

Оглавление

РЕФЕРАТ	3
1 Оглавление ВКР	5
2 Конструкторский раздел	7
2.1 Формат и метод сбора данных	7
2.2 Средства реализации ботов в мессенджере Telegram	8
2.3 Декомпозиция системы	8
2.4 Методы машинного обучения	13
2.5 Алгоритмы векторизации текстовых сообщений	18
3 План исследований	21
3.1 Условия исследований	21
3.2 Исследование применимости моделей машинного обучения в методе распознавания суицидальных паттернов поведения человека по текстовым сообщениям	21
3.2.1 Градиентный бустинг	21
3.2.2 Метод случайного леса	21
3.2.3 Метод опорных векторов	21
3.2.4 Метод К-ближайших соседей	21
3.2.5 Логистическая регрессия	21
3.2.6 Перцептрон	21
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	23

РЕФЕРАТ

Расчетно-пояснительная записка к выпускной квалификационной работе “Метод распознавания паттернов суицидального поведения человека по текстовым сообщениям” содержит 94 с., 49 рис., 3 табл., 50 ист., 0 прил.

Ключевые слова: суицидология, суицидальность, анализ текста, поведение человека, машинное обучение, классификация.

Объектом разработки является метод распознавания паттернов суицидального поведения человека.

Цель работы: разработка и программная реализация метода распознавания паттернов суицидального поведения человека по текстовым сообщениям.

В аналитическом разделе рассматриваются термины предметной области, представлен интегративный теоретический подход в суицидологии и приведена информация об определении истинности суицидальных намерений. Описываются факторы повышенного суицидального риска и мероприятия по предотвращению самоубийств. Приводится статистика совершения самоубийств. С использованием классификации признаков паттернов суицидального поведения человека описываются форматы хранения проявления поведения человека. Представлена информация о задействованных в настоящее время алгоритмах в задачах классификации сообщений в сети Интернет. Приводится формализация задачи метода распознавания суицидальных паттернов поведения человека по текстовым сообщениям.

В конструкторском разделе описывается метод распознавания суицидальных паттернов поведения человека по текстовым сообщениям, а также формат и метод сбора задействованных в нем данных. Рассматриваются средства реализации автоматизированного средства сбора данных. Приводится диаграмма вариантов использования, декомпозиция задачи распознавания суицидального сообщения, а также диаграмма “сущность-связь” в нотации Чена. Определяется перечень задействованных методов машинного обучения и векторизации, а также приводятся их схемы работы.

В технологическом разделе определяются инструменты разработки средства сбора данных и средства распознавания суицидальных паттернов поведения человека по текстовым сообщениям. Реализован метод рас-

познавания суицидальных паттернов поведения человека по текстовым сообщениям. Представлены интерфейсы разработанных средств. Приводится описание обрабатываемых данных, а также анализ тональности сообщений и облаков слов каждого класса.

В исследовательском разделе проведено сравнительное исследование задействованных в методе алгоритмов машинного обучения и определено, какой из рассмотренных методов позволяет достичь лучших метрик точности с использованием матрицы ошибок и графиков оценок классификаторов.

Разработанный метод применим в сфере суицидологии. Автоматизированные средства поиска суицидальных сообщений, а также обычные анализаторы сообщений могут задействовать данный метод для обнаружения индивидов в суицидоопасных состояниях, а также для наблюдения людей, пребывающих в периоде пресуицида и постсуицида.

1. Оглавление ВКР

РЕФЕРАТ	4
ВВЕДЕНИЕ	6
1 Аналитический раздел	7
1.1 Интегративный подход в описании суицидального поведения .	7
1.2 Предотвращение самоубийств	13
1.3 Выделение и классификация признаков паттернов суицидального поведения человека	18
1.4 Форматы хранения проявление поведения человека	21
1.5 Классификация сообщений в сети Интернет	22
1.6 Формализация задачи метода распознавания суицидальных паттернов поведения человека по текстовым сообщениям . . .	24
2 Конструкторский раздел	28
2.1 Формат и метод сбора данных	28
2.2 Средства реализации ботов в мессенджере Telegram	29
2.3 Декомпозиция системы	29
2.4 Методы машинного обучения	34
2.5 Алгоритмы векторизации текстовых сообщений	39
3 Технологический раздел	42
3.1 Выбор инструментов разработки	42
3.2 Интерфейс средства сбора данных	43
3.3 Интерфейс средства распознавания суицидальных паттернов поведения человека по текстовым сообщениям	45
3.4 Описание обрабатываемых данных	47
4 Исследовательский раздел	52
4.1 Условия исследований	52
4.2 Исследование применимости моделей машинного обучения в методе распознавания суицидальных паттернов поведения человека по текстовым сообщениям	52

4.2.1	Градиентный бустинг	52
4.2.2	Метод случайного леса	57
4.2.3	Метод опорных векторов	62
4.2.4	Метод К-ближайших соседей	67
4.2.5	Логистическая регрессия	73
4.2.6	Перцептрон	79
ЗАКЛЮЧЕНИЕ		87
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ		90

2. Конструкторский раздел

В разделе описывается метод распознавания суицидальных паттернов поведения человека по текстовым сообщениям, а также формат и метод сбора задействованных в нем данных. Рассматриваются средства реализации автоматизированного средства сбора данных. Приводится диаграмма вариантов использования, декомпозиция задачи распознавания суицидального сообщения, а также диаграмма “сущность-связь” в нотации Чена. Определяется перечень задействованных методов машинного обучения и векторизации, а также приводятся их схемы работы.

2.1 Формат и метод сбора данных

В качестве задействованных в анализе данных используются текстовые сообщения. Для сбора данных потребуется использовать автоматизированное средство сбора суицидальных сообщений в мессенджере Telegram. Интерфейс программного обеспечения позволяет направить в систему хранения два типа сообщений: суицидальные и на суицидальную тематику.

