, по отдельности. В остальных полях атомарность также соблюдается. Так же, строки таблицы не зависят друг от друга и все строки являются уникальными, что обеспечивается за счёт первичного ключа.

Условием второй нормальной формы (2НФ) является отсутствие зависимости неключевых полей от части составного ключа, при этом, должны соблюдаться все правила первой нормальной формы. 2НФ также соблюдена, так как отсутствуют неключевые атрибуты, которые вообще не зависят от потенциального ключа.

Требование третьей нормальной формы (3НФ) заключается в том, чтобы в таблицах отсутствовала транзитивная зависимость. Транзитивная зависимость – это когда неключевые столбцы зависят от значений других неключевых столбцов, то есть неключевые столбцы не должны брать на себя функцию первичного ключа. 3НФ также соблюдается.

Окончательная диаграмма IDEF1X логической модели БД представлена в приложении В2.

3.5.3 Физическое проектирование

Физическая модель базы данных — это модель данных, которая определяет, каким образом представляются данные, и содержит все детали, необходимые СУБД для создания базы данных.

На основании логической модели спроектируем физическую модель БД:

Таблица 5 – Схема отношения Role

Содержание поля	Имя поля	Формат данных	Индексация
<u>ID роли</u>	ID_role	Integer	Первичный ключ
Наименование роли	name	Varchar(30)	Обязательное поле

Таблица 6 – Схема отношения Employee

Содержание поля	Имя поля	Формат данных	Индексация
<u>ID сотрудника</u>	ID_employee	Integer	Первичный ключ
ФИО	full_name	Varchar(100)	Обязательное поле
Имя пользователя	username	Varchar(100)	Обязательное поле
Хеш пароля	password_hash	Varchar(1000)	Обязательное поле
Номер телефона	phone_number	Varchar(11)	Обязательное поле
Дата последнего входа в систему	last_login_date	Datetime	Обязательное поле
Дата выхода из системы	last_logout_date	Datetime	Обязательное поле
<u>ID роли</u>	ID_role	Integer	Внешний ключ

Таблица 7 – Схема отношения PredictedMessage

Содержание поля	Имя поля	Формат данных	Индексация
1	2	4	5
<u>ID сообщения</u>	ID_message	BigInt	Первичный ключ
Текст сообщения	message_text	Varchar(MAX)	Обязательное поле
Название чата	chat_name	Varchar(100)	Обязательное поле
Тип чата	chat_type	Varchar(30)	Обязательное поле

Отправитель	sender	Varchar(100)	Обязательное поле
Дата отправки	sending_date	Datetime	Обязательное поле
Вероятность	probability	Float	Обязательное поле
Предсказанное значение	predicted_value	Bit	Обязательное поле
Действительное значение	actual_value	Bit	Необязательное поле
Дата проверки сообщения	closing_date	Datetime	Необязательное поле
<u>ID сотрудника</u>	ID_employee	Integer	Внешний ключ

Физическое проектирование также предусматривает построение структуры физической модели. Данная диаграмма представлена на (рисунке В3 приложения В).

В результате проектирования, была построена нормализованная по трем нормальным формам БД, которая также отвечает всем необходимым требованиям, для использования её в функционировании ИС.

3.6 Задачи автоматизации, решаемые разрабатываемой ИС

При разработке ИС решались следующие задачи, которые подлежали автоматизации:

- 1 Сбор и хранение сообщений из источников информации;
- 2 Обработка сообщений с целью определение факта ЧС с использованием методов машинного обучения для NLP;
- 3 Группировка данных. Сообщения делятся на 3 основные группы:
 - а) Сообщение с низкой вероятностью ЧС. В эту группу попадают сообщения, вероятность ЧС, которых менее 35%;
 - б) Сообщение со средней вероятностью ЧС. В эту группу попадают сообщения, вероятность ЧС, которых менее 50%, но не менее 35%;
 - в) Сообщение с высокой вероятностью ЧС. В эту группу попадают сообщения, вероятность ЧС, которых равна или более 50%.

4 Сортировка данных. Например, сначала выводятся самые новые сообщения, при этом, выводятся сообщения, ещё не рассмотренные экспертом. Есть возможность не выводить сообщения, уже проверенные экспертом или убрать группировку по уже рассмотренным сообщениям. Также, возможна сортировка не только по дате поступления сообщения, но и по таким полям, как: вероятность, тип чата и отправитель;

5 Установка автоматического подтверждения ответа системы МО как ответа эксперта. Можно установить полностью автоматический режим, когда эксперт доверяет ответ системе. Режим устанавливается для каждой группы ЧС отдельно. Этот режим особенно полезен для классификации сообщений с низкой вероятностью ЧС;

6 Логирование действий пользователя ИС и ошибок, происходящих в системе;

7 Автоматическое создание резервных копий БД;

8 Автоматическое удаление сообщений с низкой вероятностью ЧС из БД;

9 Автозаполнение шаблона документа о конкретной ЧС.

4 РАЗРАБОТКА ИС

4.1 Описание разрабатываемой ИС

Данная информационная система создается для автоматизации обработки сообщений из источников этих сообщений. Программный продукт автоматизировано определяет факт наличия ЧС в тексте и уведомляет об этом оператора ИС. Данный проект является учебным и выполняется без привлечения каких-либо финансовых средств. Результаты работы предъявляются Заказчику в виде:

- функционирующей ИС, представляющего собой приложение с графическим интерфейсом;
- исходного программного кода

ИС должна быть централизованной, т.е. все данные должны располагаться в центральном хранилище. При этом, система должна иметь трехуровневую клиент-серверную архитектуру (клиент, сервер приложений, сервер БД). Сервер приложений должен иметь постоянный доступ к сети интернет для сбора информации из источников сообщений. Предполагается, что сервер собирает информацию с помощью API анализируемого сервиса. Для передачи информации от сервера к клиенту необходимо, чтобы они находились в одной локальной сети. В качестве протоколов взаимодействия используются протоколы транспортного и сетевого уровня TCP и IP соответственно. Сама ИС должна соответствовать требованиям, предъявляемым в техническом задании.

4.2 Выбор средств разработки

4.2.1 Выбор СУБД

Основными факторами при выборе СУБД для разрабатываемого приложения являлись:

- Параллелизм запросов. Поскольку операторы ИС могут запрашивать данные из БД одновременно, необходимо обработать сразу два

запроса, а не заставлять ждать одного из операторов. В связи с этим, пришлось отказаться от файл-серверных и встраиваемых СУБД, которые не удовлетворяют данному требованию;

- Производительность. От параллелизма не будет толку, если каждый запрос выполняется слишком долго;

- Повышенная безопасность и централизованное управление. Благодаря той же клиент-серверной архитектуре, приложения-клиенты не имеют непосредственного доступа к файлам данных. Доступ к БД осуществляется с помощью специально выделенного сервера;

Всем вышеперечисленным требованиям хорошо удовлетворяет РСУБД MS SQL Server, разработанная корпорацией Microsoft, которая постоянно поддерживает свой продукт. Помимо этого, данная СУБД поддерживает такие эффективные механизмы, как: триггеры, хранимые процедуры и функции, которые обеспечивает более гибкое взаимодействие с БД.

4.2.2 Выбор языка программирования

Исходя из выше установленных требований к ИС и совокупности функций, которые она должна выполнять, необходимо выбрать средства разработки с поддержкой РСУБД MS SQL Server, библиотеки для работы с большим количеством текстовых данных, их обработки и представлении в числовой форме, с дальнейшей возможностью применения методов машинного обучения к ним. При этом, важна скорость обработки этих данных. Также, необходимы библиотеки для разработки пользовательского интерфейса и серверной части приложения.

Для работы с большими данными наиболее часто используются 2 языка программирования: R и Python.

R обычно применяется в тех случаях, когда для анализа данных требуются выделенные вычислительные мощности или отдельные сервера. R отлично подходит для исследовательской работы, удобен практически при любом варианте анализа данных, поскольку в языке R существует масса

пакетов и готовые тесты, обеспечивающие нужный инструментарий для быстрого старта.

Достоинства языка R:

- широко используется для визуализации данных и имеет множество встроенных библиотек, такие как: ggplot2, ggvis, googleVis и rCharts;

- Язык создавался специально для анализа данных;

- богатая экосистема МО, огромное количество библиотек статистических методов. R широко используется в академической среде, поэтому новые методы и разработки внедряются именно в него.

- имеет мощную математическую базу, так как разрабатывался статистиками для статистиков.

Недостатки языка R:

- считается довольно медленным языком. Есть дополнительные пакеты, которые позволяют повысить производительность программ, но для языка всё равно нужны выделенные вычислительные мощности;

- язык R сложен в изучении, а кривая обучения языку R нетривиальна.

Python пригодится в случаях, когда задачи, связанные с анализом данных, вплетаются в работу веб-приложений, или если статистический код требуется инкорпорировать в рабочую базу данных. Python, будучи полнофункциональным языком программирования, отлично подходит для реализации алгоритмов с их последующим практическим использованием.

Достоинства языка Python:

- является универсальным языком программирования с простым и изящным синтаксисом. Это позволяет быстро писать программы, разрабатывать работающие прототипы приложений;

- является многоцелевым языком программирования. Python позволяет писать приложения абсолютно разных направленностей, будь то

веб-приложения, игры, скрипты для автоматизации задач или работа с данными;

- высокое качество документации;
- огромное количество сторонних пакетов, которые можно быстро устанавливать и использовать для облегчения написания, тестирования и отладки кода.

Недостатки языка Python:

- имеет динамическую типизацию, что увеличивает скорость разработки, но затрудняет поиск некоторых трудно отслеживаемых ошибок;
- визуализация данных значительно проигрывает по сравнению с визуализацией данных в R.

Языком для написания модуля, был выбран Python, благодаря высокой скорости разработки, большому количеству пакетов и встроенному объектно-ориентированному подходу. Python оперирует объектами. Это значит, что всё, даже типы данных, являются объектами, так что такой подход наиболее естественен для данного языка.

В качестве основных модулей, которые используются в разработке, являются:

- Pandas вкупе с Numpy позволят представить данные в виде таблицы и сделать первичную обработку данных;
- SpaCy для расширенной обработки естественного языка;
- Matplotlib и seaborn для визуализации данных;
- Scikit-learn для использования методов машинного обучения;
- Twisted для написания серверной логики приложения;
- PyQt5 для разработки пользовательского интерфейса;
- Threading для выполнения нескольких операций одновременно;
- Telebot для получения сообщений с помощью API бота, который находится в анализируемых группах.

В качестве инструмента для разработки экранных форм был выбран Qt Designer. Qt Designer – инструмент для проектирования и создания графических пользовательских интерфейсов (GUI) из компонентов Qt. Принцип работы в среде реализуется по принципу «What you see is what you get», WYSIWYG, «что вы видите, то и получаете» (рис. 12).

Виджеты и формы, созданные с помощью Qt Designer, интегрированы с управляющим кодом, использующий механизм сигналов и слотов Qt, который позволяет легко установить обработчики событий для элементов графики. Все свойства в Qt Designer изменяются динамически внутри программного кода.

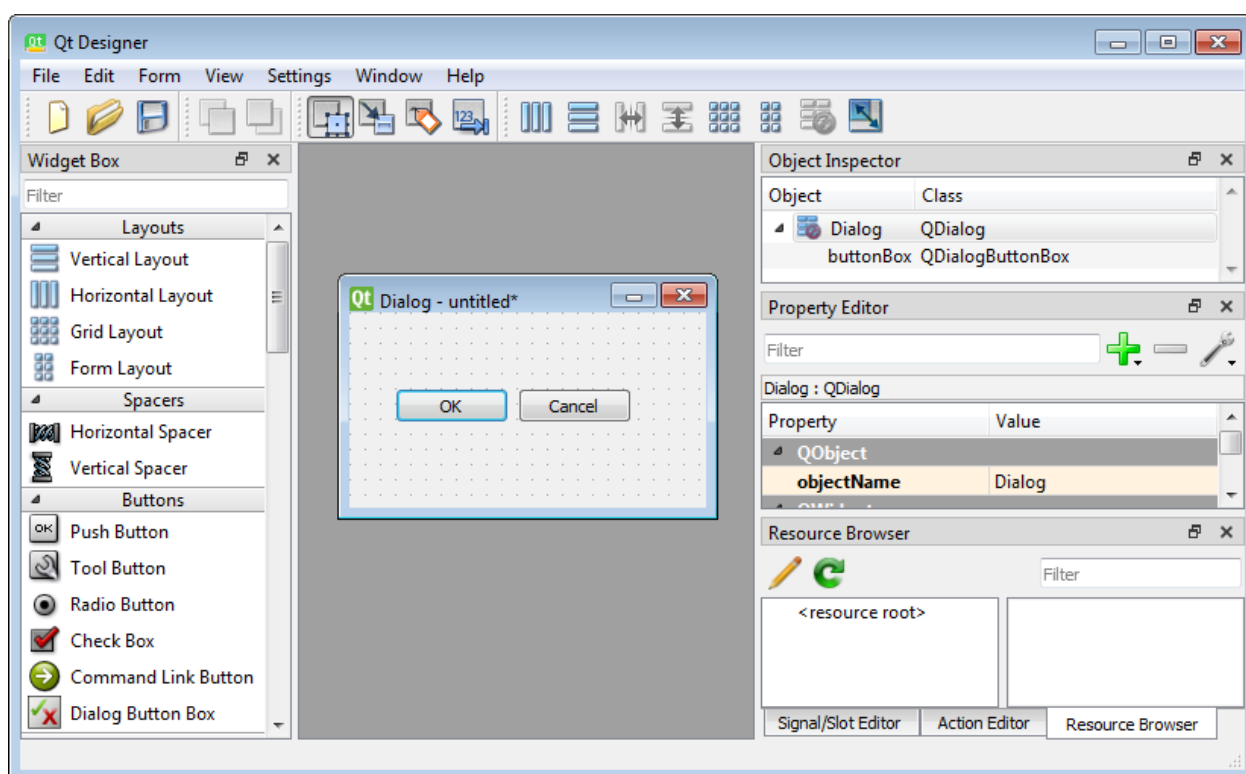


Рисунок 12 – Интерфейс Qt Designer

4.3 Работа с данными и обучение классификатора

Первым этапом разработки данного приложения стал анализ данных. Данные были собраны из социальной сети «Twitter» американской компанией «Figure Eight», которая специализируется на машинном обучении и искусственном интеллекте. Каждая запись содержит в себе числовой идентификатор сообщения, текст сообщения и метку, которая даёт

информацию о том, является ли сообщение ЧС или нет. В наборе данных 1 – это класс «ЧС», а 0, соответственно, не «ЧС». Пример неочищенных данных представлен на рисунке 13.

	russian_text	target
164	Эксперты во Франции приступили к изучению обломков самолетов, обнаруженных на острове Реюньон. В среду французские эксперты по авиационным происшествиям приступили к изучению обломков самолетов.	1
577	Jacksonville Busines FedEx прекращает поставки потенциальных патогенов биотеррора http://t.co/sHzsYmaUSi	1
618	Создание ИИ\n Изменение климата\n Биотерроризм\n Массовая автоматизация персонала\n Контакт с другой жизнью\n Неравенство богатства\n\n Да, у нас это легко	0
622	Добровольцы должны были принять участие в учениях по обеспечению готовности к чрезвычайным ситуациям, имитирующим биотеррористическую катастрофу http://t.co/NWV2RvGHf3 @HVnewsnetwork	1
710	Подпишитесь на @EdWelchMusic и посмотрите его хит-сингл Unpacked Man its BLAZING	0

Рисунок 13 – Неочищенные данные для обучения

Первым делом нужно проверить, является ли выборка достаточно сбалансированной. Это значит, что, если количество сообщений одного из классов, сильно доминирует над количеством сообщений другого класса, то модель может не найти необходимых зависимостей. Например, если количество сообщений связанных с темой ЧС было бы слишком мало, то система не смогла бы эффективно относить новые сообщения к ЧС, так как она не способна находить отличительные особенности таких сообщений.

Большинство наборов классификационных данных не имеют точно одинаковое количество экземпляров в каждом классе, но небольшая разница часто не имеет значения.

Есть проблемы, когда дисбаланс классов не просто распространен, это ожидается. Например, в наборах данных, которые характеризуют мошеннические транзакции, они не сбалансированы. Подавляющее большинство транзакций будет в классе «Не мошенничество», а очень небольшое меньшинство будет в классе «Мошенничество».

Количество данных по классам в обучающей выборке можно увидеть на рисунке 14, а распределение в процентном соотношении на рисунке 15. На рисунке 15 можно увидеть, что ЧС сообщений около 43%, а не ЧС около 57%, что находится в пределах нормы и один класс вряд ли затмит другой.

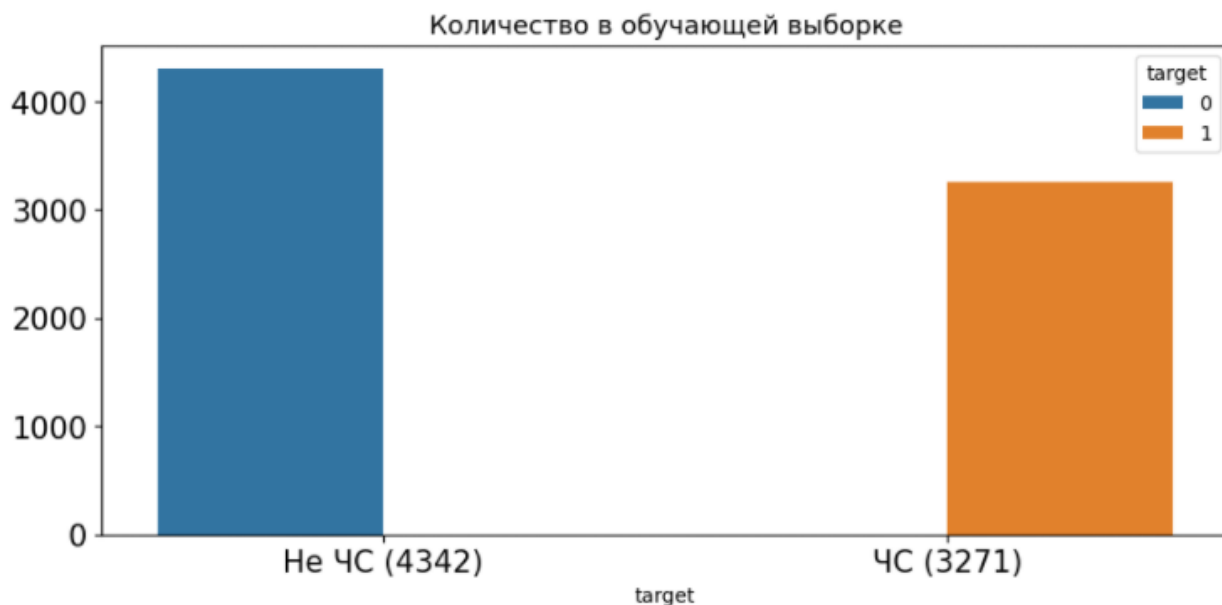


Рисунок 14 – Распределение количество сообщений по классам

Распределение в обучающей выборке

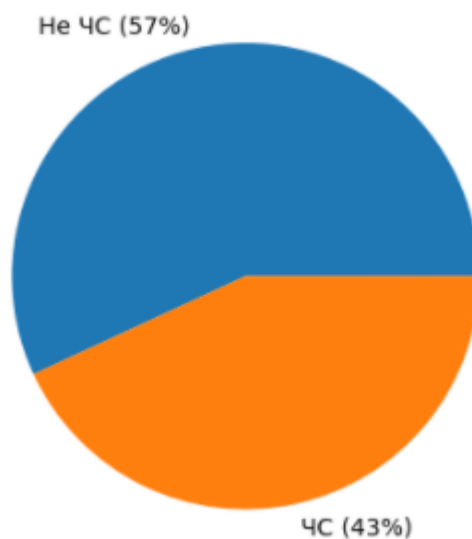


Рисунок 15 - Процентное соотношение классов в обучающей выборке

После анализа данных можно приступать к их обработке и первое, что необходимо сделать, это разбить предложения на слова (токенизация) и привести их к начальной форме (лемматизация).