Средство сбора должно предоставлять пользователю следующий функционал:

1. Получение информации о проекте;
2. Получение информации об отличиях суицидальных сообщений и сообщений на суицидальную тематику;
3. Направление примеров суицидальных сообщений;
4. Направление примеров сообщений на суицидальную тематику, но не относящихся к суицидальным.

Согласно 152-ФЗ “О персональных данных”, “персональные данные – любая информация, относящаяся к прямо или косвенно определенному или определяемому физическому лицу (субъекту персональных данных)” [1]. Таким образом, к персональным данным можно отнести фамилию, имя и отчество, дату и место рождения, адрес проживания, семейное, социальное и имущественное положение, образование, профессию, доходы и другое. В связи с этим средство сбора информации не обрабатывает и не хранит никаких персональных данных о пользователях, направивших сообщения.

2.2 Средства реализации ботов в мессенджере Telegram

В качестве представленных к использованию в качестве средства реализации бота в мессенджере телеграм могут быть задействованы популярные библиотеки:

- Python Telegram Bot [2],
- Telebot [3],
- Node Telegram Bot API [4],
- Telegram Bot Kotlin [5].

Python Telegram Bot – это библиотека, предоставляющая асинхронный интерфейс на ЯП Python для Telegram Bot API, которая совместима с версиями Python3.8 [6] и выше. Данная библиотека предоставляет высокий уровень абстракции и позволяет использовать объектно-ориентированный подход. Помимо реализации API, она также содержит ряд классов высокого уровня, упрощающих разработку ботов. Проект поддерживает строенную с асинхронным вводом-выводом. [2]

Telebot – библиотека на Python, содержащая в себе асинхронную и синхронную реализацию Telegram Bot API. Данный проект предоставляет более гибкий и низкоуровневый доступ к API Telegram, чем упомянутый Python Telegram Bot. [3]

Node Telegram Bot API – это библиотека для создания Telegram-ботов с использованием языка JavaScript [7] и платформы NodeJS [8]. Несмотря на распространенное использование, используемая версия API Telegram является устаревшей. [4]

Telegram Bot Kotlin – это библиотека для создания Telegram-ботов на ЯП Kotlin [9]. Отличительной особенностью является возможность разработки ботов на платформе Java Virtual Machine [10], а также поддержка Kotlin Coroutines. [5]

2.3 Декомпозиция системы

На рисунке 2.1 представлена диаграмма вариантов использования системы.

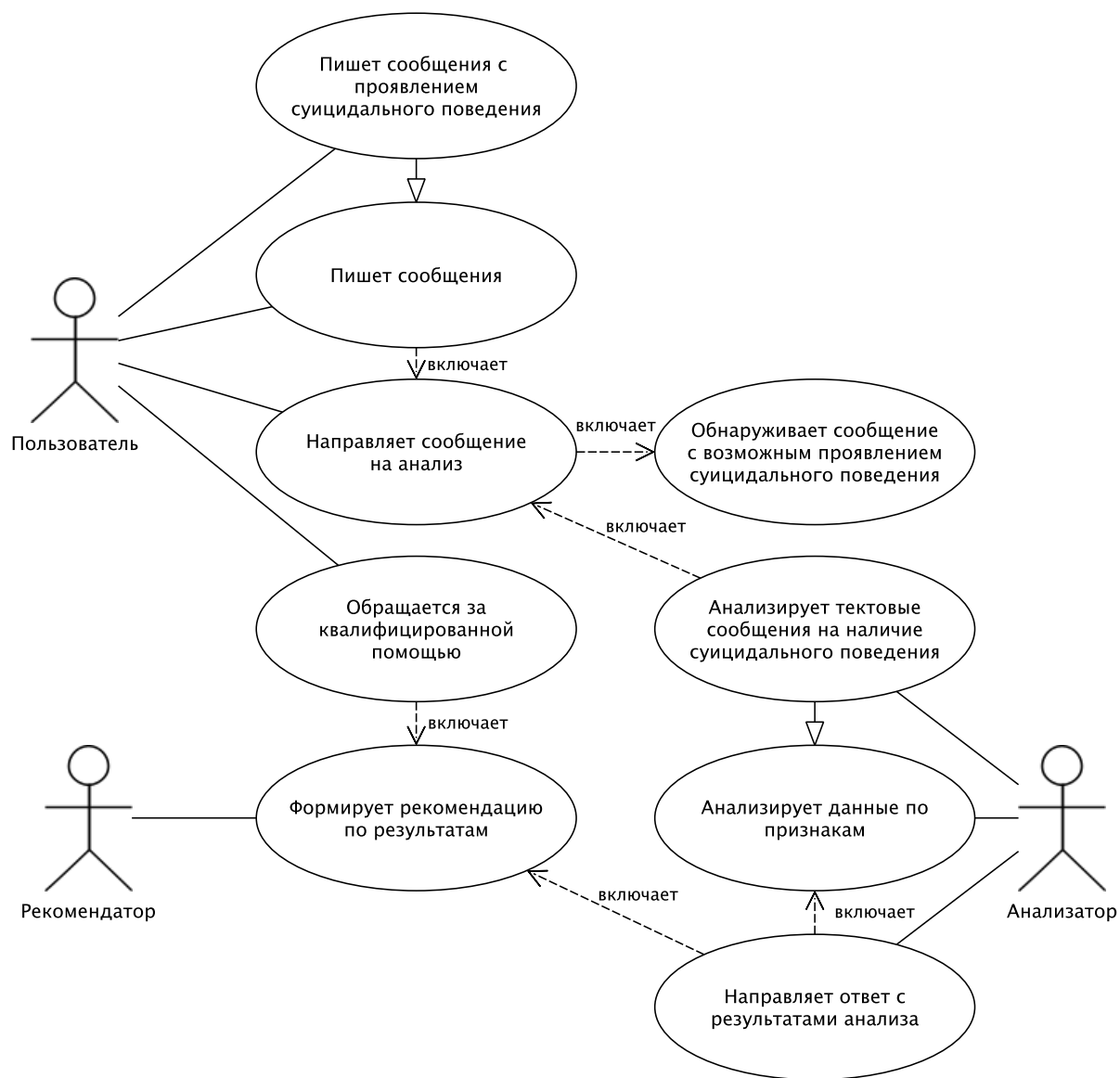


Рисунок 2.1 – Диаграмма вариантов использования системы.