4.3.1 Токенизация и лемматизация

На данном этапе необходимо пройти по всем записям, разбить их на токены и привести к начальной форме, после чего, сохранить полученный результат в отдельный столбец, чтобы воспользоваться этими данными на следующих этапах. Также, стоит сразу привести все слова к одному регистру, так как для компьютера одинаковые слова в разном регистре отличаются.

После данного этапа данные выглядят следующим образом:

		tokens	target
164	[эксперт, во, франция, приступить, к, изучение, обломок, самолёт, ,, обнаружить, на, остров, реюньон, ., в, среда, французский, эксперт, по, авиационный, происшествие, приступить, к, изучение, обломок, самолёт, .]		1
577	[jacksonville, busines, fedex, прекращать, поставка, потенциальный, патоген, биотеррора, http://t.co/shzsymausi]		1
618	[создание, ии, \n , изменение, климат, \n , биотерроризм, \n , массовый, автоматизация, персонал, \n , контакт, с, другой, жизнь, \n , неравенство, богатство, \n\n , да, .. у, нас, это, лёгкий]		0
622	[доброволец, должный, были, принять, участие, в, учение, по, обеспечение, готовность, к, чрезвычайный, ситуациям, ,, имитировать, биотеррористическую, катастрофа, http://t.co/nwv2rvghf3, @hvnewsnetwork]		1
710	[подпишитесь, на, @edwelchmusic, и, посмотреть, его, хит, -, сингл, unpacked, man, its, blazing]		0

Рисунок 16 – Предложения после обработки

Следующим этапом станет очистка данных, которая позволит избавиться от слов или символов, которые не несут смысла для обучения и даже могут уменьшить точность классификации.

4.3.2 Очистка данных

В анализируемых сообщениях необходимо убрать следующие данные:

- 1) Ссылки;
- 2) HTML-разметка;

- 3) Обращения к другим пользователям (обычно помечается каким-то отличительным символом, например @Имя_пользователя);
- 4) Эмодиконы - пиктограмма, изображающая эмоцию; чаще всего составляется из типографских знаков;
- 5) Английские слова;
- 6) Пунктуацию;
- 7) Стоп-слова или шумовые слова. К ним можно отнести предлоги, суффиксы, причастия, междометия, цифры, частицы и т. п. Общие шумовые слова всегда исключаются из поискового запроса (за исключением поиска по строгому соответствию поисковой фразы). Считается, что каждое из общих стоп-слов есть почти во всех документах коллекции.

Если не избавиться от стоп-слов, то они могут вносить информационный шум, увеличить размерность смысловых векторов на этапе приведения слов в числовую форму и в целом, они не несут смысловой нагрузки. При этом надо понимать, что не существует универсального списка стоп-слов, все зависит от конкретного случая. Для этого можно построить облако слов и построить график, на котором будут видны самые популярные слова. Облако слов – это визуальное представление списка категорий. Ключевые слова чаще всего представляют собой отдельные слова, и важность каждого ключевого слова обозначается размером шрифта или цветом. Облако слов для набора данных до очистки от стоп-слов приведено на рисунке 17.



Рисунок 17 – Облако слов до очистки данных

На облаке слов можно увидеть, как сильно выделяются предлоги, местоимения и артикли.

На рисунках 18 и 19 представлены 20 самых популярных слов и более половины из них являются стоп-словами.

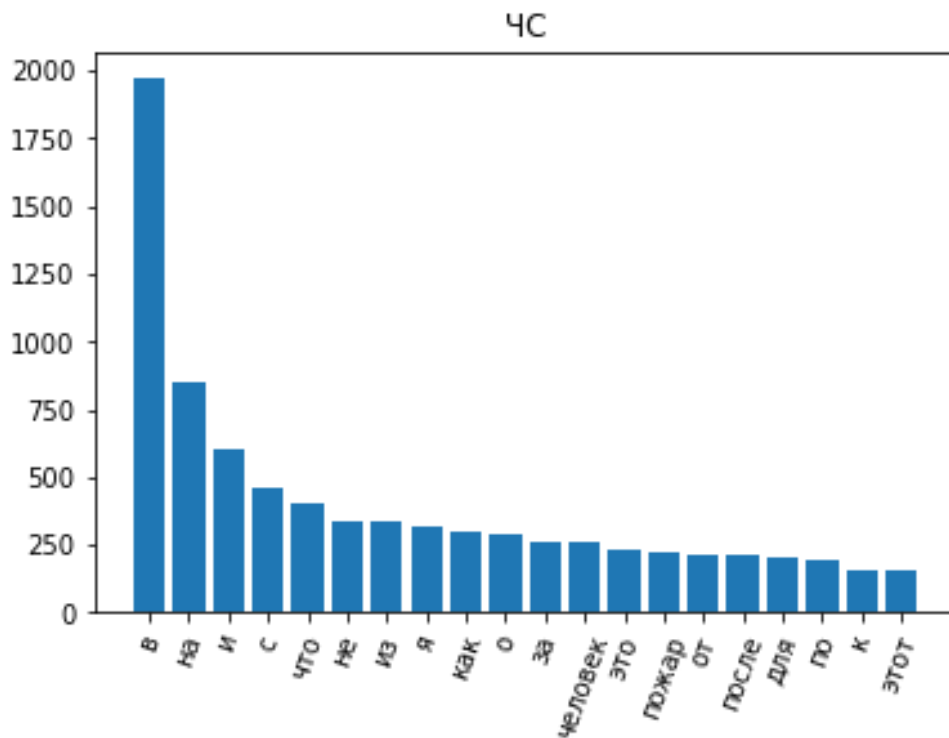


Рисунок 18 – Самые популярные слова с меткой «ЧС» до очистки

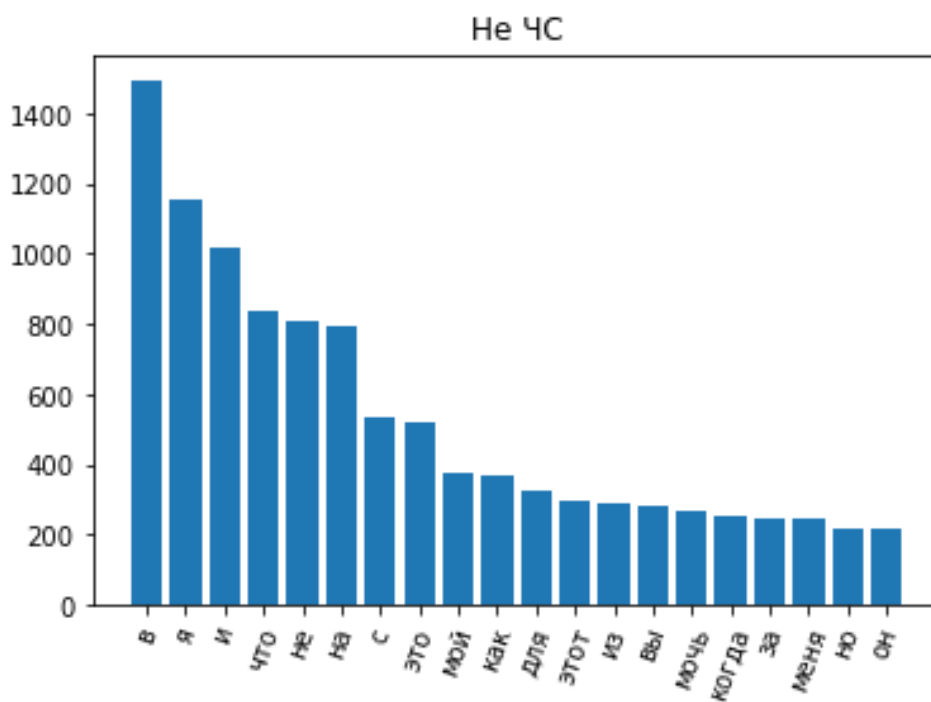


Рисунок 19 – Самые популярные слова с меткой «Не ЧС» до очистки

Для очистки всех ненужных данных использовались регулярные выражения (модуль re в python). На рисунке 20 представлено облако слов после очистки от стоп-слов.

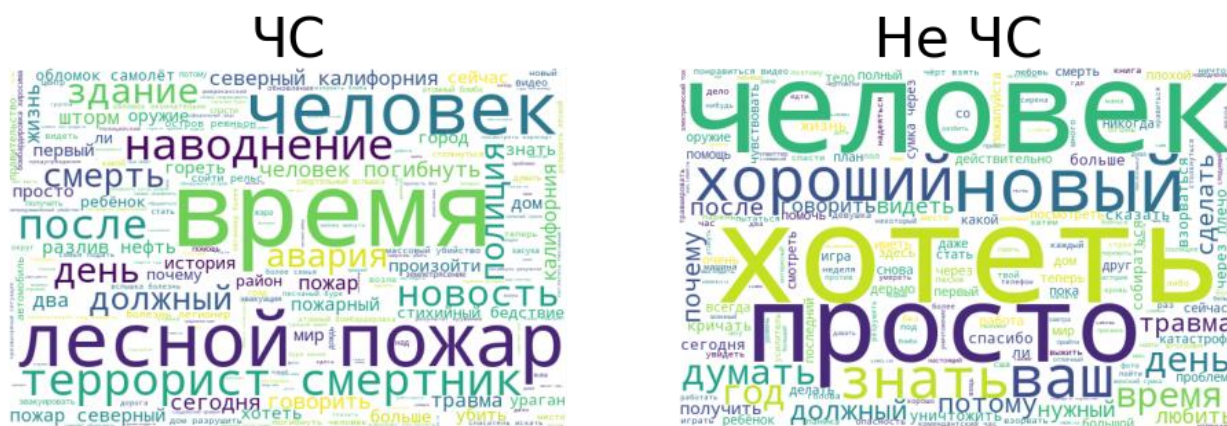


Рисунок 20 - Облако слов после очистки данных

На рисунке 20 уже видны более интересные слова, которые в сообщениях с меткой «ЧС» встречаются чаще, чем с меткой «Не ЧС». Это позволит обучить модель более эффективно. На рисунках 21 и 22 также представлены 20 самых популярных слов, но уже после очистки.

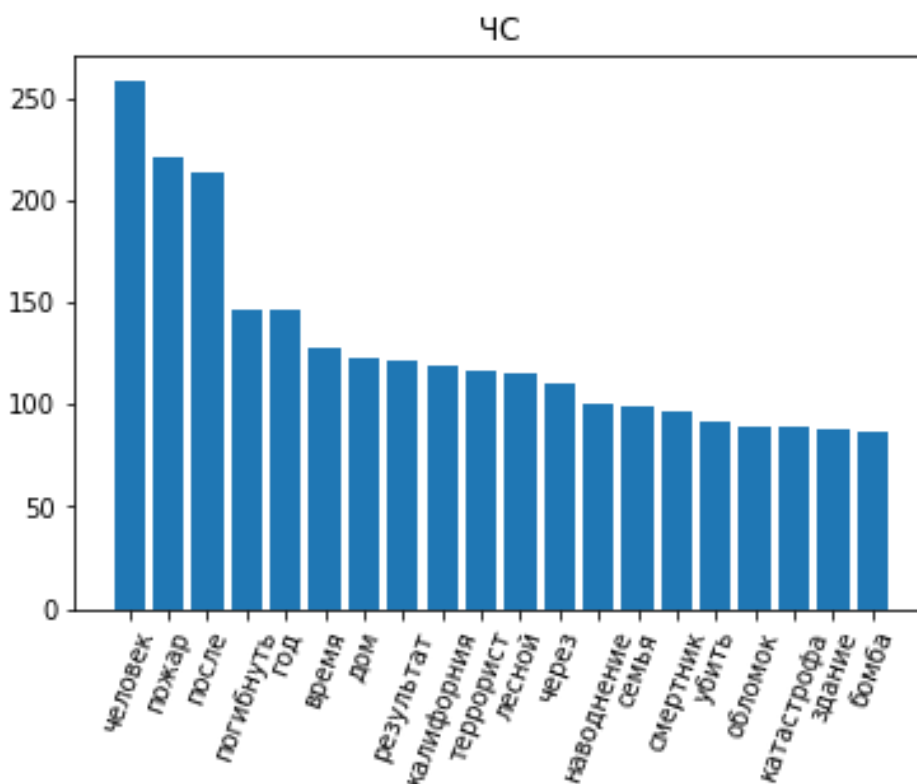


Рисунок 21 – Самые популярные слова с меткой «ЧС» после очистки

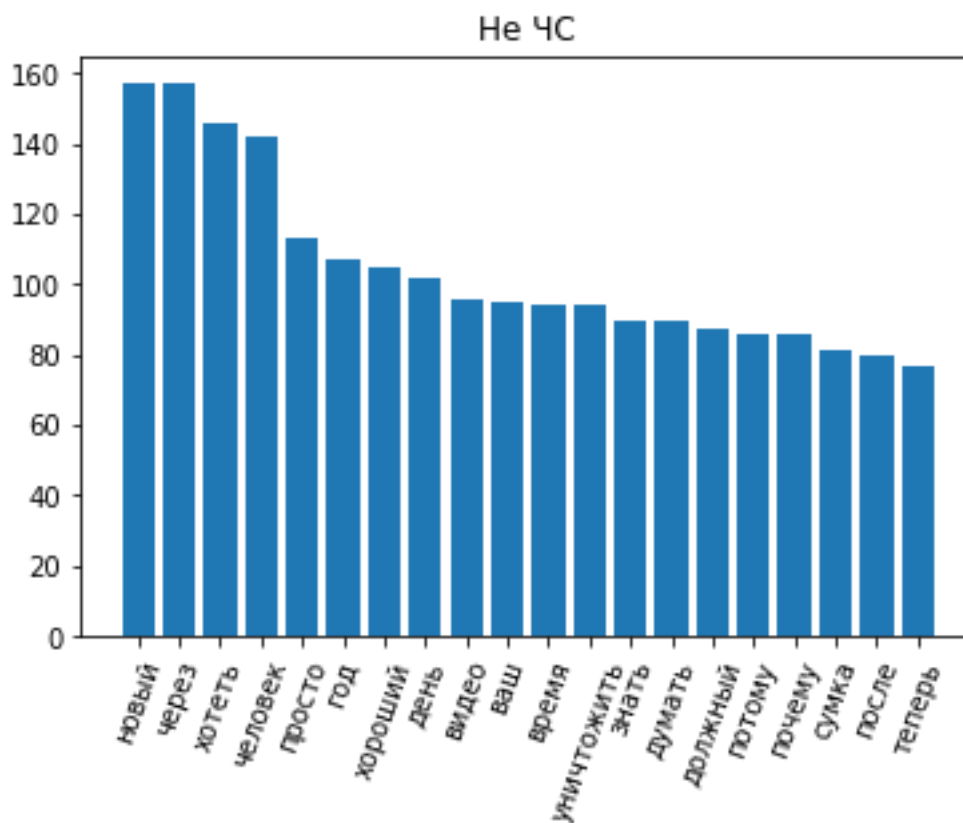


Рисунок 22 – Самые популярные слова с меткой «Не ЧС» после очистки

По рисункам видно, что очистка текста прошла успешно и данный текст пригоден для дальнейшей векторизации и обучение классификатора.

Результат очистки набора данных представлен на рисунке 23.

	cleared_tokens	target
164	[эксперт, франция, приступить, изучение, обломок, самолёт, обнаружить, остров, реюньон, среда, французский, эксперт, авиационный, происшествие, приступить, изучение, обломок, самолёт]	1
577	[прекращать, поставка, потенциальный, патоген, биотеррора]	1
618	[создание, ии, изменение, климат, биотерроризм, массовый, автоматизация, персонал, контакт, другой, жизнь, неравенство, богатство, лёгкий]	0
622	[доброволец, должный, принять, участие, учение, обеспечение, готовность, чрезвычайный, ситуациям, имитировать, биотеррористическую, катастрофа]	1
710	[подпишитесь, посмотреть, хит, сингл]	0

Рисунок 23 – Очищенные данные

4.3.3 Векторизация текста

После предыдущих этапов, текст превратился в очищенную нормализованную последовательность слов и данные слова можно преобразовать в числовые вектора. Большинство моделей машинного обучения работают с числами, а математические модели работают в векторных пространствах больших размерностей, поэтому необходимо отобразить текст в векторном пространстве. Основным подходом является мешок слов (bag-of-words): для документа формируется вектор размерности словаря, для каждого слова выделяется своя размерность, для документа записывается признак насколько часто слово встречается в нем, получаем вектор. Наиболее распространенным методом для вычисления признака является TF-IDF (TF — частота слова, term frequency, IDF — обратная частота документа, inverse document frequency). TF вычисляется, например, счетчиком вхождения слова. IDF обычно вычисляют как логарифм от числа документов в корпусе, разделённый на количество документов, где это слово представлено. Таким образом, если какое-то слово встретилось во всех документах корпуса, то такое слово не будет никуда добавлено. Плюсами мешка слов является простая реализация, однако данный метод теряет часть информации, например, порядок слов. В логике BOW два предложения могут называться одинаковыми, если содержат один и тот же набор слов. Именно метод TF-IDF из библиотеки scikit-learn и будет использоваться в данной работе.

4.3.4 Обучение классификатора

Так как с помощью такого небольшого набора данных эффективно обучить нейронную сеть не выйдет, были выбраны две классические модели машинного обучения: логистическая регрессия и наивный байесовский классификатор.

Логистическая регрессия – это статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события путём его

сравнения с логистической кривой. Она выдаёт ответ в виде вероятности бинарного события (1 или 0).

Наивный байесовский классификатор – простой вероятностный классификатор, основанный на применении теоремы Байеса со строгими предположениями о независимости. Наивный байесовский классификатор предполагает, что наличие функции в классе не связано с какой-либо другой функцией.

Данные алгоритмы не требуют огромных объемов данных и способны довольно быстро обучаться.

После обучения моделей с помощью вышеназванных алгоритмов, необходимо будет проверить точность построенных моделей и сравнить их между собой. Для этого необходимо ввести несколько понятий.