На рисунке 2.2 представлена IDEF0 диаграмма первого уровня задачи определения наличия суицидальных паттернов в текстовом сообщении.

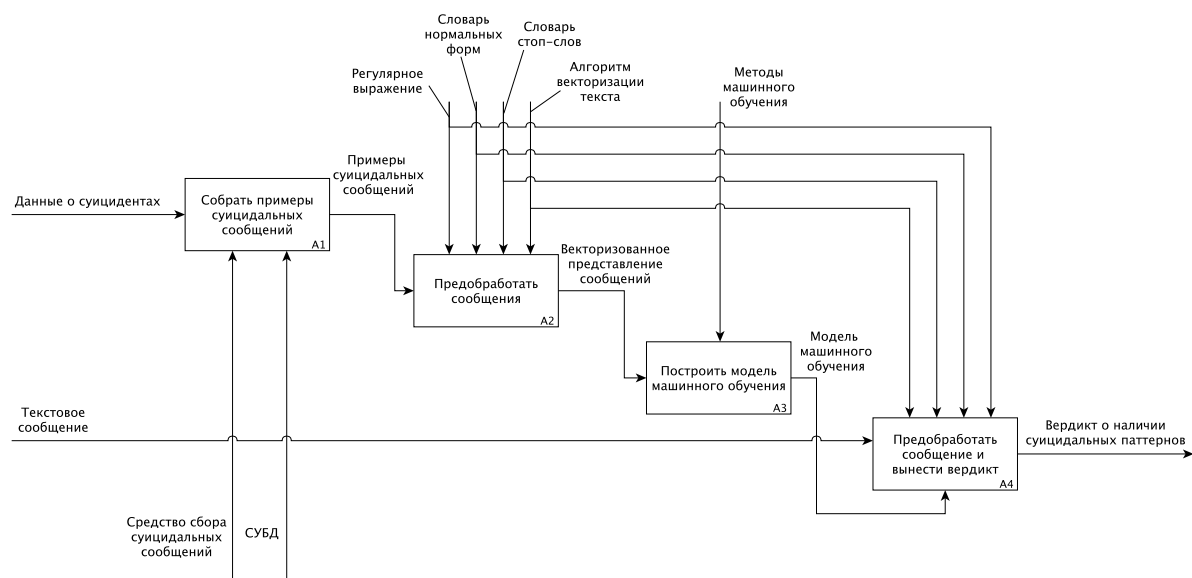


Рисунок 2.2 – IDEF0 диаграмма первого уровня.

Модуль A1 на рисунке 2.2 отвечает за сбор примеров суицидальных сообщений с использованием средства сбора суицидальных сообщений и СУБД. Средство сбора суицидальных сообщений в данном контексте подразумевает программное обеспечение, позволяющее пользователям вносить размеченные сообщения, то есть текст с указанным классом, к которому он относится. Затем эти данные будут задействованы в блоке предобработки сообщений.

Модуль A2 на рисунке 2.2 отвечает за предобработку данных, собранных на этапах работы модуля A1. Данный блок включает в себя следующие шаги: токенизацию, лемматизацию, удаление стоп-слов и векторизацию. В результате выполнения блока будет получено векторизованное представление поступивших в него примеров суицидальных сообщений, которое будет задействовано в дальнейшем в блоке построения модели машинного обучения. Декомпозиция рассматриваемого блока приведена на рисунке 2.3.



Рисунок 2.3 – IDEF0 диаграмма, декомпозиция блока A2.

Модуль A3 на рисунке 2.2 отвечает за построение модели машинного обучения, играющую роль классификатора сообщений. В качестве контроля блока выступают методы машинного обучения: градиентный бустинг, метод случайного леса, метод опорных векторов, метод К-ближайших соседей, логистическая регрессия и перцептрон. В результате выполнения будет получена модель машинного обучения, задействованная в блоке предобработки поступившего вне обучающей выборки сообщения и вынесения вердикта о его суицидальности.

Модуль A4 на рисунке 2.2 отвечает за предобработку поступившего в систему сообщения, а также его оценку на суицидальность. В качестве механизмов используются словарь нормальных форм, регулярное выражение и словарь стоп-слов для выполнения предобработки, а также модель машинного обучения для оценки сообщения.

На рисунке 2.4 представлена диаграмма “сущность-связь” в нотации Чена.