Валидационные данные – это часть набора данных основа для проверки работоспособности модели машинного обучения. На валидационных данных и будет происходить тестирование точности модели.

Мера Accuracy (точность) – доля правильных ответов алгоритма. Не является самой информативной, особенно при явном дисбалансе классов, так как, если бы количество сообщений одного класса сильно превышало количество сообщений другого класса, то системе достаточно всегда относить сообщения к первому классу и при этом, Accuracy будет очень высокой.

Precision можно интерпретировать как долю объектов, названных классификатором положительными и при этом действительно являющимися положительными, а recall показывает, какую долю объектов положительного класса из всех объектов положительного класса нашел алгоритм.

Precision и recall не зависят, в отличие от accuracy, от соотношения классов и потому применимы в условиях несбалансированных выборок.

Часто в реальной практике стоит задача найти оптимальный (для заказчика) баланс между этими двумя метриками. Так как для модели важно

находить сообщения, которые относятся к теме ЧС, то нужно максимизировать метрику Recall.

Чем выше точность и полнота, тем лучше. Но в реальной жизни максимальная точность и полнота не достижимы одновременно и приходится искать некий баланс. Поэтому, хотелось бы иметь некую метрику, которая объединяла бы в себе информацию о точности и полноте выбранного алгоритма. Именно такой метрикой является F1-score (F-мера).

F1-score представляет собой гармоническое среднее между точностью и полнотой. Она стремится к нулю, если точность или полнота стремятся к нулю.

$$F = 2 * \frac{precision * recall}{precision + recall} \quad (1)$$

Данная формула придает одинаковый вес точности и полноте, поэтому F-мера будет падать одинаково при уменьшении и точности и полноты.

Были выбраны алгоритмы и метрики, по которым будет оцениваться точность классификации. Результат обучения представлен в таблице 9.

Наивный байесовский классификатор смог обнаружить более 2/3 сообщений, которые относятся к ЧС (recall), при этом, с 82% вероятностью алгоритм отвечал верно (precision).

Логистическая регрессия показала похожий результат. Она смогла определить меньшее количество сообщений, которые относятся к классу «ЧС» (recall), но реже ошибалась при отнесении к этому классу (precision).

Результат проверки в виде графиков можно увидеть на рисунках 24 и 25, которые представляют собой матрицу ошибок для наивного байесовского классификатора и линейной регрессии соответственно.

Именно F1-score и Recall использовалась при выборе классификатора. Поскольку Recall выше именно у байесовского классификатора, как и F1-score

для сообщений класса «ЧС» то был выбран именно этот алгоритм для дальнейшей работы.

Таблица 8 – Результат обучения моделей

Алгоритм	Класс	precision	recall	F1-score	Accuracy
Наивный байесовский классификатор	Не ЧС	0.8	0.83	0.82	0.80
	ЧС	0.76	0.71	0.73	
Логистическая регрессия	Не ЧС	0.77	0.91	0.84	0.79
	ЧС	0.84	0.64	0.73	

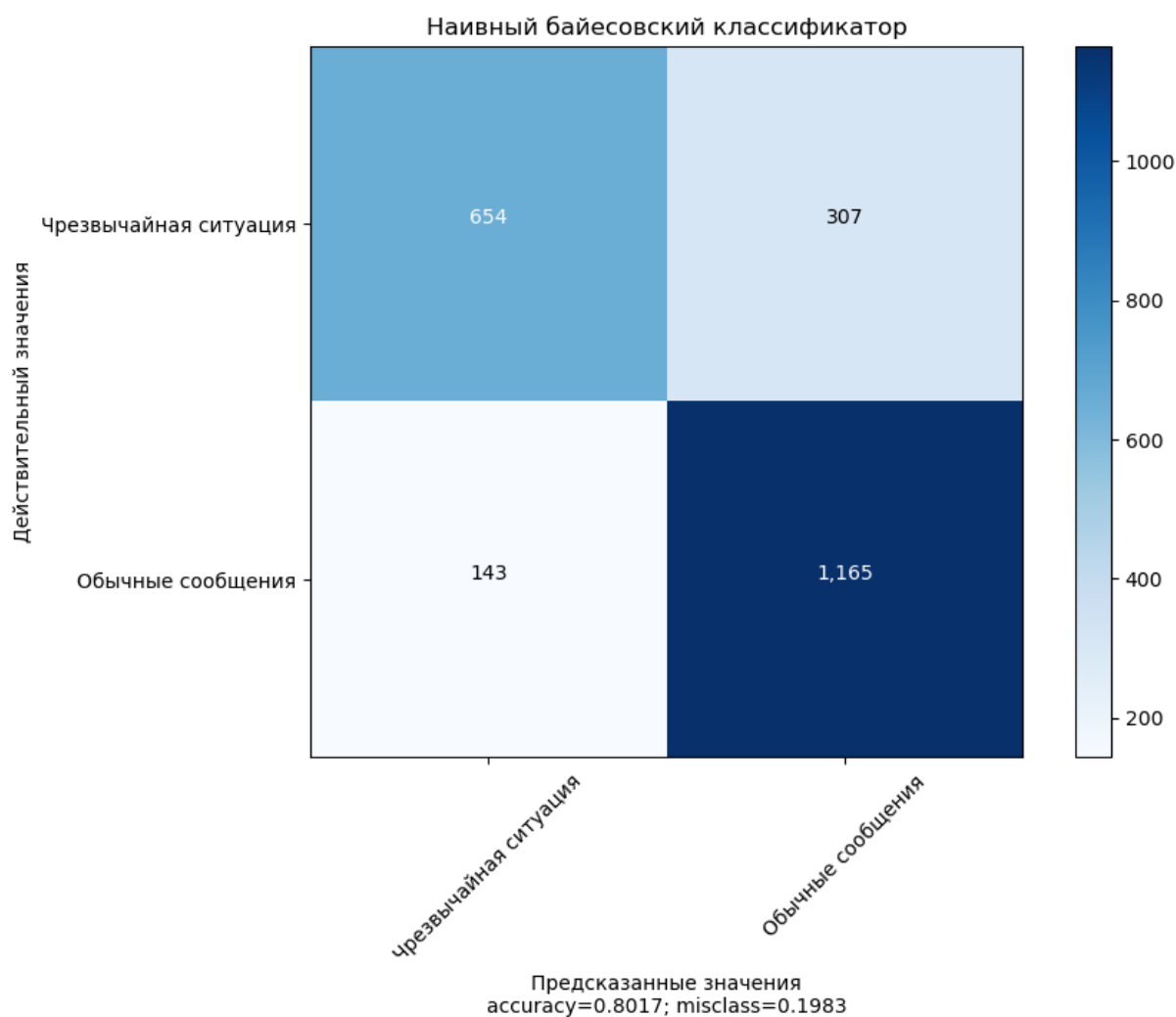


Рисунок 24 – Матрица ошибок наивного байесовского классификатора

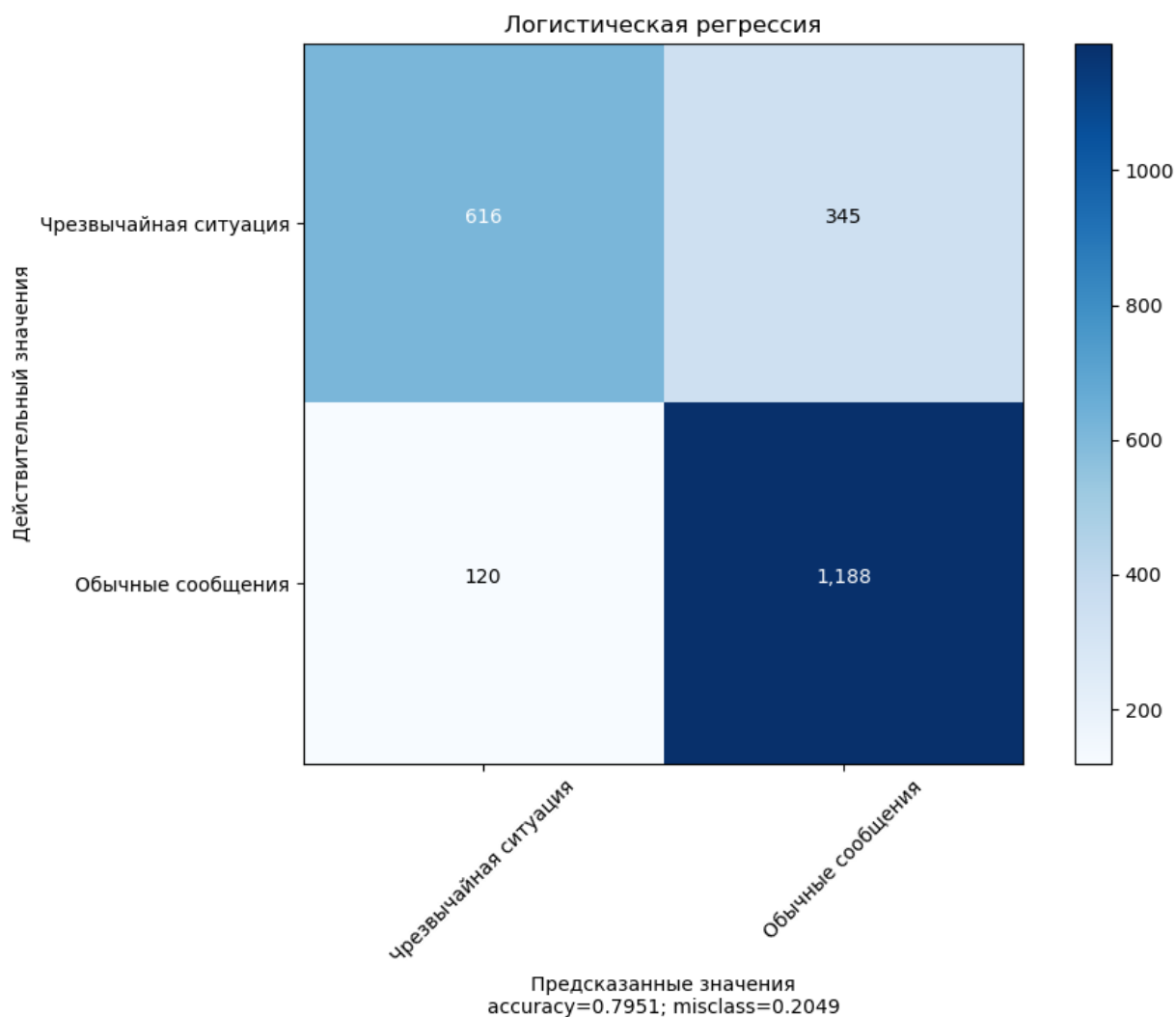


Рисунок 25 – матрица ошибок логистической регрессии

4.4 Разработка экранных форм приложения

При входе в систему пользователя встречает окно авторизации (рисунок 26). Вход в систему осуществляется по логину и паролю, который выдаётся системным администратором каждому сотруднику. Данные для авторизации передаются на сервер, где сервер выдаёт JWT токен, позволяющий пользователю совершать операции в системе.

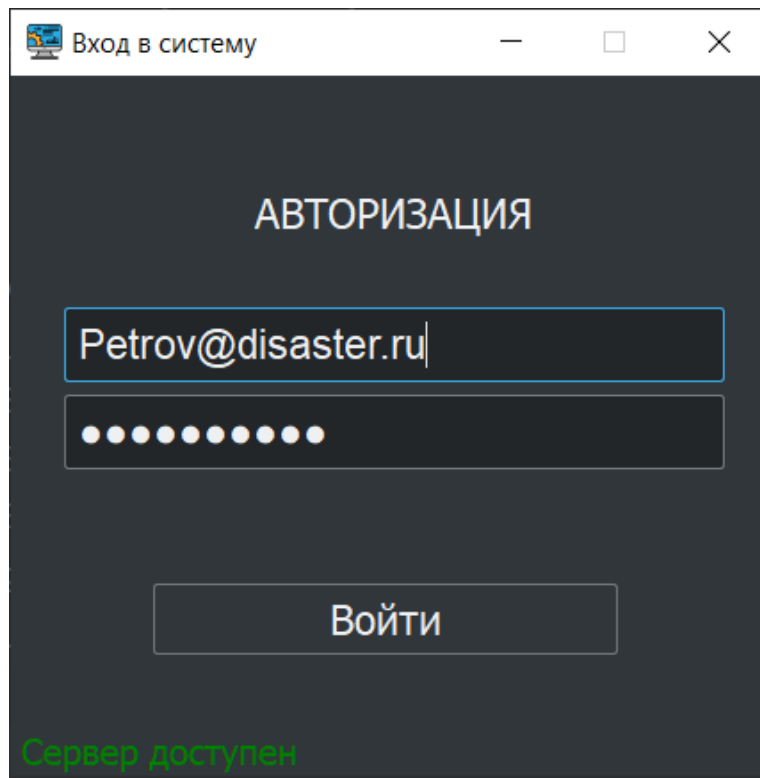


Рисунок 26 – Окно авторизации

После выдачи прав, пользователь получает доступ к основным функциям приложения. Главное окно представляет из себя форму с 3-мя вкладками: активность, графики и настройки.

На вкладке «Активность» (рисунок 27) также есть виджет с 3-мя вкладками, который предназначен для разделения ЧС по вероятности. В пределах данного окна, эксперт видит информацию по всем сообщениям, видит ответ системы о принадлежности конкретного сообщения к ЧС и может управлять этими сообщениями. Если нажать на сообщение 2 раза, то оно откроется, и пользователь увидит перед собой окно для анализа конкретного сообщения (рисунок 28), просмотра контекста сообщения и эксперт может в пределах данного окна подтвердить или отклонить факт ЧС.

Вы вошли как Petrov

Активность | Графики | Настройки

Выводить только непроверенные сообщения
 Отменить сортировку по завершенности
 Сортировать по

Высокая вероятность ЧС
 Средняя вероятность ЧС
 Низкая вероятность ЧС

	ID	Сообщение	Источник	Дата	Отправитель	Вероятность	ЧС	Эксперт
1	180	Подросток погиб при нападении медведя на туристов в ...	Рабочая группа	Геннадий Плахонин(NejdanX)	2021-06-22 19:53:22	0.87	Да	Нет данных
2	162	Введенный после разлива нефтепродуктов режим ЧС ...	Рабочая группа	Геннадий Плахонин(NejdanX)	2021-06-22 19:27:04	0.983	Да	Нет данных
3	134	В Удмуртии женщину заподозрили в убийстве сына	Ученическая группа	Геннадий Плахонин(NejdanX)	2021-06-22 17:06:13	0.56	Да	Нет данных
4	133	Названа предположительная причина убийства вожатой в ...	Ученическая группа	Геннадий Плахонин(NejdanX)	2021-06-22 17:02:31	0.949	Да	Нет данных
5	132	Здание было всё в огне, страшное зрелище.	Ученическая группа	Геннадий Плахонин(NejdanX)	2021-06-22 08:29:48	0.664	Да	Нет данных
6	131	Вся улица была в крови после жестокой бойни.	Ученическая группа	Геннадий Плахонин(NejdanX)	2021-06-21 22:48:34	0.625	Да	Нет данных
7	129	Я попал в зону лесного пожара.	Ученическая группа	Геннадий Плахонин(NejdanX)	2021-06-21 22:47:10	0.987	Да	Нет данных
8	128	В аэропорту террорист захватил самолет и угрожает его взорват...	Ученическая группа	Геннадий Плахонин(NejdanX)	2021-06-21 22:44:18	1.0	Да	Нет данных
9	126	В здании нашли бомбу. Произошла экстренная эвакуац...	Ученическая группа	Геннадий Плахонин(NejdanX)	2021-06-21 22:35:32	0.97	Да	Нет данных
10	125	Кабинет 307 был разрушен после масштабного землетрясения.	Ученическая группа	Геннадий Плахонин(NejdanX)	2021-06-21 22:33:00	0.953	Да	Нет данных
11	123	Вся крыша кардиологии сгорела после пожара на улице Горького	Ученическая группа	Геннадий Плахонин(NejdanX)	2021-06-21 22:27:50	0.986	Да	Нет данных
12	121	Террорист захватил самолёт	Ученическая группа	Геннадий Плахонин(NejdanX)	2021-06-21 22:10:29	0.976	Да	Нет данных

Рисунок 27 – Вкладка «Активность»

Сообщение

На основании приведенных ниже данных, подтвердите или отклоните факт ЧС

ID сообщения: 128
 Обнаружена ЧС с вероятностью 1.0
 Дата: 2021-06-21 22:44:18
 Источник: Ученическая группа
 Тип источника: Группа
 Отправитель: Геннадий Плахонин(NejdanX)
 Статус: Открыт
 Дата закрытия: Не определена

В аэропорту террорист захватил самолет и угрожает его взорвать, если правительство не удовлетворит его требованиям

[Контекст](#)

От Геннадий Плахонин(NejdanX) 2021-06-21 22:35:32
В здании нашли бомбу. Произошла экстренная эвакуация учеников

От Геннадий Плахонин(NejdanX) 2021-06-21 22:33:00
Кабинет 307 был разрушен после масштабного землетрясения.

От Геннадий Плахонин(NejdanX) 2021-06-21 22:27:50
Вся крыша кардиологии сгорела после пожара на улице Горького

Рисунок 28 – Просмотр информации о сообщении

Также, в пределах вкладки «Активность», возможно сформировать отчёт о ЧС за конкретный период. Пример отчёта приведен в приложении Г.

На вкладке «Графики» эксперт может визуализировать данные о ЧС. Доступна визуализация матрицы ошибок (рисунок 29), количество ЧС по часам в конкретный день (рисунок 30), а также по дням, месяцам и годам (рисунок 31). Можно выбрать конкретную дату или диапазон дат для визуализации.