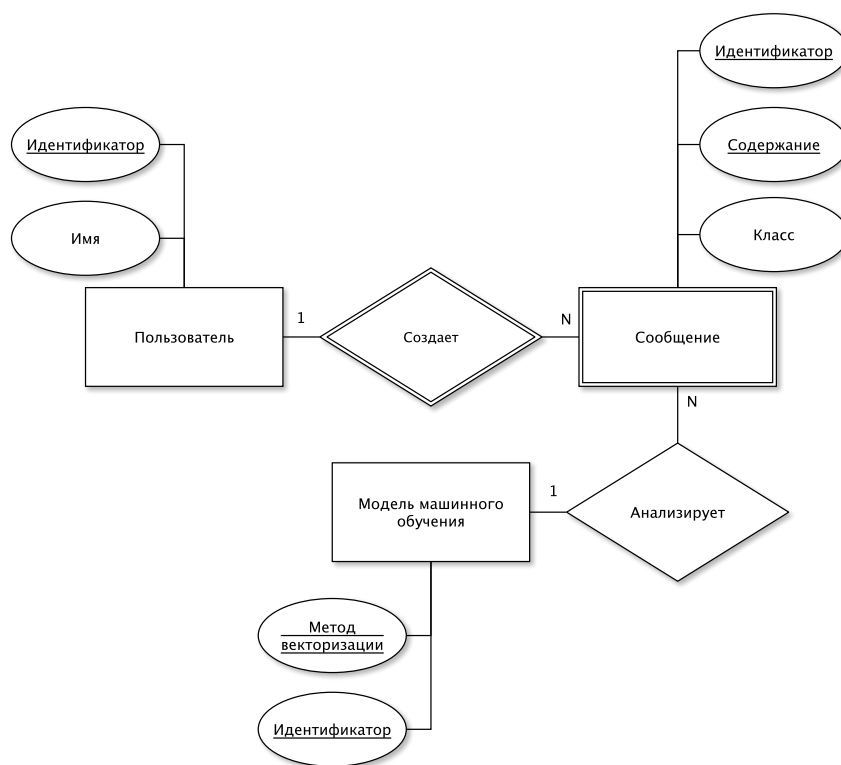


Рисунок 2.4 – Диаграмма “сущность-связь” в нотации Чена.

На рисунке 2.5 представлена схема программного обеспечения.

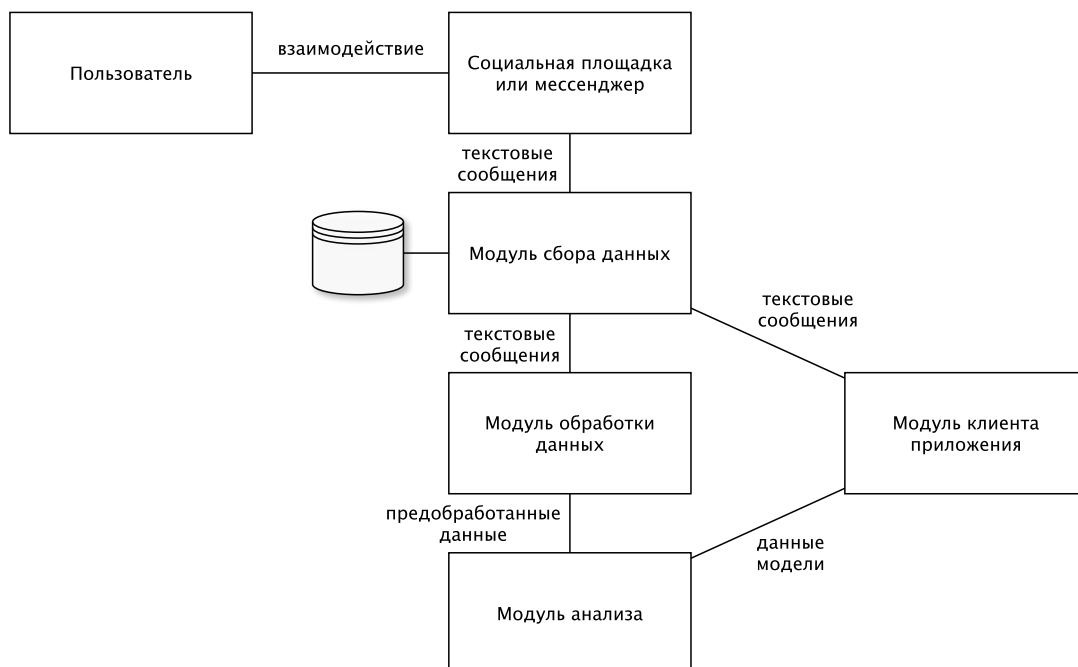


Рисунок 2.5 – Схема программного обеспечения.

2.4 Методы машинного обучения

В качестве задействованных в методе алгоритмов рассматриваются: градиентный бустинг (схема алгоритма представлена на рисунке 2.6), метод случайного леса (схема алгоритма представлена на рисунке 2.7), метод опорных векторов (схема алгоритма представлена на рисунке 2.8), метод К-ближайших соседей (схема алгоритма представлена на рисунке 2.9), логистическая регрессия (схема алгоритма представлена на рисунке 2.10) и перцептрон (схема алгоритма представлена на рисунке 2.11). Данный список обусловлен потребностью в целях исследования охватить широкий спектр методов машинного обучения для дальнейшего изучения возможности и потребности в создании ансамблевых моделей для решения поставленной задачи.