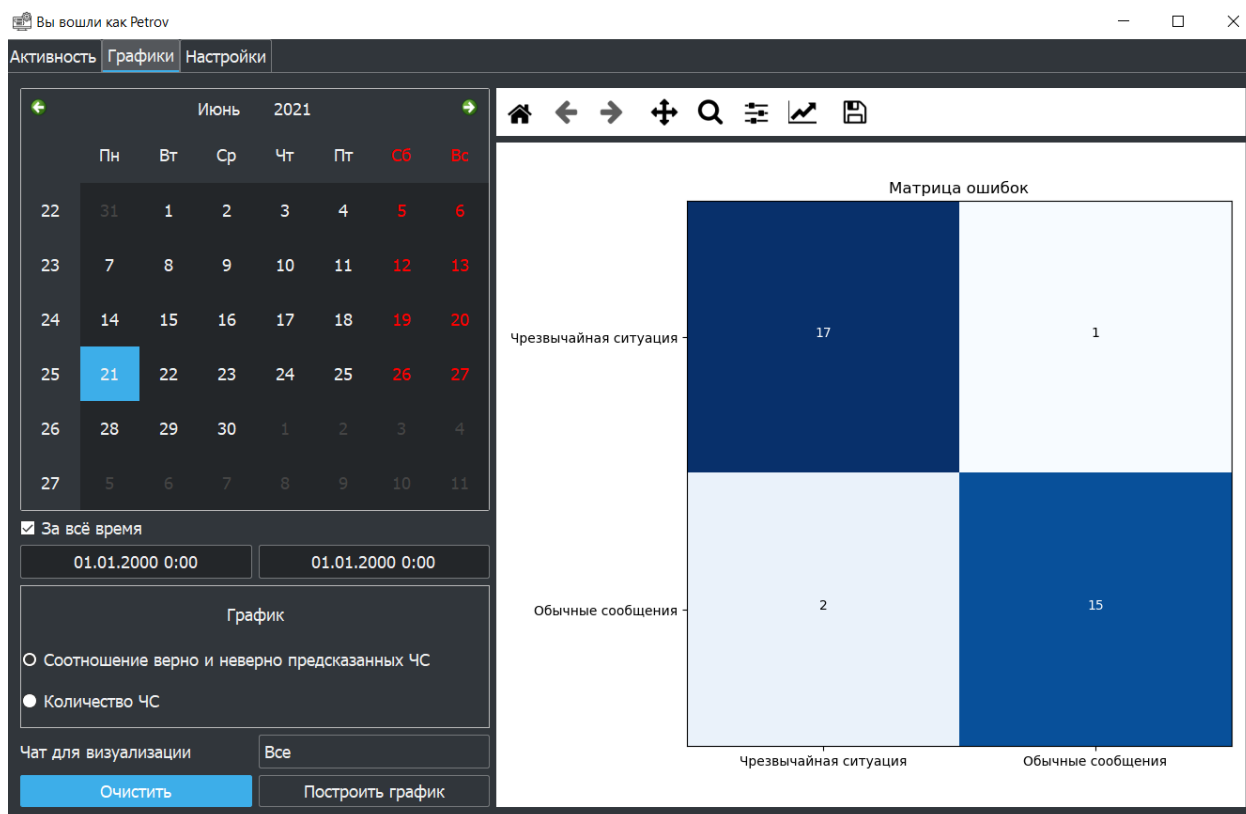


Рисунок 29 – Вкладка «Графики»

Данная функциональность представлена в пределах роли «Эксперт», который анализирует сообщения, приходящие в систему. При возникновении различных ошибок, связанных с работой сервера или ошибочными действиями пользователя, последний получает информационные сообщения об ошибках.

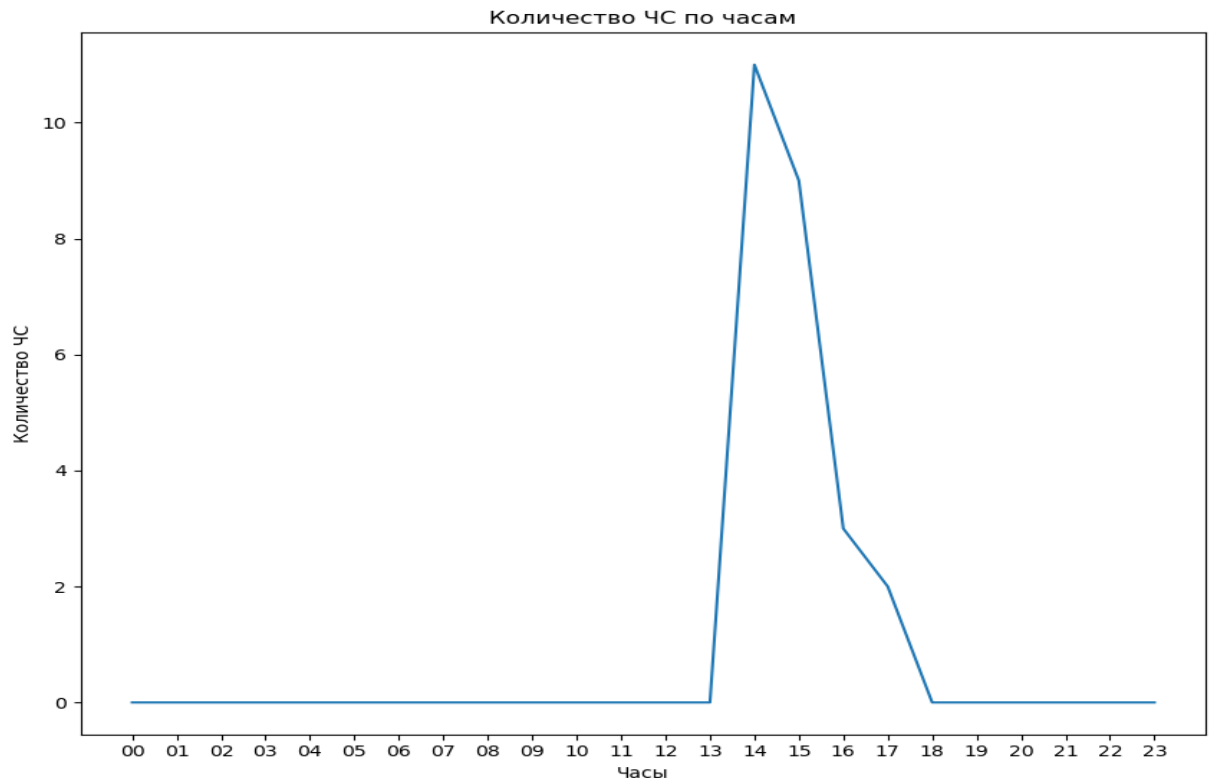
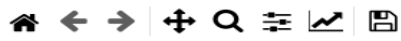


Рисунок 30 – Вид графика «Количество ЧС по часам»

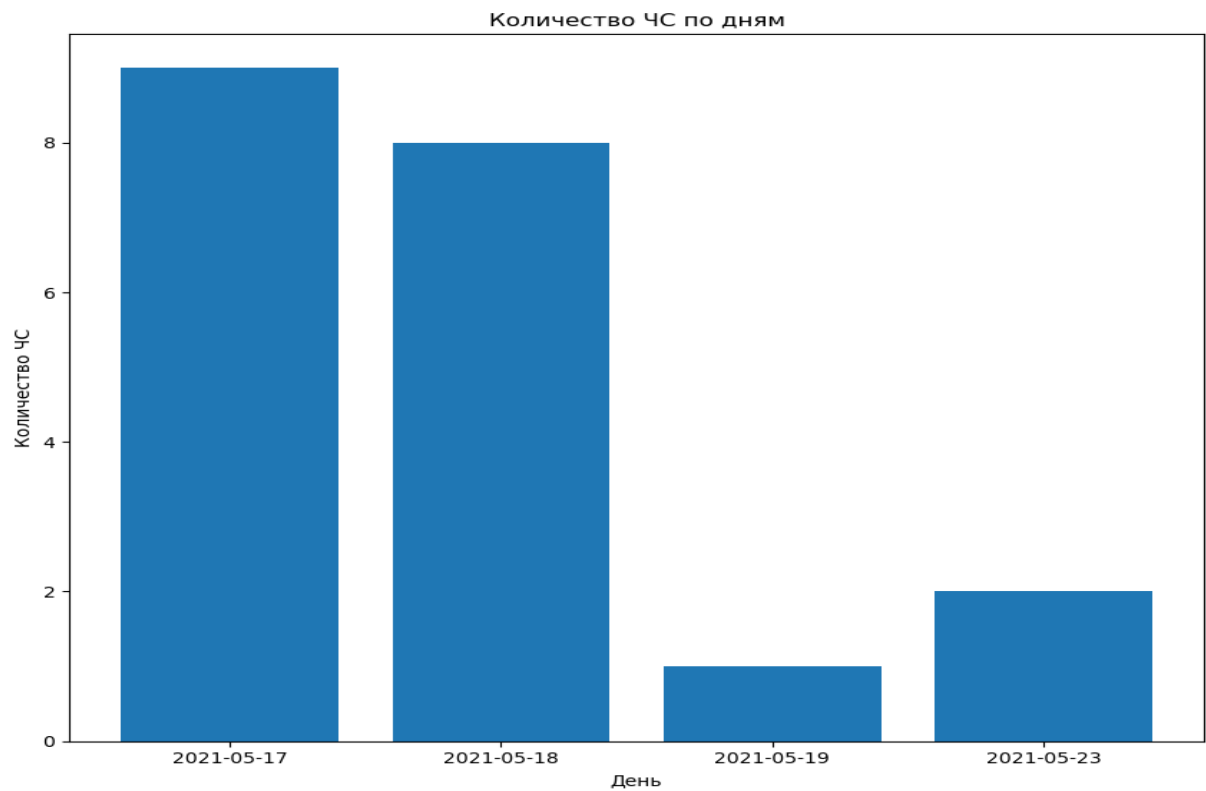
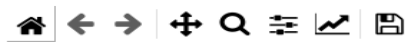


Рисунок 31 – Вид графика «Количество ЧС по дням»

Для администратора ИС также разработан пользовательский интерфейс, который позволяет управлять пользователями системы, а также менять настройки сервера.

На рисунке 32 представлено окно «Настройки сервера», которое позволяет настроить строгость классификации сообщений и провести тестирование выбранной модели, настроить логирование и резервное копирование, автоматическую очистку БД от сообщений, которые не относятся к теме ЧС, а также, выбрать чаты для анализа.

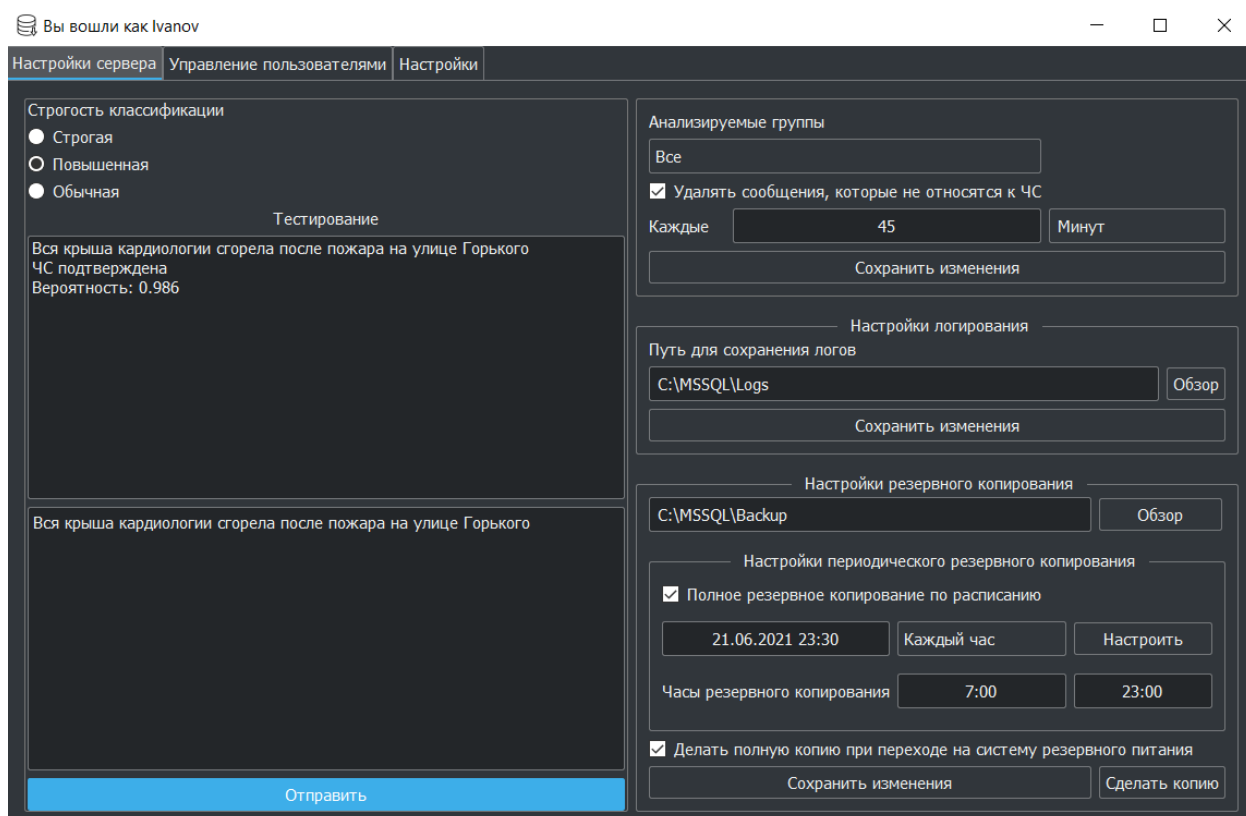


Рисунок 32 – Вкладка «Настройки сервера»

На вкладке «Управление пользователями» можно управлять текущими подключениями к серверу, добавлять новых пользователей или изменять данные уже существующих. Данное окно представлено на рисунке 33. Также, в данном окне возможно заблокировать конкретного пользователя или IP-адрес на время или перманентно, если нажать на кнопку «Заблокировать» (рисунок 34).

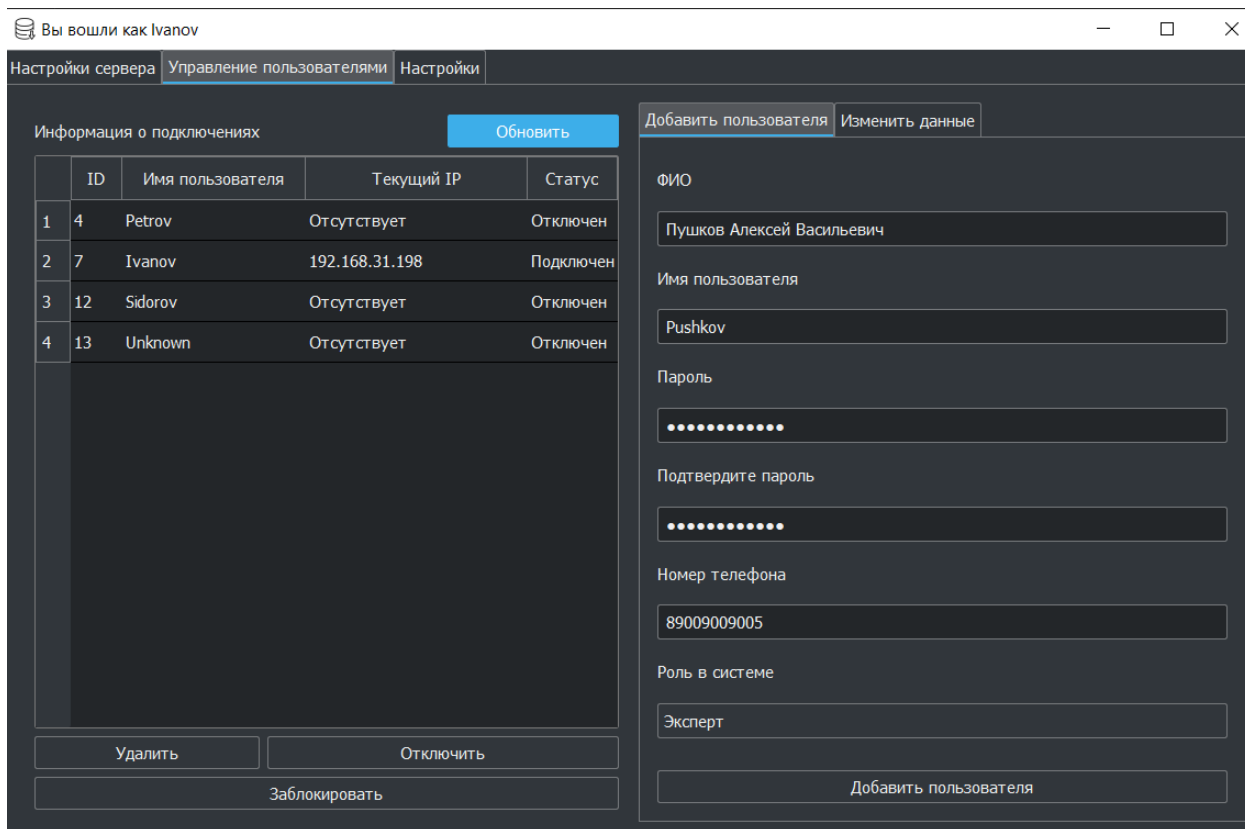


Рисунок 33 – Вкладка «Управление пользователями»

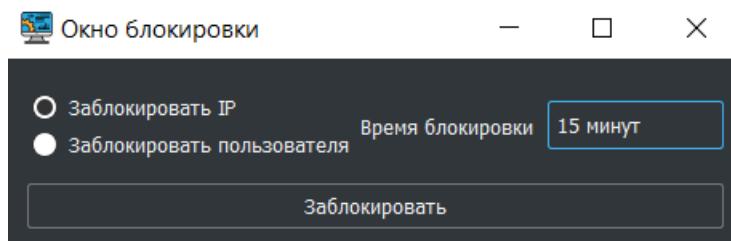


Рисунок 34 – Окно блокировки

При возникновении различных ошибок, связанных с работой сервера или ошибочными действиями администратора, последний получает информационные сообщения об ошибках.

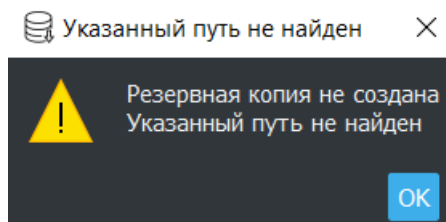


Рисунок 35 – пример ошибки при резервном копировании

5 УГРОЗЫ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ ПРИЛОЖЕНИЯ

5.1 Предложения по обеспечению ИБ системы

В главе 3 пункте 4.1 подробно описывается процесс авторизации пользователя в системе. Это позволит предотвратить преднамеренные или непреднамеренные попытки доступа к основным функциям приложения. Также предлагаются следующие предложения по обеспечению ИБ:

- Разграничение прав доступа пользователей и администраторов ИС должно строиться по принципу "что не разрешено, то запрещено".
- Запись последнего время входа и выхода пользователя в ИС должна помещаться в БД;
- Права на редактирование внутренних настроек системы должны быть доступны только администратором системы;
- Разграничение по получаемым сообщениям должно обеспечиваться на стороне сервера для каждого активного пользователя;

5.2 Требования к надежности

5.2.1 Состав показателей надежности системы в целом.

Надежность системы обеспечивается за счёт применения организационного-технических мероприятий и программно-аппаратных средств.

Надежность должна обеспечиваться за счёт:

- защиты от некорректных действий пользователя системы;
- своевременного выполнения процессов администрирования системы;
- соблюдения правил эксплуатации и планового технического обслуживания ИС, предварительного обучения пользователей и обслуживающего персонала;
- выполнения резервных копий БД и предварительно обученных моделей для классификации на внешние носители информации;

5.2.2 Перечень аварийных ситуаций, по которым регламентируются требования к надежности.

Под аварийной ситуацией понимается нарушение работоспособности той или иной подсистемы ИС.

В ходе функционирования системы возможны следующие аварийные ситуации, которые могут повлиять на надежность работы системы:

- сбой в электроснабжении сервера, рабочей станции пользователей системы или обеспечения локальной сети;
- ошибки в ИС, которые не были выявлены при отладке и испытании системы;
- сбои в сервисах, откуда система получает сообщения для анализа.

5.2.3 Требования к надежности технических средств и программного обеспечения

Надежность системы должна обеспечиваться благодаря использованию современных и проверенных технических средств, постоянной профилактикой технических средств.

Вопросы надежности электроснабжения необходимо решить с помощью:

- комплектация сервера источником питания до полного резервного копирования БД на внешний носитель;
- оповещения администратора и пользователей ИС о переходе на автономный режим.

К организационным мероприятиям обеспечение надежности относятся:

- ознакомление пользователей ИС с Руководством пользователя;
- своевременного выполнения процессов администрирования;
- соблюдения правил эксплуатации и технического обслуживания программно-аппаратных средств;

- своевременное выполнение процедур резервного копирования данных.

Надежность программного обеспечения подсистем должна обеспечиваться за счет:

- надежности общесистемного ПО и ПО, разрабатываемого Разработчиком;

- проведением комплекса мероприятий отладки, поиска и исключения ошибок.

- ведением журналов системных сообщений и ошибок по подсистемам для последующего анализа и изменения конфигурации.

5.3 Требования по сохранности информации при авариях

Благодаря дополнительному источнику питания сервера предполагается, что будет сделана полная резервная копия на внешний носитель. Также, после аварии на сервере необходимо средствами самой СУБД проверить БД на наличие ошибок и по возможности их исправить. Ущерб программному обеспечению мало вероятен, только если пользовательские настройки, которые записаны в файле конфигурации на локальном компьютере самого пользователя, не будут искажены в последствии. В этом случае, необходимо по возможности восстановить пользовательские настройки или сбросить их до заводских. Тоже самое касается конфигурации самого сервера. Если при запуске сервер не найдёт файл конфигурации, до просто сбросится до заводских настроек.

6 БЕЗОПАСНОСТЬ И ЭКОЛОГИЧНОСТЬ

Информационной безопасностью называется процесс обеспечения конфиденциальности, целостности и доступности информации. При отслеживании ЧС используются рабочие места для сотрудников, выполняющих анализ сообщений и настройку сервера и соответственно необходимо задействовать помещения, в которых они будут находиться. Рабочие места создаются в соответствии с нормативными документами и стандартами, в которых предусмотрены вопросы о сохранении здоровья сотрудников при работе на ЭВМ. СанПиНы и другие нормативно – правовые акты, принятые в нашей, направлены на обеспечение условия труда, отвечающих требованиям сохранения жизни и здоровья работников в процессе трудовой деятельности. Они содержат положения, обеспечивающие работникам гарантии прав на охрану труда. Обеспечение здоровых и безопасных условий труда работникам является главной задачей работодателя.

Безопасность жизнедеятельности – наука о комфортном и безопасном взаимодействии человека со средой обитания.

Цель этого раздела является изучение и решения вопросов техники безопасности и экологии труда для работников, которые занимаются эксплуатацией разрабатываемую систему.

При создании рабочего места таким работникам также необходимо разработать рекомендации и комплексы упражнений с целью охраны здоровья сотрудников. В работе будут исследоваться 3 основных положения:

- безопасность;
- экологичность;
- защита от ЧС.

6.1. Безопасность

6.1.1. Определение безопасных и вредных факторов на рабочем месте пользователя ПЭВМ.

Стандартом СанПин 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания» есть требования, которые стоит соблюдать для эффективной защиты и профилактики вопросов организации труда и отдыха.

В соответствии с ГОСТом 12.03.003-2015 «опасные и вредные производственные факторы» выделяются при работе с ПЭВМ следующие опасные факторы, наносящие вред здоровью пользователя:

- электростатические поля;
- электромагнитное излучение;
- опасность поражения электрическим током;
- повышенная или пониженная температура воздуха на рабочем месте;
- повышенный уровень шума на рабочем месте;
- монотонность рабочего процесса;
- утомляемость глаз;
- отсутствие или недостаток в помещении естественного света;
- недостаточная искусственная освещенность рабочего места;
- нервно-эмоциональный перегрузки пользователя.

К рабочему месту пользователя, эксплуатирующим информационную систему, предъявляются следующие требования:

- рабочая поверхность стола для взрослых пользователей по высоте должна регулироваться в пределах 680-800 мм, при отсутствии такой возможности высота должна составлять 725мм;
- пространство для ног рабочего стола должно составлять не менее 600мм, шириной не менее 500 мм, глубиной на уровне колен не менее 450 мм и на уровне вытянутых ног не менее 650 мм;

- поверхность сидения должна иметь 400 мм, передний край должен быть закругленным и регулироваться в пределах 400-500 мм, углы наклона вперед до 15 градусов и назад до 5 градусов;
- подлокотники могут быть стационарными или съемными, их длина не менее 250 мм и ширина 50-70 мм, они регулируются над сиденьями в пределах 230 мм, внутреннее расстояние между подлокотниками в пределах 350 – 500 мм;
- клавиатура располагается на поверхности стола на расстоянии 100-300 мм от края, обращенного к пользователю или на специальной рабочей поверхности, которая регулируется по высоте и отделена от стола;
- рабочее место оборудуется подставкой для ног, имеющей ширину не менее 300 мм, глубину не менее 400 мм и которое регулируется по высоте в пределах 150 мм и по углу наклона подставки до 20 градусов.

Также, необходимо учитывать площадь на одно рабочее место. В случае отсутствия периферийного оборудования и использовании жидкокристаллического монитора на одно рабочее место требуется 4,5 кв.м. площади. Если используются периферийные устройства площадь увеличивается до 6 кв.м.

Рекомендуемое размещение пользователя ПЭВМ представлено на рисунке 36.

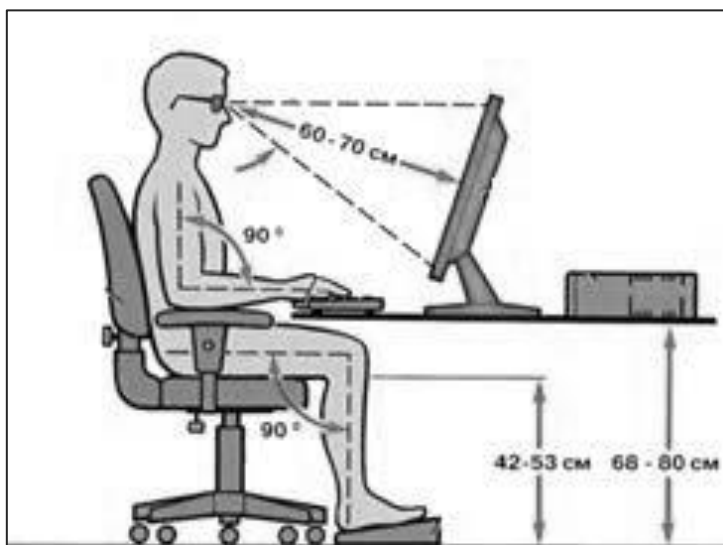


Рисунок 36 – Рекомендуемое размещение пользователя ПЭВМ

6.1.2. Освещение

Для оптимальной световой среды применяют различные системы освещения.

Под освещением понимают свет от какого – либо источника, который создает освещенность поверхности, предметов и обеспечивает зрительное восприятие этих предметов.

Выделяют следующие виды освещения:

– естественное освещение – это освещение светом, проникающим через световые проемы в наружных ограждающих конструкциях, в зависимости от расположения оно может быть боковым, верхним или комбинированным;

– искусственное освещение – это освещение помещений и других мест иными источниками, в случае недостаточного естественного освещения. Искусственное освещение подразделяется на рабочие, аварийное, охранное, дежурное, общее, местное и комбинированное. Искусственное освещение должно обеспечивать равномерное освещение всего рабочего стола.

При работе на ПЭВМ используется совмещенное освещение. Оно применяется при недостатке естественного освещения.

Аварийное освещение используется при чрезвычайных ситуациях, в случаи отключения искусственного освещения.

Согласно ГОСТ Р 55710-2013 «освещение рабочих мест внутри зданий» выделяются требования, при которых коэффициент естественной освещенности не должен быть менее 1,2% в зонах с неустойчивым снежным покровом и не ниже 1,5 % на остальной территории. На рабочем месте должно быть одностороннее естественное освещение. Если естественного освещения недостаточно необходимо применять искусственное освещение. Для этих целей применяются люминесцентные лампы, которые имеют высокую световую отдачу. Освещенность рабочей поверхности стола должна быть в пределах 300 – 500 лк.

6.1.3. Механические колебания

К механическим колебаниям относятся вибрация и шум.

Шум – это беспорядочное сочетание различных по частоте и уровню звуков. Это одна из причин быстрого утомления работающих, которая может вызвать головокружение и даже привести к несчастному случаю.

В помещениях жилых, общественных зданий и на территории жилой застройки имеет четкие ограничения. В таблице 10 представлены значения уровней звукового давления в октавных полосах частот и уровней звука на рабочих местах с ПЭВМ.

Таблица 9 - Допустимые значения уровней звукового давления в октавных полосах частот и уровня звука на рабочих местах

Уровни звукового давления, в дБ, в октавных полосах со среднегеометрическими частотами, Гц									Уровень звука в дБА
31,5	63	125	250	500	1000	2000	4000	8000	65
Гц	Гц	Гц	Гц	Гц	Гц	Гц	Гц	Гц	
96	83	74	68	63	60	57	55	54	
дБ	дБ	дБ	дБ	дБ	дБ	дБ	дБ	дБ	

Вибрация – механическое колебание упругих тел при низких частотах (3-100Гц) с большими амплитудами (0,5 – 0,003 мм).

В случаи если на рабочем месте на работника длительный период воздействует такой фактор как вибрация может наступить потеря трудоспособности, вызванная вибрационной болезнью (невритом). Эта проявляется в виде головных болей, белях в суставах, судорог пальцев, спазмов сосудов и нарушении питания тканей тела. В тяжелых случаях это приводит к полной потери трудоспособности и инвалидности.

6.1.4. Электромагнитное и ионизирующие излучение

При включенном ПЭВМ на рабочем месте создается электромагнитное и ионизирующие излучение. При работе монитор ПЭВМ излучает мягкое рентгеновское излучение. В течении работы на корпусе накапливается статическое электричество. При включенном ПЭВМ запрещается:

- прикасаться к монитору;
- вытирать пыль;
- производить любые манипуляции с отключением отдельных частей ПЭВМ.

Экран монитора должен находиться на расстоянии 60-70 см. от глаз пользователя.

Допустимые уровни электромагнитного излучения можно найти в СанПиН 1.2.3685-21.

6.1.5. Микроклимат

Микроклимат производственных помещений – это характеристика внутренней среды помещений, в которых находятся рабочие места. Микроклимат определяется совместно действующими на организм человека температурой, относительной влажностью и скоростью движения воздуха окружающих поверхностей. Эти показатели определены в нормативно – правовом документе ГОСТ 12.1.005 «Общие санитарно-гигиенические требования к воздуху рабочей зоны».

Задача работодателя поддерживать микроклимат рабочего места в пределах гигиенических норм, с целью охраны труда работников.

В помещениях, оборудованных ПЭВМ, должна проводиться ежедневная влажная уборка и систематическое проветривание после каждого часа работы на ПЭВМ.

Оптимальная температура воздуха в помещениях согласно ГОСТу 12.1.005-76 «Воздух рабочей зоны. Общие санитарно-гигиенические требования» в холодные и переходные периоды должна находится в пределах

20 – 30⁰С, в теплый период года 20-25⁰С. Относительная влажность воздуха 60-40%, скорость движения воздуха не более 0,2м/с во все периоды года.

При работе на ПЭВМ на рабочем месте необходимо поддерживать оптимальную температуру воздуха. ПЭВМ на рабочих местах являются источником существенных выделений, которые повышают температуру тела работника, что приводит к снижению работоспособности. ПЭВМ повышает температуру всего помещения в целом. Исходя из вышеизложенного, поддержание температуры на требуемом уровне позволяет обеспечить безопасность и комфортность при работе на ПЭВМ.

В целях поддержания микроклимата в помещении используются системы вентиляции. Система вентиляции – это смена воздуха в помещении, предназначенная для поддержания оптимальных метеорологических параметров помещения и подача чистого воздуха с наружи. В рабочих помещениях используют систему естественной вентиляции, в случаи её недостаточности применяют систему кондиционирования для регулирования микроклиматических параметров в рабочих помещениях для создания комфортных условий труда.

В рабочих помещениях в холодный период года постоянную и равномерную температуру поддерживает система отопления. Виды систем отопления:

- водяные;
- паровые;
- воздушные;
- комбинированные.

Наиболее эффективными в санитарно-гигиеническом отношении признаны системы водяного отопления, их достоинство заключается в том, что они надежны и обеспечивают возможность регулировать температуру в помещении. В помещениях с установленными ПЭВМ в холодный период температура не должна превышать 22-24⁰С, в теплый период – 20-25⁰С.

Относительная влажность должна составлять 40-60%, скорость движения воздуха не должна превышать 0,1м/с.

6.1.6. Анализ помещения с установленными ПЭВМ

Для автоматизированной системы выявления и оповещения ЧС используется помещение, которое имеет площадь 24м². В этом помещении находится одно рабочее место с 2 ПЭВМ, содержащих ЖК монитор, клавиатуру, мышь, планшет. Рабочая поверхность, сидения соответствуют всем требованиям нормативного документа. Оконные проемы закрыты вертикальными пластиковыми жалюзи. Помещение оснащено защитным заземлением. Температура в помещении круглый год поддерживается в диапазоне 22 – 25⁰С.

Схема размещения ПЭВМ представлена на рисунке 37:

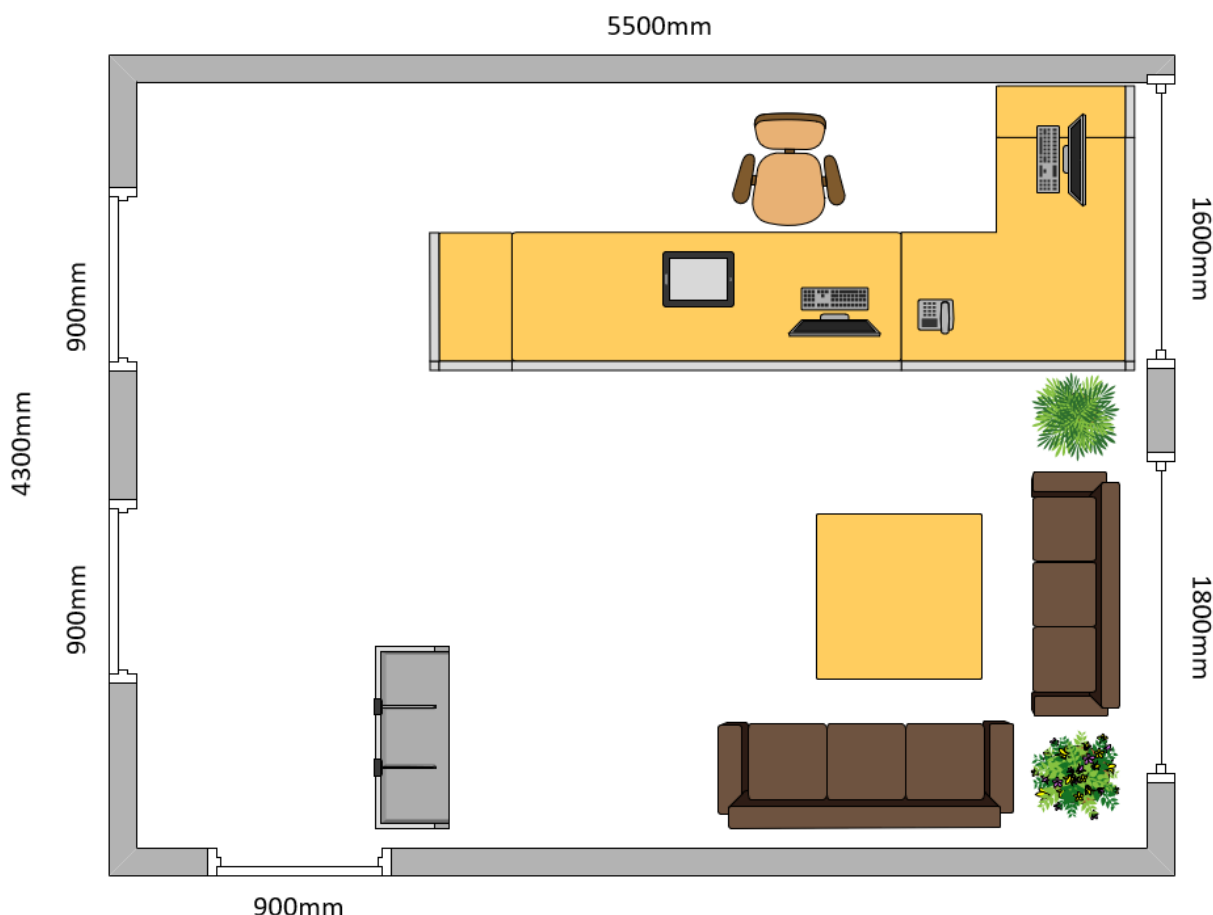


Рисунок 37 - Схема помещения с ПЭВМ

Исходя из анализа всех параметров, делаем вывод, что данное помещение соответствует всем требованиям и рекомендациям, предъявляемым в ГОСТах и СанПиНе.

6.2. Экологичность

Экологичность – качество чего-либо, отражающее его способность не наносить вред окружающей природе. Конструкция ПЭВМ состоит из многих компонентов. Эти компоненты содержат токсические вещества, которые вредны для окружающей среды.

К таким веществам относятся:

- ртуть, которая находится в подсветке ЖК-мониторов. Она может поражать мозг и нервную систему;
- щелочи, находятся в щелочных аккумуляторах. Они поражают слизистые оболочки и кожу;
- никель, цинк располагаются в материнской плате, в батареях питания для ноутбуков. Их воздействие на организм пользователя может вызвать дерматит;
- поливинилхлорид есть в составе кабелей. Воздействие разрушает нервную систему и вызывает раковые заболевания.

Принимая во внимание наличие вредных веществ в ПЭВМ для утилизации пришедших в негодность ПЭВМ используется специальный комплекс методов утилизации. Комплекс включает следующие последовательные действия:

- сортировка металлических и неметаллических частей;
- отправка металлических частей на переплавку для последующего использования в качестве вторичного сырья;
- неметаллические части утилизируются специальным способом.

Утилизируемое оборудование хранится в подсобном помещении, при хранении оно не выделяет вредных веществ. Транспортировка к месту утилизации производится без специальных средств и методов.

При утилизации отходов, содержащих свинец или ртуть, таких как ртутные лампы и аккумуляторы источников бесперебойного питания заключается договор со специальной организацией, которая имеет лицензию на действия, связанные с утилизацией этого класса отходов. Вывоз и непосредственно утилизация осуществляется этой организацией.

Таким образом можно сделать вывод, что в настоящее время создана и внедрена малоотходная технология при утилизации отходов при списании ПЭВМ. В государстве поставлена глобальная задача в перспективе решить вопрос о переходе на безотходную технологию. Однако это требует комплекса сложных технологических, конструкторских и организационных задач.