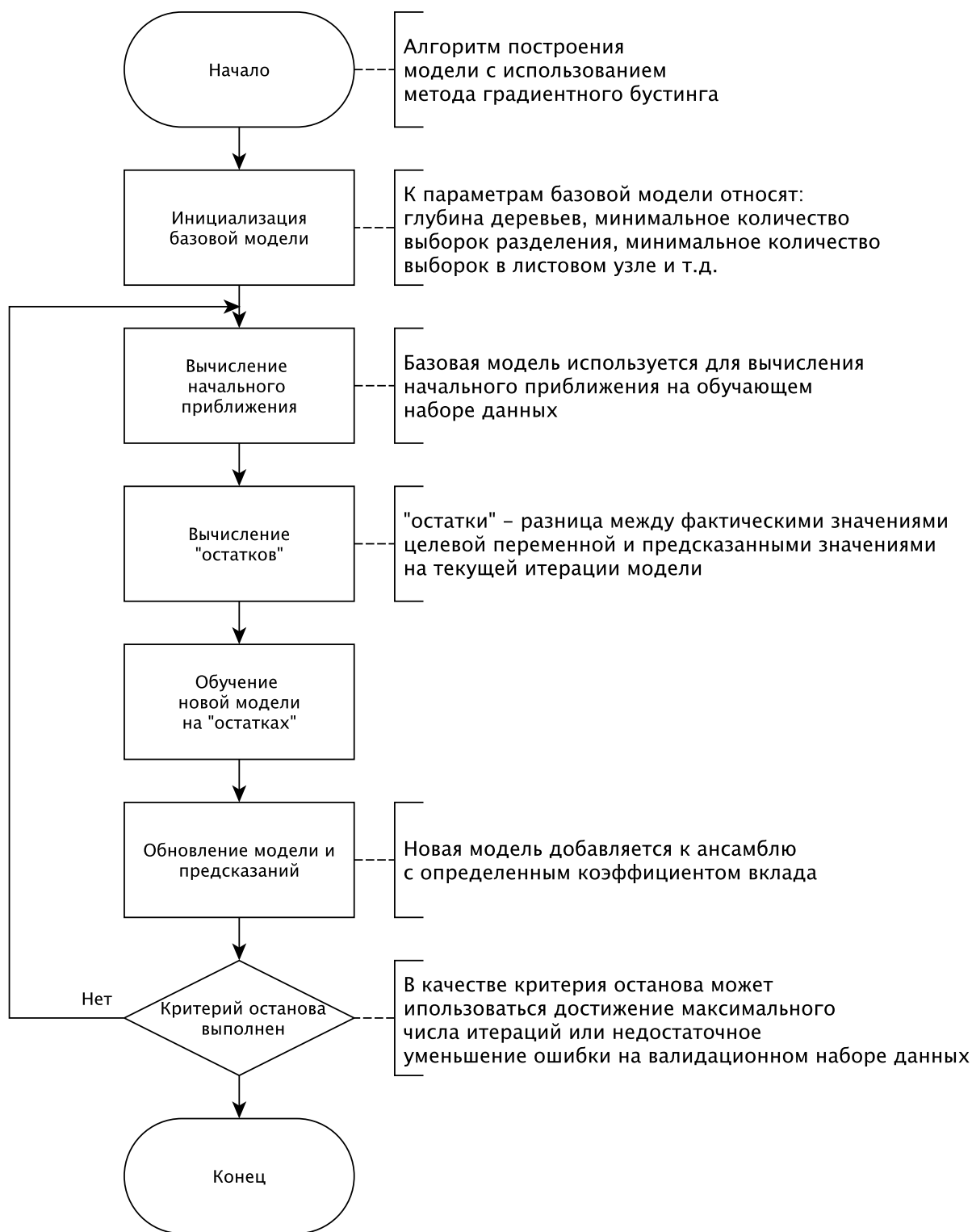


Рисунок 2.6 – Схема алгоритма работы градиентного бустинга.

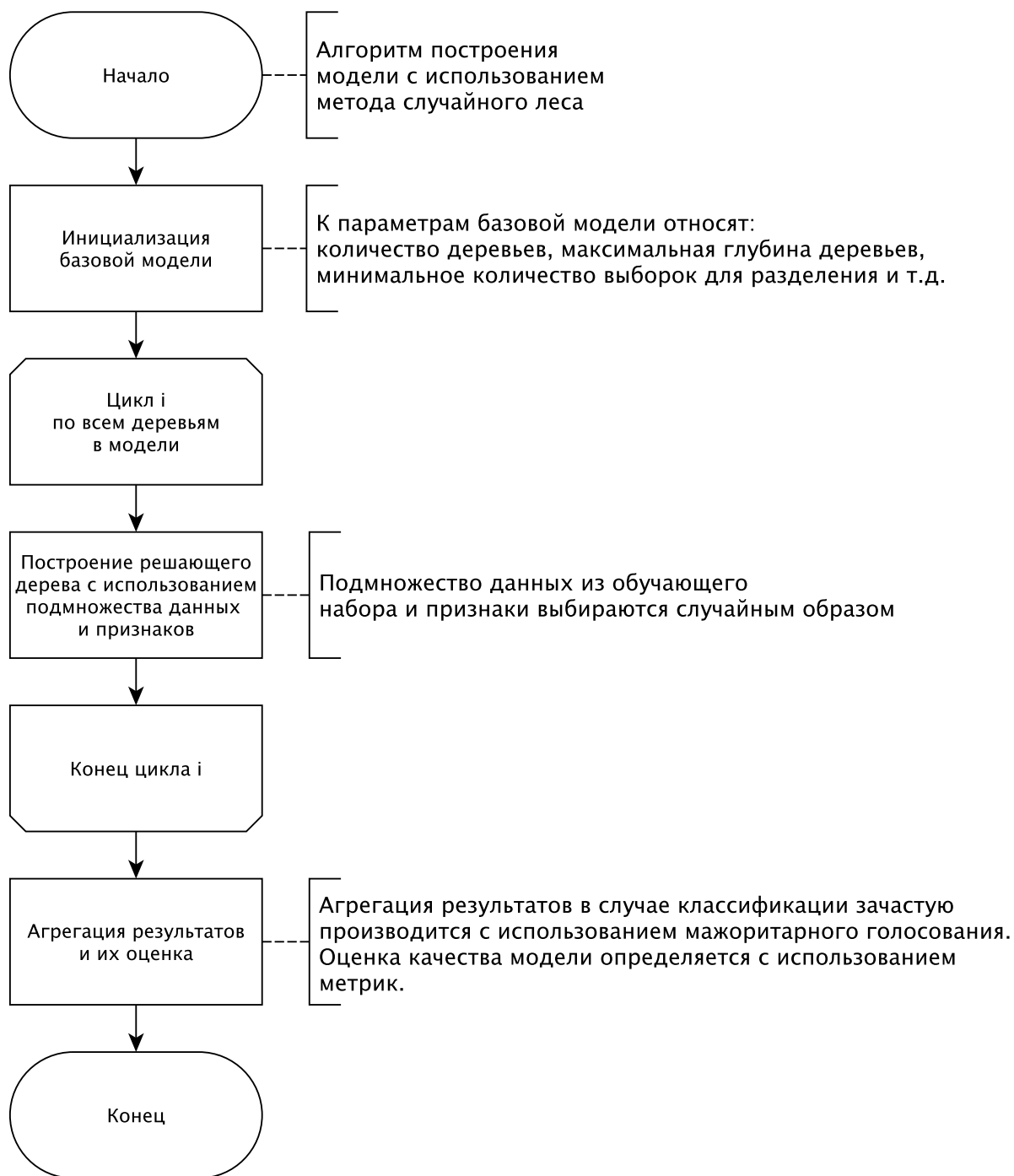


Рисунок 2.7 – Схема алгоритма работы метода случайного леса.

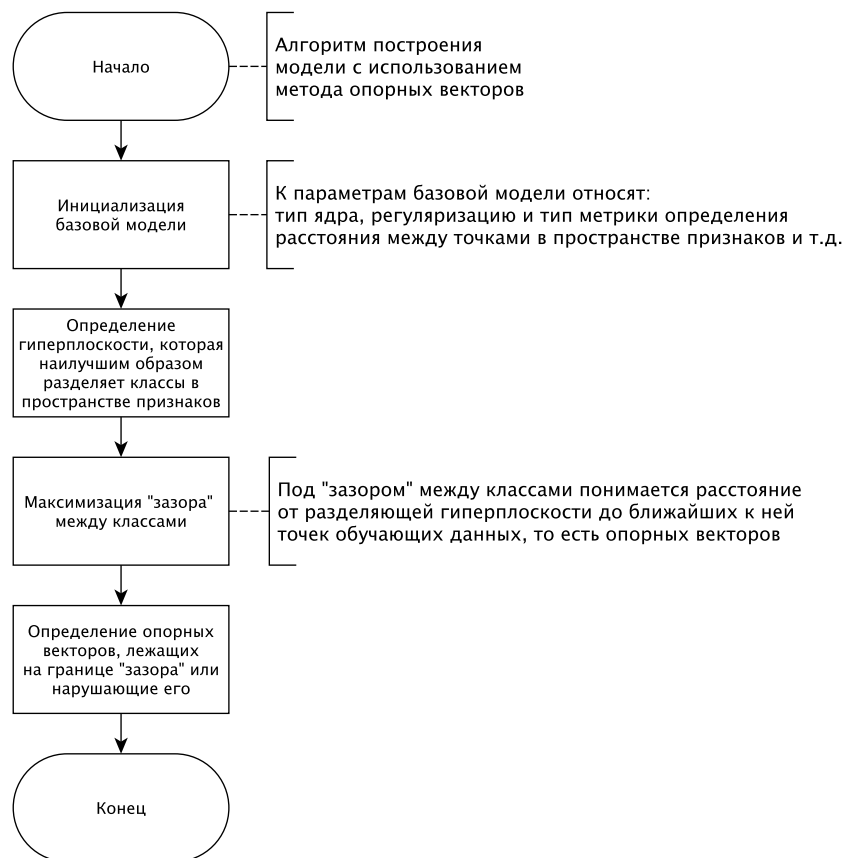


Рисунок 2.8 – Схема алгоритма работы метода опорных векторов.

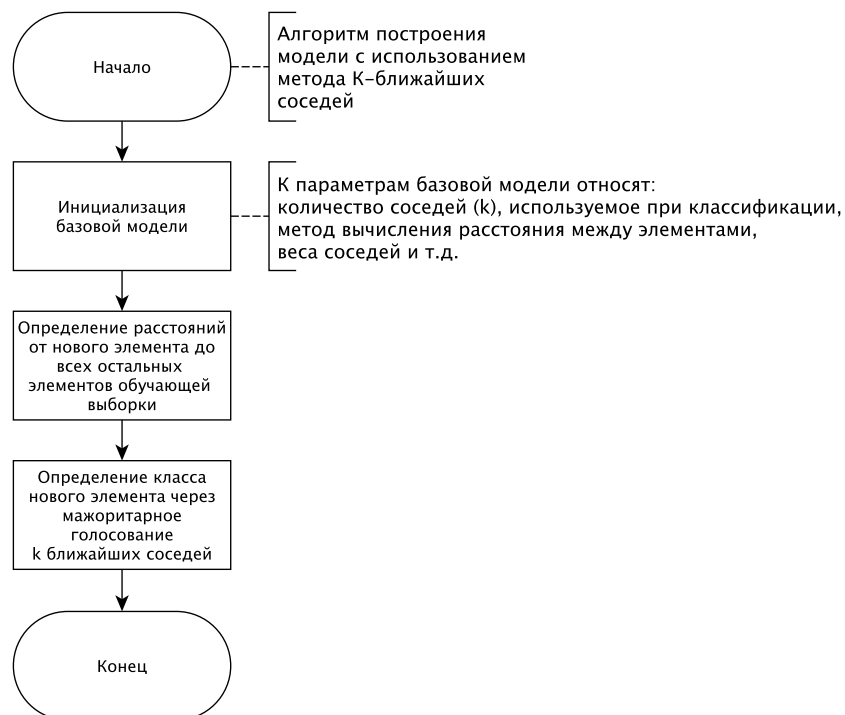


Рисунок 2.9 – Схема алгоритма работы метода K-ближайших соседей.



Рисунок 2.10 – Схема алгоритма логистической регрессии.