6.3. Чрезвычайные ситуации

В помещении, где находятся ПЭВМ, при выполнении работы могут произойти различные чрезвычайные ситуации: пожар, взрыв в здании, разрушение здания от сейсмической активности, затопление, получение урона вследствие электрического разряда. Во время выполнения своей основной деятельности, в «Доме научной коллаборации» также могут произойти различные ЧС, помимо перечисленных выше. К ним относятся: акты терроризма, вооруженного конфликта и других ситуаций, связанных с насилием. В следствии этого возможны следующие аварийные ситуации:

- обрыв проводов;
- неисправность заземления;
- повреждение электрооборудования;
- повреждение инженерных коммуникаций;
- различные акты насилия;

При возникновении аварийной ситуации или при резком ухудшении самочувствия, а также при любых ситуациях, когда возникает угроза жизни и здоровью людей, необходимо выполнение комплекса действий:

- остановить работы;

- обеспечить при необходимости оказания первой помощи пострадавшим;
- отключить электроэнергию;
- открыть аварийные выходы для эвакуации работников;
- доложить о создавшейся ситуации административным лицам организации;
- доложить оперативному дежурному.

Наиболее вероятная ситуация в помещении – пожар.

6.3.1 Меры пожарной безопасности на рабочих местах

В помещении, где работают ПЭВМ существует электропроводка, которая обеспечивает питание все электроприборы, а также систему освещения.

Помещения относятся к категории В – пожароопасные, согласно СП 12.13130.2009. Поэтому проблема противопожарной безопасности в них является одной из основополагающих при рассмотрении аспектов БЖД.

Эксплуатация ПЭВМ подразумевает наличие большого количества электроприборов, токопроводящих кабелей и высоких нагрузок на электросеть. Их установка, эксплуатация, техническое обслуживание, проверка, замена, утилизация должны соответствовать принятым законодательным нормам и стандартам.

Эксплуатирую ПЭВМ необходимо проверять целостность токопроводящих кабеле, вилки, розетки на отсутствие их повреждений.

В помещениях, в которых располагаются ПЭВМ, необходима установка средств пожаротушения. Это могут быть огнетушители различных конструкций.

Для более быстрого реагирования на чрезвычайную ситуацию в виде пожара необходимо оснастить помещения автоматической системой пожаротушения, дымовыми датчиками, пожарной сигнализацией, звуковыми извещателями. По всему учреждению должны быть размещены планы

эвакуации при чрезвычайной ситуации в обозначенными на них эвакуационными выходами.

При оборудовании рабочего места для работы ПЭВМ необходимо учесть следующие факторы, которые обеспечат предотвращение нанесения вреда здоровью сотрудникам при пожаре:

- обеспечение наличие проходов к путям эвакуации и эвакуационным выходам;
- установление ПЭВМ и монитора на надежную опору, не допускать его падения;
- запрещено устанавливать в нишах мебельных стенок, в тумбочках и т.п.;
- запрещено установление ближе 1 метра от нагревательных приборов, от горючих предметов (тюлей, декоративных украшений, новогодних елок и т.п.)
- запрещено установление на путях передвижения и эвакуации людей.

Для безаварийной работы перед началом эксплуатации ПЭВМ требуется провести следующие мероприятия:

- провести внешний осмотр места установки ПЭВМ и монитора, убедиться в исполнении требований безопасности, предъявляемых выше;
- провести внешний осмотр ПЭВМ, монитора, электровилки, электрошнура, убедиться в их исправности. Если есть сомнения в их исправности ПЭВМ эксплуатировать запрещается;
- осмотреть место установки на предмет горючих предметов. По возможности убрать их от места установки. При невозможности выбрать другое место установки;
- проверить вентиляционные отверстия в ПЭВМ, в случае их закрытия какими – либо предметами, открыть их или изменить условия установки;

– проверить наличие в помещении, где устанавливается ПЭВМ средства огнетушения, в случае их отсутствия обеспечить их наличие.

Исполнение всех перечисленных выше правил сократит риск возникновения чрезвычайной ситуации, а именно пожара.

6.3.2 Роль разрабатываемой системы в предупреждении ЧС

Автоматизированная система раннего обнаружения чрезвычайных ситуаций и оповещения предназначена для автоматического непрерывного наблюдения и предварительной обработки информации относительно текущего состояния потенциально опасных объектов и зданий, инженерных сооружений и сетей, расположенных на территориях с риском проявления опасных природных явлений и процессов, оперативного предоставления пользователям фактической и прогнозируемой информации, а также оповещения, в случае необходимости, населения, работников и руководителей потенциально опасных объектов, ответственных за состояние техногенной безопасности, должностных лиц органов исполнительной власти и местного самоуправления при непосредственном участии человека-оператора.

Система обучена различать разные виды ЧС, будь то: природные (землетрясение, бури), техногенные (пожары, взрывы), биологические (эпидемии), социальные (терроризм, насилие), антропогенные (ошибочные действия людей). Это возможно благодаря тому, что система обучалась на большом объеме данных и может отличать такие сообщения от тех, которые не относятся к чрезвычайным ситуациям.

Если при анализе сообщения чрезвычайная ситуация подтверждена, то об этом уведомляется оператор ИС, который даёт окончательный ответ. Оператору ИС необходимо предпринять все необходимые меры для предотвращения ЧС или полной минимизации ущерба.

Система способна оповестить только непосредственных пользователей, которые следят за сообщениями, приходящими в систему, поэтому от их своевременных действий зависит окончательный ущерб.

ЗАКЛЮЧЕНИЕ

Автоматизированная система раннего обнаружения ЧС и оповещения позволяет предотвратить самые нежелательные последствия – людские жертвы и серьёзные материальные потери. Такие системы будут актуальны всегда, и их модернизация с помощью новейших и эффективных разработок также является очень важным направлением развития.

В качестве анализируемых источников данных были выбраны чаты детей и преподавателей в мессенджере «Telegram», в котором находится бот. Он собирает сообщения и отправляет их в систему, а анализ и обработка данных осуществляется с помощью методов машинного обучения для NLP. Данная комбинация позволила автоматизировать множество рутинных задач, таких как сбор и семантический анализ данных, а также их удобное представление для оператора ИС, что максимально упрощает его трудовую деятельность.

Также, в ходе работы была создана автоматизированная система объектового уровня раннего обнаружения ЧС и оповещения для учреждения центр развития современных компетенций детей «АмурТехноЦентр» «Дом научной коллаборации имени академика РАН М.Т.Луценко». Инкрементная модель ЖЦ ПО, которая была выбрана для разработки данной ИС, позволила видеть заказчику результат создания системы на основе прототипов и вовремя корректировать требования к самой системе, а с помощью трехуровневой клиент-серверной архитектуры стало возможным максимально снизить требования к РМ операторов ИС и обеспечить централизованное управление.

Основная цель работы достигнута, и информационная система для бинарной классификации сообщений к теме ЧС методами машинного обучения находится на стадии внедрения в ЦРСКД «АмурТехноЦентр» (ДНК РАН М.Т.Луценко).

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

- 1 Желябин Д. В. Применение методов машинного обучения для решения задачи NLP классификации текста на основе анализа семантики естественного языка //Связь. – 2020. – Т. 4. – С. 440.
- 2 Акжолов Р. К., Верига А. В. предобработка текста для решения задач NLP //Вестник науки. – 2020. – Т. 1. – №. 3. – С. 66-68.
- 3 Гусев А. В. Перспективы нейронных сетей и глубокого машинного обучения в создании решений для здравоохранения //Врач и информационные технологии. – 2017. – №. 3.
- 4 Качнов С.А. Методы машинного обучения для обработки естественного языка //Наука молодых-будущее России. – 2020. – С. 41-44.
- 5 Bird S., Klein E., Loper E. Natural Language Processing with Python. Published by O'Reilly Media, 2019. - 504 p.
- 6 Рашид Т. Создаём нейронную сеть, Вильямс – СПб.: ООО “Альфа-книга”, 2017. - 272 с.
- 7 Python 3 и PyQt5. Разработка приложений. – 2-е изд., перераб. и доп. / Н.А. Прохоренко, В.А. Дронов. – СПб.: БХВ-Петербург, 2019. – 832 с.: ил. – (Профессиональное программирование)
- 8 Зоткин А. С., Ворожцов А. С. Большие данные: современные технологии обработки информации // Информационные технологии. – 2016. – 150 с.
- 9 Шлюйкова Д. П. Большие данные: современные подходы к хранению и обработке // Наука, техника и образование. – 2016. – №. 1. – с. 81.
- 10 Назаренко Ю. Л. Обзор технологии «большие данные» (big data) и программно-аппаратных средств, применяемых для их анализа и обработки //European Science. – 2017. – №. 9. – с. 30.
- 11 Ковалевский А. Е., Ефремов Е. А. Большие данные //Новая наука: Стратегии и векторы развития. – 2016. – №. 6-1. – с. 28.

- 12 Емельянова Н., Партыка Т., Попов И. Устройство и функционирование информационных систем. – М.: Форум, Инфра, - М, 2016. - 448 с.
- 13 Data Mining. Извлечение информации из Facebook, Twitter, LinkedIn, Instagram, GitHub. — СПб.: Питер, 2020. — 464 с.
- 14 Data Science. Наука о данных с нуля: пер. с англ. СПб.: БХВ-Петербург, 2017. – 335 с.
- 15 Анализ социальных медиа на Python: пер. с англ. – М.: ДМК Пресс, 2018. – 288 с.
- 16 Telegram API. [Электронный ресурс]. Режим доступа: <https://core.telegram.org> (дата обращения: 05.06.2021).
- 17 Pandas: User Guide [Электронный ресурс]. URL: https://pandas.pydata.org/docs/user_guide/index.html (дата обращения 20.05.2021)
- 18 NumPy: Quickstart tutorial [Электронный ресурс]. URL: <https://numpy.org/doc/stable/user/quickstart.html> (дата обращения 20.05.2021)
- 19 Scikit-learn: BernoulliNB [Электронный ресурс]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html (дата обращения 23.05.2021)
- 20 Spacy: Advanced NLP with Spacy [Электронный ресурс]. URL: <https://spacy.io/usage/spacy-101> (дата обращения 10.05.2021)
- 21 ГОСТ 12.0.003-2015. Система стандартов по безопасности труда. Опасные и вредные производственные факторы. Классификация. – Введ. 2017-03-01. – М.: Стандартинформ, 2016. – 10 с.
- 22 СанПиН 1.2.3685-21. Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания : утв. постановлением гл. гос. санитар. Врача Рос. Федерации от 28.01.2021 № 2.