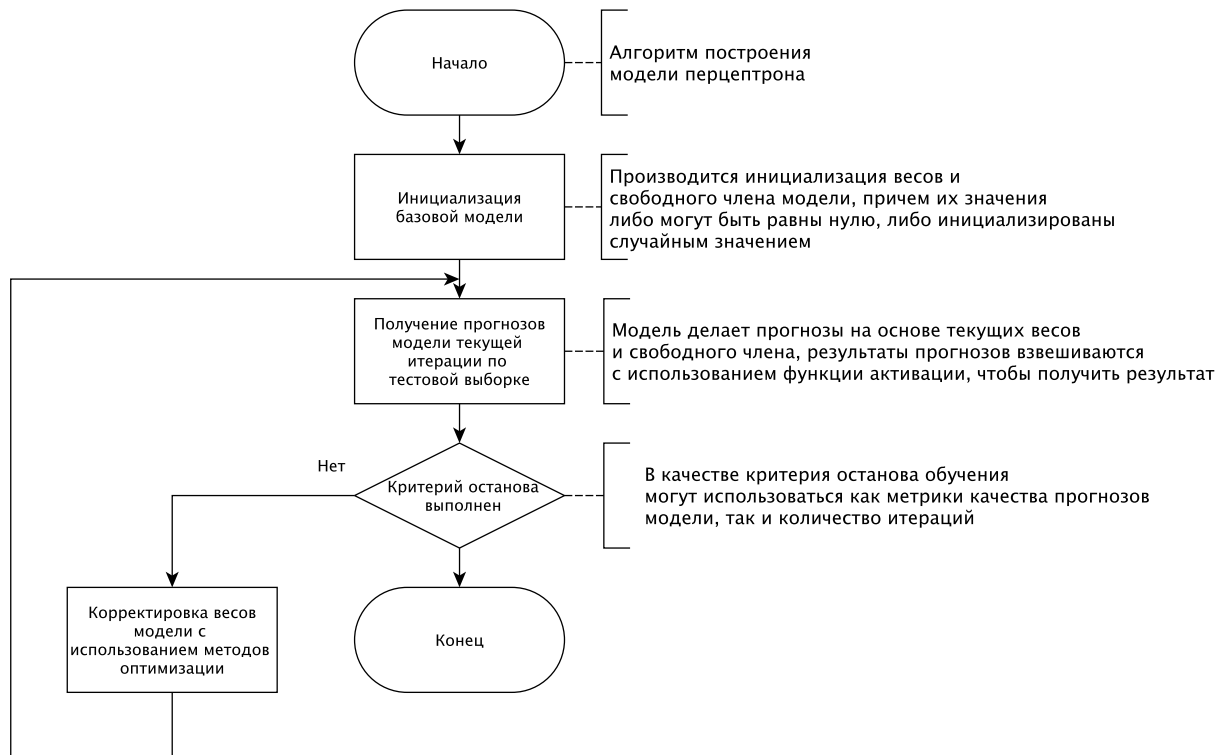


Рисунок 2.11 – Схема алгоритма работы метода с использованием перцептрона.

2.5 Алгоритмы векторизации текстовых сообщений

Классификация текстовых сообщений может производиться с использованием выделения ключевых слов, либо с использованием векторного представления данных. Первый подход является наиболее простым, но к его недостаткам относят необходимость участия экспертов в создании словарей для описания каждого классифицируемого класса сообщений, кроме того для получения приемлемых результатов требуется, чтобы классы сообщений мало пересекались по словарному набору [11]. Второй подход в классификации является более сложным, в нём могут использоваться различные модели векторизации для дальнейшей обработки текстовой информации. В качестве задействованной здесь может выступать модель текста “мешок слов” [12] с его расширением для частотных характеристик встречаемости слов в сообщениях TF-IDF [13]. Метод базируется на создании векторов сообщений с учетом весов встречаемости каждого слова, как в самом сообщении, так и во всех сообщениях выборки. Его использование при построении классификатора значительно повышает точность классификации в некотором наборе задач [11].

Кроме того, в качестве способа представления могут быть задействованы векторизации вида word embeddings. В качестве основного механизма они задействуют модель word2vec [14], которая представляет собой нейронную сеть, ставящую подаваемому на вход слову в соответствие выходной вектор заданной длины. Обучение нейронной сети производится таким образом, чтобы получить близкие в пространстве векторы для слов, встречающихся в одинаковых контекстах. Таким образом, слова близкие по значению, либо употребляющиеся совместно, будут иметь близкие векторы в пространстве. Существенным недостатком данного метода заключается в необходимости обучения нейронной сети на значительном корпусе текстов, что является ресурсозатратной задачей. Эта проблема нивелируется существованием предобученных моделей. Также к недостаткам относят невозможность представления слова, которого не было в обучающем корпусе. [11]

К векторизации word embeddings также относят и проект ELMo [15]. Данная модель обладает преимуществами word2vec и дополняет их возможностью формирования одного вектора на все сообщения, а также получения вектора для неизвестного слова путем его разложения на отдельные слоги и буквы. В основе метода лежит многослойная нейронная сеть, которая на первом

слое получает векторы для букв, а затем слои для работы со словами в составе всего сообщения. В качестве вектора используются веса последнего слоя сети. Для данного подхода также существуют предобученные модели для многих языков, что значительно упрощает его использование. [11]

BERT [16] – языковая модель, применяющаяся для анализа текста. Ее применение для построения классификатора требует обучения модели с кодированием отдельным слоем нейронной сети, что не дает возможность использовать данную технологию только для векторизации. В отличие от прежних классических языковых моделей, BERT обучает контекстно-зависимые представления, причем он учитывает двусторонний контекст, что помогает модели лучше понимать смысл многозначных слов. [11]

Для векторизации собранных текстовых сообщений будут задействованы алгоритм “мешок слов” и языковая модель BERT. Схемы работы выбранных алгоритмов представлены на рисунках 2.12 и 2.13 соответственно. Алгоритм TF-IDF не входит в рассмотрение в силу того, что он является модернизацией алгоритма “мешок слов”. Модели word2vec и ELMo также не попадают в рассмотрение в связи с тем, что данная работа первично имеет цель определить наиболее подходящий для решения задачи алгоритм машинного обучения.



Рисунок 2.12 – Схема работы алгоритма “мешок слов”.

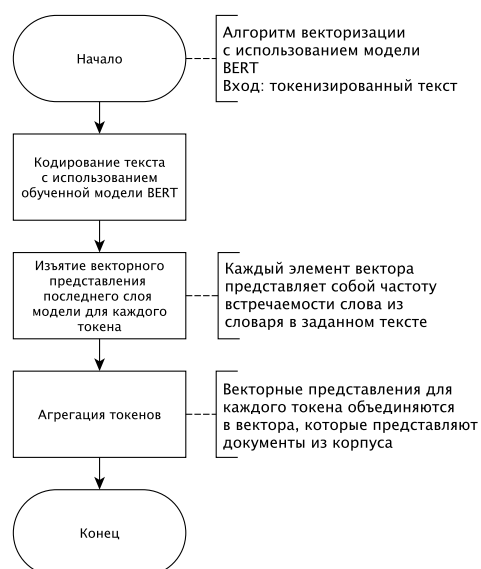


Рисунок 2.13 – Схема алгоритма векторизации с использованием модели BERT.

Вывод

Был описан метод распознавания суицидальных паттернов поведения человека по текстовым сообщениям, а также формат и метод сбора задействованных в нем данных. Определено, что в качестве средства сбора данных будет использоваться бот в мессенджере Telegram. Рассмотрены доступные средства реализации ботов в выбранном мессенджере.

Приведена диаграмма вариантов использования. Для системы было определено три действующих лица: пользователь, рекомендатор и анализатор. Приведена IDEF0 диаграмма, декомпозирована главная задача метода – распознавание суицидального сообщения. Диаграмма “сущность-связь” в нотации Чена позволила на абстрактном уровне описать систему распознавания.

Был определен перечень задействованных методов машинного обучения, который включил в себя: градиентный бустинг, метод случайного леса, метод опорных векторов, метод К-ближайших соседей, логистическая регрессия и перцептрон. В качестве методов векторизации выбраны: алгоритм “мешок слов” и языковая модель BERT. Для каждого выбранного алгоритма приведена схема его работы.

3. План исследований

3.1 Условия исследований

Исследование проводилось на персональном компьютере со следующими характеристиками:

- процессор Apple M1 Pro,
- операционная система macOS Ventura 13.5.2 (22G91),
- 32 Гб оперативной памяти.

Для определения гиперпараметров каждой модели применялся метод поиска по сетке с опорой на значение F1-меры. Разбиение данных на выборки производилось на 4 части, 1 из которых используется в качестве тестовой, результаты приводятся для каждого разбиения.

3.2 Исследование применимости моделей машинного обучения в методе распознавания суицидальных паттернов поведения человека по текстовым сообщениям

3.2.1 Градиентный бустинг

В данном подразделе представлены параметры модели, матрицы ошибок и оценки классификатора для задействованных методов векторизации.

3.2.2 Метод случайного леса

В данном подразделе представлены параметры модели, матрицы ошибок и оценки классификатора для задействованных методов векторизации.

3.2.3 Метод опорных векторов

В данном подразделе представлены параметры модели, матрицы ошибок и оценки классификатора для задействованных методов векторизации.

3.2.4 Метод К-ближайших соседей

В данном подразделе представлены параметры модели, матрицы ошибок и оценки классификатора для задействованных методов векторизации.

3.2.5 Логистическая регрессия

В данном подразделе быть представлены параметры модели, матрицы ошибок и оценки классификатора для задействованных методов векторизации.

3.2.6 Перцептрон

В данном подразделе представлены параметры модели, матрицы ошибок и оценки классификатора для задействованных методов векторизации.

Вывод

В данном подразделе представлен анализ результатов, сравнение матриц ошибок, сравнительная таблица средних значений оценок классификаторов. Определяется, какой метод имеет самые высокие оценки.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

Список литературы

1. Федеральный закон от 27.07.2006 г. № 152-ФЗ “О персональных данных” // СПС “КонсультантПлюс”.
2. Python-telegram-bot official page [Электронный ресурс]. Режим доступа: <https://python-telegram-bot.org/> (дата обращения 10.12.2023).
3. Telebot documentation [Электронный ресурс]. Режим доступа: <https://pytba.readthedocs.io/en/latest/> (дата обращения 10.12.2023).
4. Официальная страница модуля node-telegram-bot-api [Электронный ресурс]. Режим доступа: <https://www.npmjs.com/package/node-telegram-bot-api> (дата обращения 10.12.2023).
5. Репозиторий проекта kotlin-telegram-bot [Электронный ресурс]. Режим доступа: <https://github.com/kotlin-telegram-bot/kotlin-telegram-bot> (дата обращения 10.12.2023).
6. Python official page [Электронный ресурс]. Режим доступа: <https://www.python.org/> (дата обращения 10.05.2023).
7. mdn web docs: JavaScript [Электронный ресурс]. Режим доступа: <https://developer.mozilla.org/en-US/docs/Web/JavaScript> (дата обращения 10.12.2023).
8. Node.js documentation [Электронный ресурс]. Режим доступа: <https://nodejs.org/docs/latest/api/> (дата обращения 10.12.2023).
9. Kotlin language specification [Электронный ресурс]. Режим доступа: <https://kotlinlang.org/spec/introduction.html> (дата обращения 09.10.2020).

10. Java Virtual Machine SE8 Specification [Электронный ресурс]. Режим доступа: <https://docs.oracle.com/javase/specs/jvms/se8/html/> (дата обращения 08.09.2023).
11. Спивак А.И. Лапшин С.В. Лебедев И.С. Классификация коротких сообщений с использованием векторизации на основе ELMo // Известия ТулГУ. Технические науки. 2019. № 10.
12. Zhang Y. Jin R. Zhou Z.H. Understanding bag-of-words model: a statistical framework // International Journal of Machine Learning and Cybernetics. 2010. T. 1.
13. K.S. Jones. A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation. 2004. T. 60, № 1.
14. Tomas Mikolov Kai Ceen. Efficient Estimation of Word Representations in Vector Space. 2013.
15. Peters M.E. Neumann M. Iyyer M. Deep contextualized word representations // Proceedings of the 2018 Conference of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies. 2018. T. 1.
16. Cornell University arxiv [Электронный ресурс]. Режим доступа: <https://arxiv.org/abs/1810.04805v2> (дата обращения 10.05.2023).