ПРИЛОЖЕНИЕ А

Внешний и внутренний документооборот учреждения

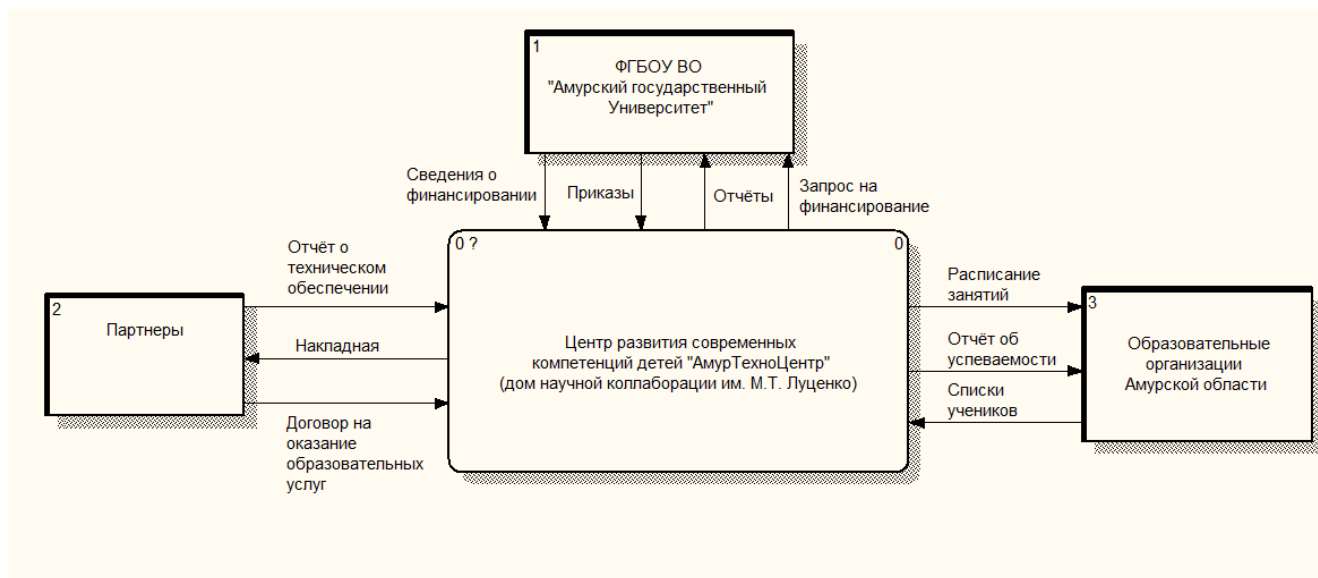


Рисунок А.1 – Внешний документооборот учреждения

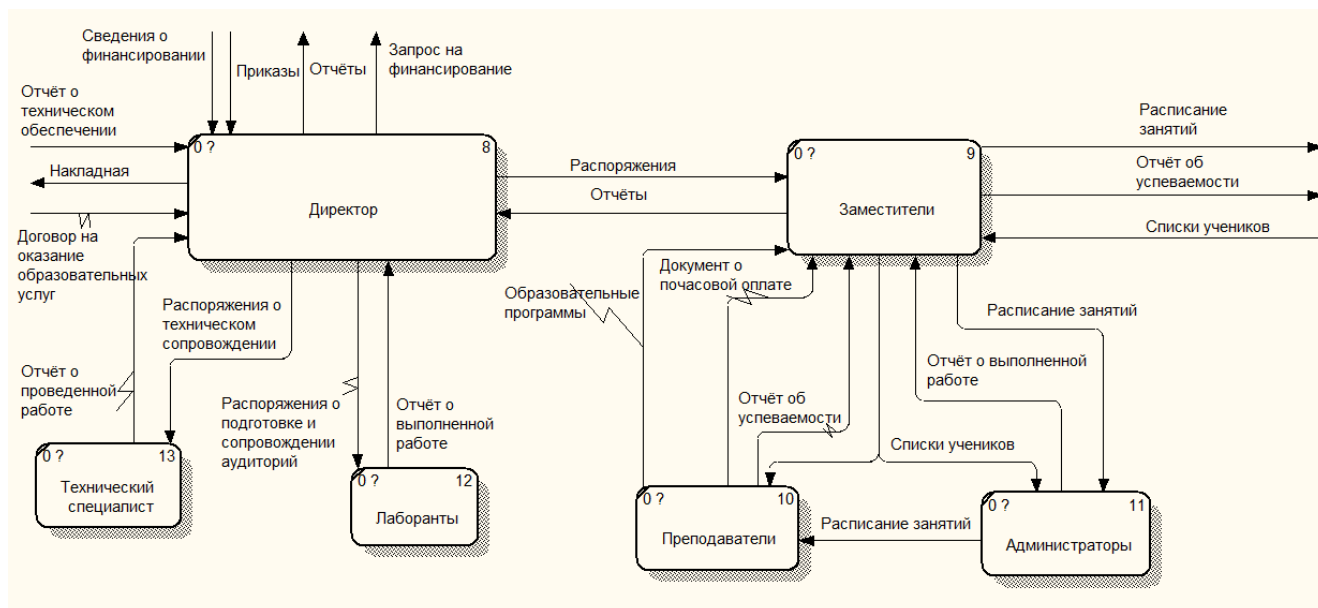


Рисунок А.2 – Внутренний документооборот учреждения

ПРИЛОЖЕНИЕ Б

Декомпозиция логики клиентского и серверного приложений

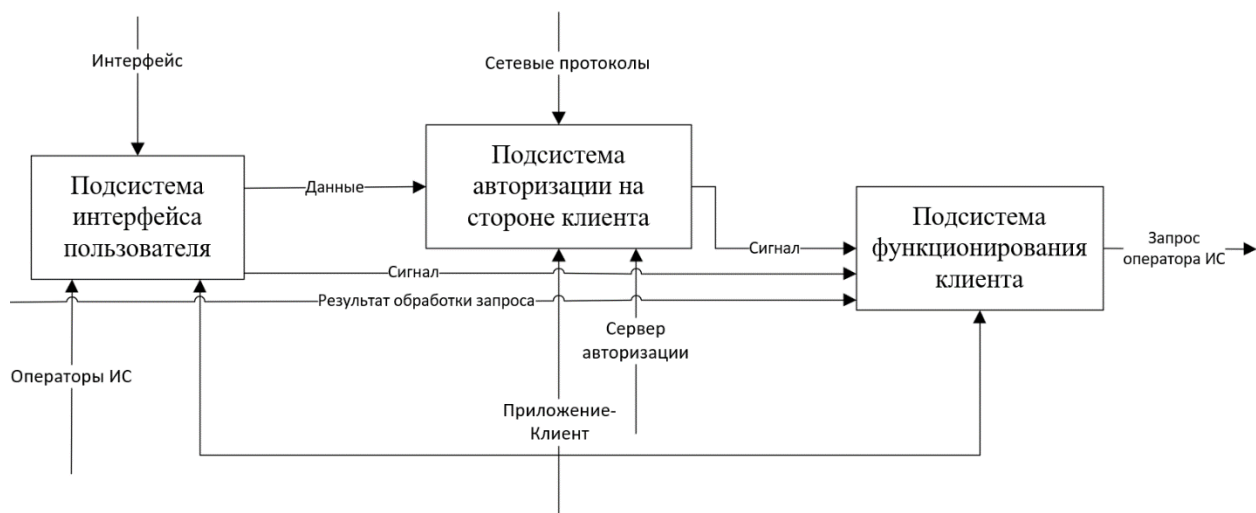


Рисунок Б.1 – Декомпозиция логики клиентского приложения

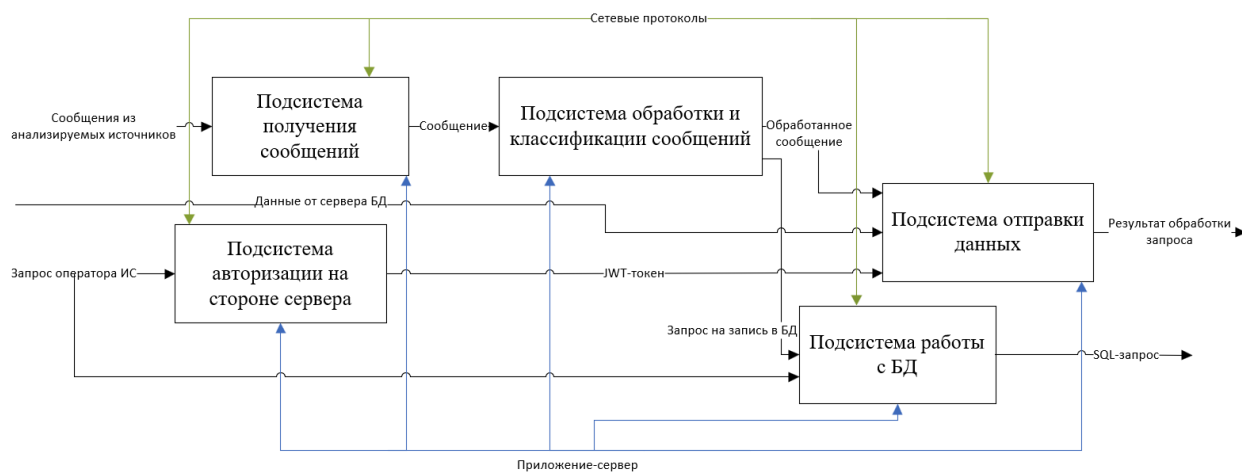


Рисунок Б.2 – Декомпозиция логики серверного приложения

ПРИЛОЖЕНИЕ В

Схемы проектирования базы данных

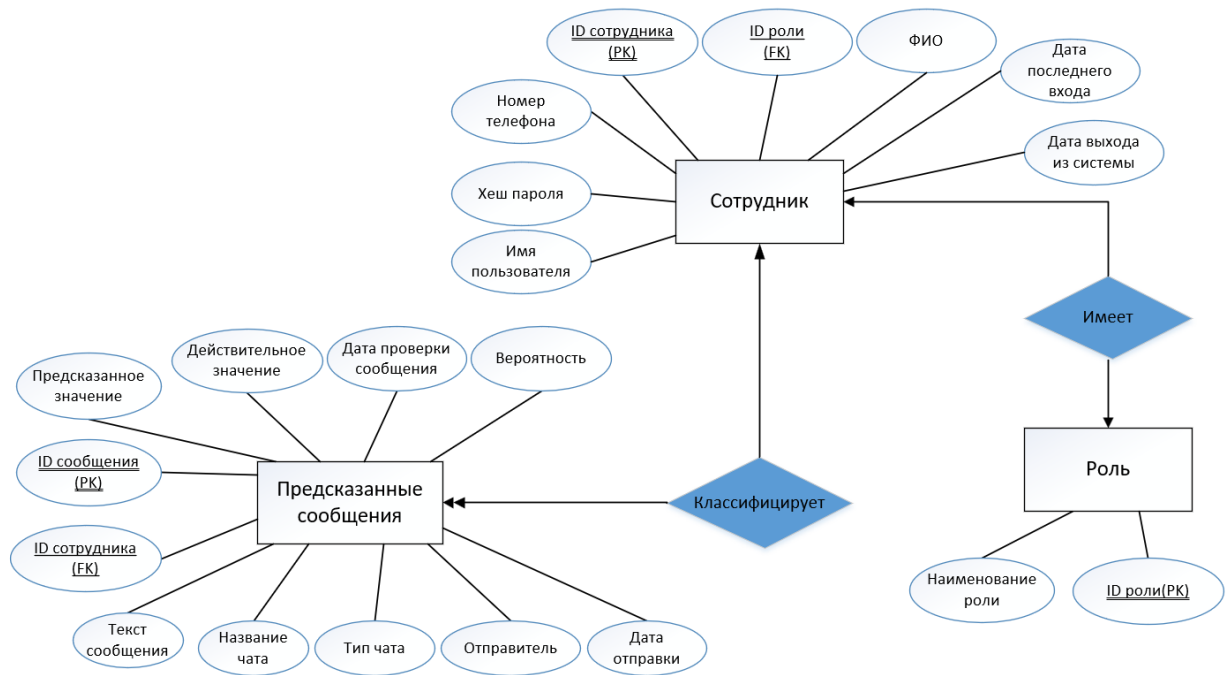


Рисунок В.1 – инфологическая модель БД в нотации Чена

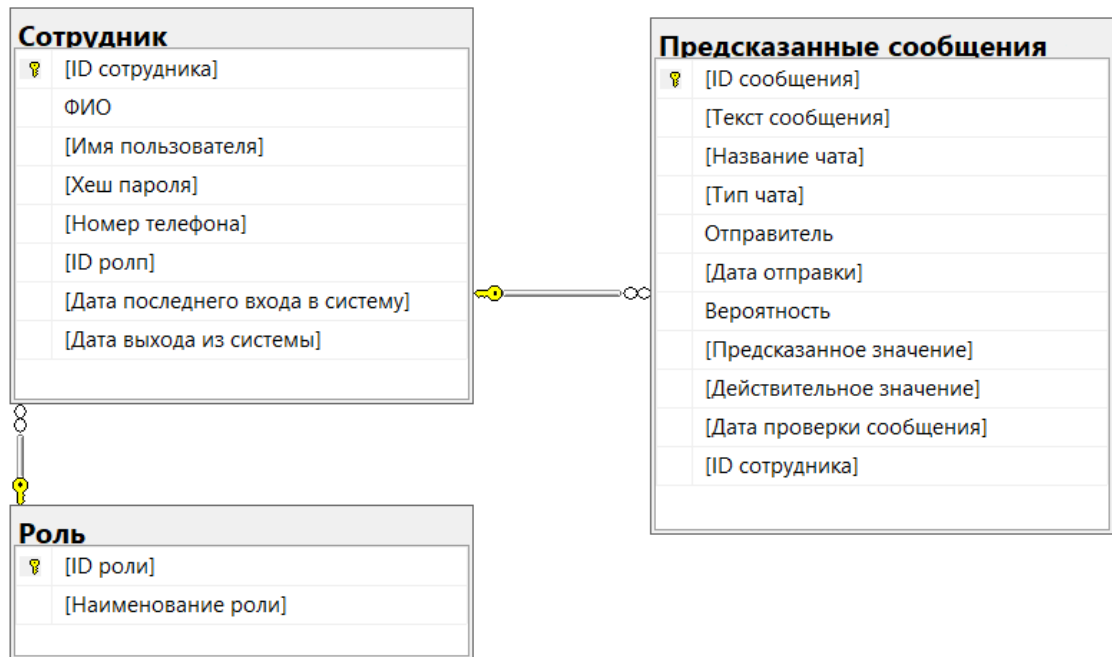


Рисунок В.2 – Логическая модель БД

ПРОДОЛЖЕНИЕ ПРИЛОЖЕНИЯ В

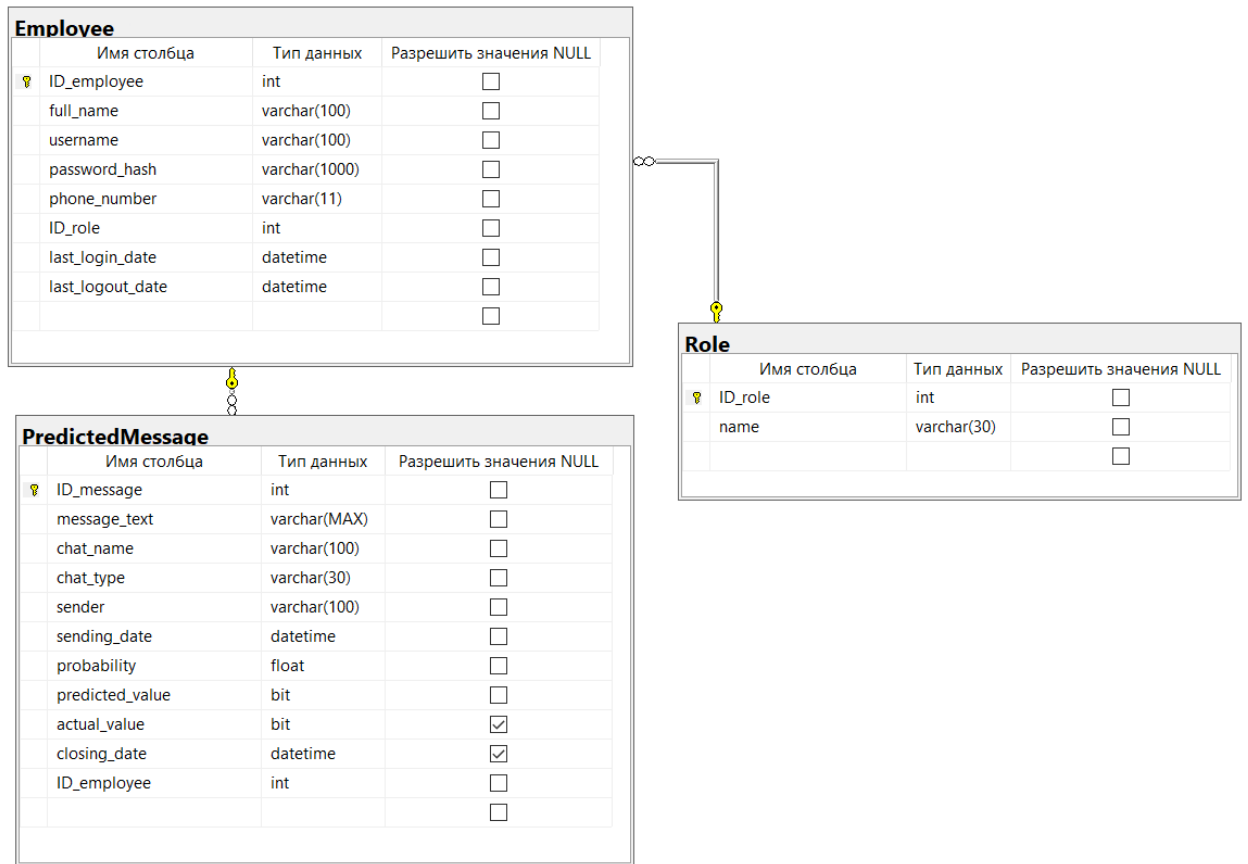


Рисунок В.3 – Физическая модель БД

ПРИЛОЖЕНИЕ Г

Пример автоматически сгенерированного отчёта

Отчёт о чрезвычайной ситуации №122

Ответственное лицо Петров Пётр Петрович
Дата ЧС 2021-06-21 22:10:41
Отправитель сообщения Геннадий Плахонин(NejdanX)
Источник сообщения Ученическая группа
Ответ системы о ЧС Сообщение является ЧС
Действительное значение Сообщение является ЧС
Класс ЧС Социальная
Виды ЧС (по классу) Терроризм

Краткое описание ЧС, причины её возникновения

Террорист захватил самолёт и один из преподавателей, который находился в этом самолёте, написал об этом сообщение.

Последствия ЧС и её масштаб

Люди сильно испугались, но обошлось без пострадавших. Из-за того, что полиция смогла связаться с пилотом и дать ему четкие указания, удалось убедить террористов в том, что нужна экстренная посадка из-за погодных условий. В результате, полиция смогла предотвратить акт терроризма.

Предпринятые действия для предотвращения последствий ЧС или их минимизации

Система сообщила о факте ЧС. Оператор ИС уведомил полицию и дал необходимую информацию о номере рейса и пассажире, который написал данное сообщение. В результате, удалось своевременно среагировать на данный инцидент.

Ответственное лицо

«__»_____ 20____ Петров П.П.

Директор

«__»_____ 20____ Еремина В.В.

Рисунок Г.1 – Сгенерированный отчёт

ПРИЛОЖЕНИЕ Д

Техническое задание

1 ОБЩИЕ СВЕДЕНИЯ

1.1 Полное наименование системы

Информационная система для бинарной классификации сообщений к теме ЧС методами машинного обучения

1.2 Область применения

Данная информационная система создается для автоматизации обработки сообщений из источников сообщений организации. Программный продукт автоматизировано определяет факт наличия ЧС в тексте и уведомляет об этом оператора ИС.

1.3 Наименование предприятий разработчика и заказчика системы

Разработчик: студент группы 755-об, факультета математики и информатики, Амурского государственного университета Плахонин Геннадий Павлович.

Заказчик: Центр развития современных компетенций детей «АмурТехноЦентр» «ДОМ НАУЧНОЙ КОЛЛАБОРАЦИИ ИМЕНИ АКАДЕМИКА РАН М.Т.ЛУЦЕНКО»

Фактический адрес: 675028, Амурская область, г. Благовещенск, ул. Игнатъевское шоссе, 7/1,

Телефон: +74162234747

1.4 Перечень документов

Документы, на основании которых создается ИС:

ГОСТ 34.601-89 – техническое задание на проектирование автоматизированной системы;

РД 50-34.698-90. Методические указания. Информационная технология. Комплекс стандартов на автоматизированные системы. Автоматизированные системы. Требования к содержанию документов.

1.5 Плановые сроки начала и окончания работы

Срок начала работ: 20.02.2021г.

Срок окончания работ: 24.06.2021г.

В процессе разработки сроки могут быть уточнены.

1.6 Сведения об источниках и порядке финансирования работ

Данный проект является учебным и выполняется без привлечения каких-либо финансовых средств.

1.7 Порядок оформления и предъявления заказчику результатов работ по созданию системы

Результаты работы предъявляются Заказчику в виде:

- функционирующей ИС, представляющего собой приложение с графическим интерфейсом;
- исходного программного кода;

Результаты предоставляются Исполнителем в сроки, установленные Заказчиком. Приемка системы осуществляется Заказчиком в установленном порядке. Порядок предъявления системы, ее испытаний и окончательной приемки определен в разделе 7 настоящего технического задания. Одновременно с предъявлением Системы производится сдача разработанного Исполнителем комплекта документации согласно разделу 8 настоящего ТЗ.

2 НАЗНАЧЕНИЕ И ЦЕЛИ СОЗДАНИЯ СИСТЕМЫ

2.1. Назначение системы

Проектируемая система должна быть ориентирована на решение следующих задач:

- автоматизированное получение сообщений;
- автоматизированная обработка и бинарная классификация сообщений по факту ЧС;
- сортировка сообщений по факту ЧС, вероятности ответа системы о принадлежности сообщения к ЧС и по дате получения этих сообщений;

- распределение нагрузки между операторами ИС и повышение скорости реагирования на ЧС разного характера;
- хранение и визуализация данных для удобной работы с ними оператора ИС;
- предоставление отчётов на основе данных о ЧС за выбранный период;
- автоматическое создание резервных копий БД;
- введение логов о действиях и ошибках в системе.

2.2. Цели создания системы

Целью создания системы является повышение скорости реагирования на ЧС, распределение нагрузки между несколькими операторами ИС, снижение временных затрат на мониторинг и классификацию сообщений из разных источников информации организации.

3 ХАРАКТЕРИСТИКА ОБЪЕКТА АВТОМАТИЗАЦИИ

3.1 Краткие сведения об объекте автоматизации

Объектом автоматизации является центр развития современных компетенций детей «АмурТехноЦентр» «дом научной коллаборации».

Автоматизации подлежат процессы:

- распределение нагрузки между несколькими операторами ИС;
- получение, обработка, хранение и представление информации операторам ИС;
- автоматическая очистка БД от сообщений не относящихся к теме ЧС или автоматическое подтверждение таких сообщений;
- ведение логов в системе и создание резервных копий;
- формирование отчётности о ЧС за определенные периоды.

3.2 Сведения об условиях эксплуатации и о характеристике окружающей среды.

Помещения (рисунок Д.1), в который предполагается размещение рабочего места, а также технических средств, должны, соответствовать согласованным показателям температуры, влажности и освещённости.

Условия эксплуатации должны соответствовать нормальным климатическим условиям, определённым в ГОСТ 27201-87 и иметь следующие значения:

- искусственное освещение в помещениях эксплуатации компьютеров должно осуществляться системой равномерного освещения;
- температура воздуха в помещении от 15 °С до 25 °С;
- относительная влажность воздуха в помещении от 40% до 60% при температуре 25 °С;
- атмосферное давление от 630 мм. Рт. Ст. до 800 мм Рт. Ст.

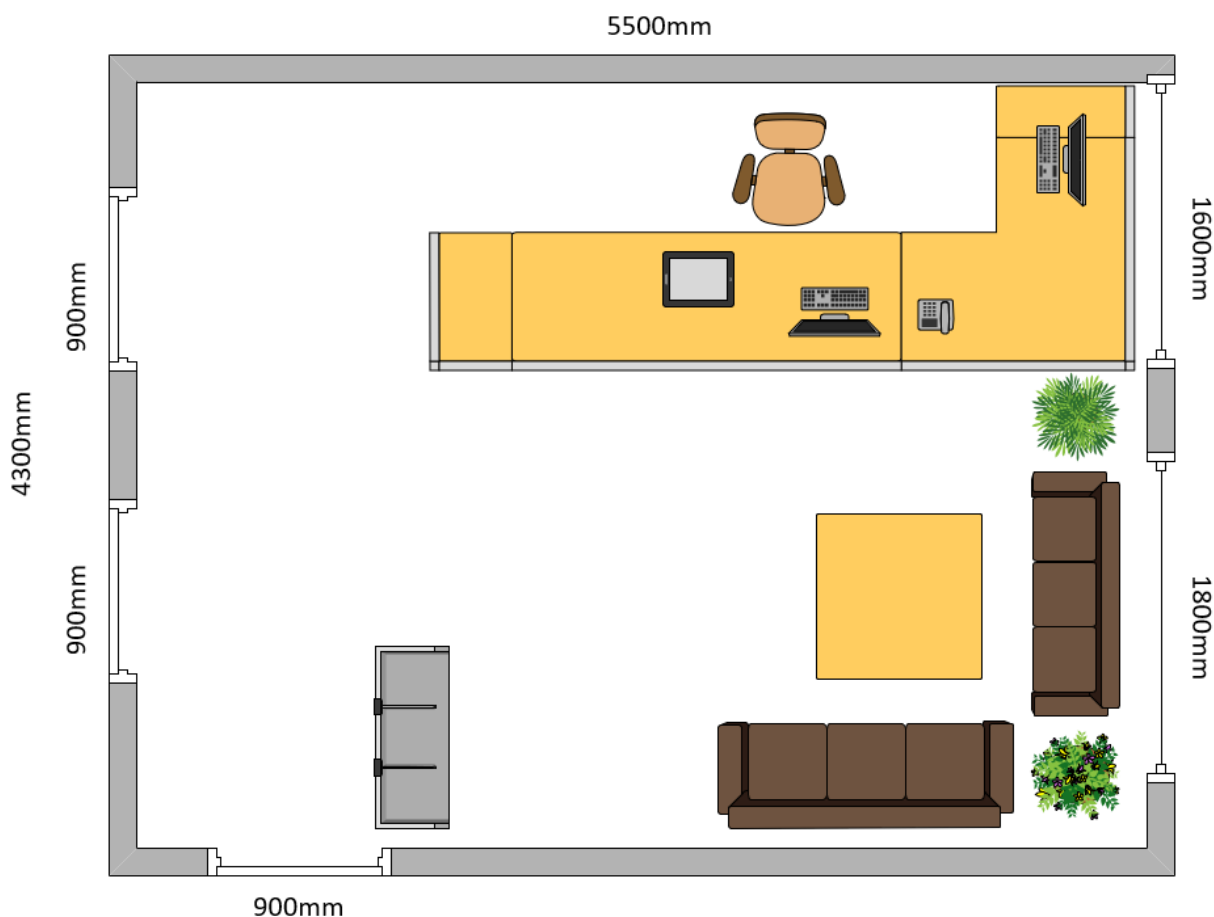


Рисунок Д.1 - Схема помещения с ПЭВМ

4 ТРЕБОВАНИЯ К СИСТЕМЕ

4.1 Требования к системе в целом

4.1.1 Требования к структуре и функционированию системы

4.1.1.1 Перечень подсистем и их назначение

Разрабатываемая система должна состоять из следующих функциональных подсистем:

- подсистема авторизации;
- подсистема получения сообщений;
- подсистема обработки и классификации сообщений;
- подсистема хранения данных;
- подсистема оповещения должностных лиц на терминалы связи;
- подсистема оповещения лиц, попадающих в зону поражения;
- подсистема по работе оператора ИС с сообщениями;
- подсистема пользовательского интерфейса.

4.1.1.2 Требования к числу уровней иерархии и степени централизации

ИС должна быть централизованной, т.е. все данные должны располагаться в центральном хранилище. При этом, система должна иметь трехуровневую клиент-серверную архитектуру (клиент, сервер приложений, сервер БД).

4.1.1.3 Требования к способам и средствам связи для информационного обмена между компонентами системы

Сервер приложений должен иметь постоянный доступ к сети интернет для сбора информации из источников сообщений. Предполагается, что сервер собирает информацию с помощью API анализируемого сервиса. Для передачи информации от сервера к клиенту необходимо, чтобы они находились в одной локальной сети. В качестве протоколов взаимодействия используются протоколы транспортного и сетевого уровня TCP и IP соответственно.

4.1.1.4 Требования к режимам функционирования системы

Система должна поддерживать следующие режимы функционирования:

- Основной режим, когда все подсистемы ИС выполняют все свои основные функции.

- Режим ограниченного функционирования, когда одна или несколько подсистем не выполняют свои функций.

В основном режиме, функционирование сервера должно обеспечиваться на протяжении 24 часов, 7 дней в неделю, а работоспособность клиента на протяжении всего рабочего дня, определенного Заказчиком.

В режиме ограниченного функционирования проводится техническое обслуживание, модернизация аппаратно-программного комплекса и устранение аварийных ситуаций. Вышеуказанные работы предпочтительнее проводить вечером или ночью, когда рабочий день окончен или в выходные дни, когда количество входящих сообщений сильно сокращается. Время устранения аварийных ситуаций не должно превышать более суток.

4.1.1.5 Перспективы развития, модернизация системы

Модернизация системы может происходить в двух направлениях: модернизация программного обеспечения и модернизация аппаратного обеспечения комплекса.

При модернизации программного обеспечения могут вноситься изменения или осуществляться дополнения в необходимые для функционирования программной системы. К ним относятся:

- увеличение эффективности классификации сообщений методами машинного обучения с помощью большего объема данных или применение новейших алгоритмов, которые показывают более высокую точность классификации;

- деление сообщений, которые относятся к теме ЧС, на несколько классов, для более точечной работы с каждым классом. Например, ЧС могут делиться по следующим признакам: природного, техногенного, биолого-социального, экологического характера. В свою очередь, каждая из этих групп

также может быть декомпозировано на более мелкие, в зависимости от нужд конкретного предприятия;

– распределение нагрузки между экспертами с учетом текущей загруженности каждого из них;

Модернизация аппаратного обеспечения комплекса должна происходить путем приобретения новых или модернизации старых аппаратных средств.

4.1.2 Требования к численности и квалификации персонала

4.1.2.1 Требования к численности персонала

В рамках разрабатываемой системы отсутствуют ограничения к численности персонала, так как количество операторов ИС, непосредственно работающих с системой, неограниченно.

4.1.2.2 Требования к квалификации персонала

Перед внедрением и эксплуатацией новой ИС на предприятие, необходимо произвести инструктаж персонала, ознакомить сотрудников с руководством пользователя, назначить администратора и пользователей ИС, с занесением их данных в БД.

Администратор ИС

Необходимо выделить сотрудника, который будет поддерживать функционирование в штатном режиме технических и программных средств ИС, проводить её модернизацию при необходимости. Администратор ИС должен иметь опыт работы с БД и обладать базовыми знаниями в области информационных и сетевых технологий. Для него отдельно создаётся интерфейс для более удобного управления настройками сервера и пользователями системы.

Пользователь ИС

Пользователю ИС необходимо обладать базовыми знаниями работы с компьютером и иметь знания предметной области, для своевременного и грамотного реагирования на сообщения.

4.1.2.3 Требования к режиму работы персонала

Пользователи и администратор ИС, работающие с ИС и выполняющие функции её сопровождения и обслуживания, должны работать в соответствии с основным рабочим графиком подразделений Заказчика.

4.1.3 Требования к надежности

4.1.3.1 Состав показателей надежности системы в целом.

Надежность системы обеспечивается за счёт применения организационного-технических мероприятий и программно-аппаратных средств.

Надежность должна обеспечиваться за счёт:

- защиты от некорректных действий пользователя системы;
- своевременного выполнения процессов администрирования системы;
- соблюдения правил эксплуатации и планового технического обслуживания ИС, предварительного обучения пользователей и обслуживающего персонала;
- выполнения резервных копий БД и предварительно обученных моделей для классификации на внешние носители информации;

4.1.3.2 Перечень аварийных ситуаций, по которым регламентируются требования к надежности.

Под аварийной ситуацией понимается нарушение работоспособности той или иной подсистемы ИС.

В ходе функционирования системы возможны следующие аварийные ситуации, которые могут повлиять на надежность работы системы:

- сбой в электроснабжении сервера, рабочей станции пользователей системы или обеспечения локальной сети;
- ошибки в ИС, которые не были выявлены при отладке и испытании системы;

- сбои в сервисах, откуда система получается сообщения для анализа.

4.1.3.3 Требования к надежности технических средств и программного обеспечения

Надежность системы должна обеспечиваться благодаря использованию современных и проверенных технических средств, постоянной профилактикой технических средств.

Вопросы надежности электроснабжения необходимо решить с помощью:

- комплектация сервера источником питания до полного резервного копирования БД на внешний носитель;
- оповещения администратора и пользователей ИС о переходе на автономный режим.

К организационным мероприятиям обеспечение надежности относятся:

- ознакомление пользователей ИС с Руководством пользователя;
- своевременного выполнения процессов администрирования;
- соблюдения правил эксплуатации и технического обслуживания программно-аппаратных средств;
- своевременное выполнение процедур резервного копирования данных.

Надежность программного обеспечения подсистем должна обеспечиваться за счет:

- надежности общесистемного ПО и ПО, разрабатываемого Разработчиком;
- проведением комплекса мероприятий отладки, поиска и исключения ошибок.
- ведением журналов системных сообщений и ошибок по подсистемам для последующего анализа и изменения конфигурации.

4.1.4 Требования к эргономике и технической эстетике

Система должна иметь человеко-машинный интерфейс, удовлетворяющий следующим требованиям:

- взаимодействие системы и пользователя должно осуществляться на русском языке, за исключением системных сообщений, не подлежащих русификации;
- должно быть реализовано отображение на экране только тех возможностей, которые доступны конкретному пользователю в соответствии с его функциональной ролью в системе;
- допустима видимость предоставляемой информации на экране;
- допустимая цветопередача.

Интерфейс должен соответствовать современным эргономическим требованиям и обеспечивать удобный доступ к основным функциям и операциям системы.

Подсистема визуализации должна обеспечивать удобный и понятный для конечного пользователя интерфейс построения графиков разных типов, с общими настройками для каждого типа графика. Например, можно указать диапазон дат или конкретный источник информации для разных типов графиков. Также, интерфейс должен быть локализованным.

Управление системой осуществляется преимущественно с помощью манипулятора типа «мышь». Для ввода и/или редактирования текстовых данных и числовых полей используется клавиатура.

4.1.5 Требования к эксплуатации, техническому обслуживанию, ремонту и хранению

Пользователи обязаны быть проинформированы о правилах использования технических средств и работы с программой и с оборудованием, на котором используется данная программа.

Устройство хранения должно быть защищено от внешних физических воздействий, в качестве переноса и хранения может быть любой диск для хранения данных.

В случае аварии требования заключаются в сохранности информации при сбоях в работе системы, а также при допущении ошибок пользователей при работе с программой.

Программные средства администратора системы должны обеспечивать:

- 1 при выходе технических средств из строя, должна обеспечиваться ее замена без потери функциональной подсистемы;
- 2 полное или частичное восстановление потерянной информации;
- 3 протокол действий при возникновении нештатной ситуации.

4.1.6 Требования к защите информации от несанкционированного доступа

- Разграничение прав доступа пользователей и администраторов ИС должно строиться по принципу "что не разрешено, то запрещено";
- Пользователь ИС должен пройти процесс идентификации и аутентификации с целью авторизации в системе;
- Запись последнего время входа и выхода пользователя должна помещаться в БД;
- Права на редактирование внутренних настроек системы должны быть доступны только администраторам системы;
- Разграничение по получаемым сообщениям должно обеспечиваться на стороне сервера для каждого активного пользователя;

4.1.7 Требования по сохранности информации при авариях

Благодаря дополнительному источнику питания сервера предполагается, что будет сделана полная резервная копия на внешний носитель. Также, после аварии на сервере необходимо средствами самой СУБД проверить БД на наличие ошибок и по возможности их исправить. Ущерб программному обеспечению мало вероятен, только если пользовательские настройки, которые записаны в файле конфигурации на локальном компьютере самого пользователя, не будут искажены в последствии. В этом случае, необходимо по возможности восстановить

пользовательские настройки или сбросить их до заводских. То же самое касается конфигурации самого сервера. Если при запуске сервер не найдёт файл конфигурации, то просто сбросится до заводских настроек.

4.2 Требования к функциям, выполняемым системой

4.2.1 Требования к подсистеме. Перечень функций подсистем

- Подсистема авторизации: идентификация пользователя, выдача формирования JWT-токена;
- Подсистема получения сообщений: получение сообщений из анализируемых источников с помощью API или парсинга.
- Подсистема обработки и классификации сообщений: предобработка сообщений, написанных на естественном языке, перевод отдельных токенов в векторную форму, классификация сообщений.
- Подсистема работы с БД: запись и хранение данных в БД, чтение данных из БД.
- Подсистема передачи данных операторам ИС: отправка запросов на рассмотрение сообщений активным операторам ИС.
- Подсистема интерфейса пользователя.
- Подсистема функционирования клиентской логики: подтверждение или отклонение факта ЧС сообщения, визуализация данных, запрос на запись ответов в БД, изменение настроек сервера, управление пользователями ИС.

4.2.2 Временной регламент реализации каждой функции

При выполнении функций, связанных с загрузкой сервера или долгим ответом от сервиса, который является источником сообщений, возможная естественная временная задержка в обработке данных. При передаче уже обработанных сообщений от сервера до клиента, также возможна естественная временная задержка, которая зависит от качества соединения.

При поиске доступного оператора ИС может произойти задержка, если не один из операторов недоступен. Тогда сообщения накапливаются в

специальном временном хранилище и при доступности оператора, отправляются ему постепенно.

4.3 Требования к видам обеспечения

4.3.1 Требования к информационному обеспечению

В состав информационного обеспечения программы входит база данных (внутримашинное обеспечение), входная, внутренняя и выходная информация.

В качестве входной информации выступает:

- сообщения из анализируемых источников;

В качестве внутренней информации выступает:

- предварительно обученный классификатор сообщений.

Выходной информацией служат:

- изменения в объектах БД и добавление новых записей;
- файл БД с внесенными в него изменениями;
- отчет о ЧС.

4.3.2 Требования к лингвистическому обеспечению

Для создания данной программы необходимы знания языка программирования Python, языка разметки HTML, формального языка описания внешнего вида документа CSS и языка запросов SQL.

Для описания предметной области (объекта автоматизации) должен использоваться Erwin.

Язык интерфейса должен быть полностью локализован на русский язык, взаимодействие системы и пользователя должно осуществляться на русском языке, за исключением системных сообщений, не подлежащих русификации.

4.3.3 Требования к программному обеспечению

Для реализации и эксплуатации ИС, пользователь должен иметь установленную операционную систему семейства Windows не старше Windows 7. Рекомендуется использовать Windows 10. Для просмотра и

редактирования отчётов, созданных в системе, необходим текстовый редактор Microsoft Word.

4.3.4 Требования к техническому обеспечению

Аппаратные средства (серверное оборудование) Системы предоставляется Заказчиком. Система может располагаться как на выделенном физическом, так и на виртуальном сервере.

Список рекомендуемых технических характеристик сервера:

- объём ОЗУ не менее 8 ГБ типа DDR4-2666, с возможностью расширения;
- RAID1-массив из двух SAS-дисков, которые обеспечивают multipath, с объёмом 500 ГБ каждого;
- Intel Xeon E2226G с тактовой частотой 3.4 ГГц;
- серверная плата Intel M10JP2SB;
- бесперебойный блок питания на 500 Вт;
- монитор;
- клавиатура и компьютерная мышь;
- сетевой коммутатор;

Минимальные требования для работы на персональных компьютерах, имеющих следующие минимальные характеристики:

- процессор с тактовой частотой от 2.0 ГГц;
- ОЗУ – 2 ГБ или более
- К дополнительным требованиям относятся:
- устройство ввода информации: клавиатура, мышь;
- монитор;
- доступ в интернет.

4.3.5 Требования к математическому обеспечению

Требования к математическому обеспечению не предъявляются.

4.3.6 Требования к организационному обеспечению

Для корректной эксплуатации системы следует разработать руководство пользователей и провести инструктаж сотрудников. В соответствии с функционалом внедряемой системы, полномочия некоторых сотрудников могут быть изменены, о чём сотрудники должны быть заранее проинформированы.

Также, создаётся отдельное руководство для администратора системы, так как он несёт наибольшую ответственность за контролем её функционирования и обеспечение.

5 СОСТАВ И СОДЕРЖАНИЕ РАБОТ ПО СОЗДАНИЮ СИСТЕМЫ

5.1 Перечень стадий и этапов работ по созданию системы

Этапы, которые необходимо выполнить по созданию программы:

1 этап – разработка технического задания, определение требований к программе, и стадий, этапов и сроков разработки программы, выбор языков программирование, согласование и утверждение технического задания;

2 этап – Изучение предметной области, анализ процессов деятельности организации, обоснование необходимости создания ИС. В конце этого этапа будет принято решение о необходимости создания ИС и разработаны диаграммы внешнего и внутреннего документооборота;

3 этап – Проектирование ИС, описание характеристик функциональных подсистем, основание выбора средств разработки и проектирование БД, включающее инфологическое, логическое и физическое проектирование;

4 этап – Разработка программного продукта. В конце этого этапа будет разработан прототип работающего приложения;

5 этап – согласование созданного прототипа с требованиями заказчика;

6 этап – доработка программного продукта, проведение испытаний системы;

7 этап – согласование созданной ИС с требованиями заказчика;

8 этап – внедрение и сопровождение системы, обучение пользователей к работе с системой, устранение обнаруженных неполадок, установка и настройка ПО.

5.2 Состав организации исполнителя работ

Все работы выполняются студентом Амурского государственного университета Плахониным Геннадием Павловичем.

6 ТРЕБОВАНИЯ К ПРИЕМКЕ-СДАЧЕ ПРОЕКТА

6.1 Виды, состав, объем и методы испытаний системы и ее составных частей

Испытания должны быть организованы и проведены в соответствии с ГОСТ 34.603.92 «Информационная технология. Виды испытаний автоматизированных систем».

Должны быть проведены следующие виды испытаний:

- предварительные испытания;
- опытная эксплуатация (ОЭ);
- приёмочные испытания.

На этапе предварительных испытаний проводится автономное тестирование всей системы в целом или её модулей. База данных заполняется данными, которые могут (должны) приходиться из внешних источников.

На этапе опытной эксплуатации проверяется работоспособность системы на реальных данных, пользователях и выполнении реальных задач. В ходе этого этапа устраняются основные недостатки системы.

Приёмочные испытания проводят для определения соответствия системы техническому заданию, оценки качества опытной эксплуатации и решения вопроса о возможности приемки системы в постоянную эксплуатацию.

6.2 Общие требования к приемке работ по стадиям

Предварительные испытания и эксплуатация проводятся на аппаратных средствах Исполнителя.

По результатам испытаний возможны доработки и исправления. Выявленные в ПО и документации недостатки Исполнитель исправляет в специально оговоренные после проведения испытаний сроки.

Сдача-приёмка работ производится поэтапно, в соответствии с календарным планом.

Все создаваемые в рамках настоящей работы программные изделия (за исключением покупных) передаются Заказчику, как в виде готовых модулей, так и в виде исходных кодов, представляемых в электронной форме на стандартном машинном носителе.

7 ТРЕБОВАНИЯ К СОСТАВУ И СОДЕРЖАНИЮ РАБОТ ПО ПОДГОТОВКЕ ОБЪЕКТА АВТОМАТИЗАЦИИ К ВВОДУ СИСТЕМЫ В ДЕЙСТВИЕ

Перед вводом в эксплуатацию готового продукта разработчик должен договориться с руководителем организации о временном промежутке, в течение которого он обязан внедрить разработанный программный продукт. Под внедрением понимается комплекс мероприятий, включающий обучение персонала, настройку системы для дальнейшего использования, предоставление им необходимой документации для системы, ознакомление инструктора с его обязанностями.

7.1 Преобразование входной информации к машиночитаемому виду

Вся поступающая в ИС информация должна быть приведена к виду, пригодному для обработки ЭВМ.

7.2 Организационные мероприятия

Под организационными мероприятиями понимаются ознакомление пользователей с «Руководством пользователя» и определение списка работ, результаты которых должны заноситься в ИС. Список определяется в рабочем порядке представителями Заказчика. Необходимо произвести реорганизацию или набор нового персонала, в случае необходимости.

7.3 Технические мероприятия

Под техническими мероприятиями понимаются установка на аппаратные средства соответствующую операционную систему и само программное обеспечение. Также, необходимо произвести настройку системы доступа и создание учетных записей.

7.4 Сроки и порядок комплектования штатов и обучения персонала

Сроки и порядок комплектования штатов и обучения персонала осуществляются в соответствии с требованиями Заказчика.

8 ТРЕБОВАНИЯ К ДОКУМЕНТИРОВАНИЮ

Документирование проекта автоматизации по ГОСТ 34 регламентируется следующими стандартами:

- ГОСТ 34.201-89 Виды, комплектность и обозначения документов при создании автоматизированных систем.
- РД 50-34.698-90 Методические указания. Информационная технология. Комплекс стандартов и руководящих документов на автоматизированные системы. Автоматизированные системы. Требования к содержанию документов.
- ГОСТ 34.602-89 – разработка технического задания для создания ИС.

При вводе программы в эксплуатацию пакет сопроводительных документов должен включать:

- техническое задание;
- описание программного продукта;
- исходный программный код;
- руководство пользователя;
- руководство администратора.