МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

# АЛГОРИТМ И ПРОГРАММА ДЛЯ ФОРМАЛИЗАЦИИ ТЕКСТОВ ИНСТРУМЕНТАЛЬНЫХ МЕДИЦИНСКИХ ОБСЛЕДОВАНИЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Сидорова Сергея Александровича

Научный руководитель:
проф. кафедры информатики и
программирования, д.т.н. _____ А.С.Фалькович

подпись, дата

Зав. кафедрой информатики и
программирования
к.ф.-м.н., доцент _____ М.В.Огнева

подпись, дата

Саратов 2023

# ВВЕДЕНИЕ

**Актуальность темы.** В современном мире, где постоянно увеличивается количество получаемой информации, возникает потребность извлечения из единицы некоторой ключевой части, которая наиболее точно описывает основную тему данного документа. Такую часть принято называть ключевой фразой.

Поэтому перед новостными порталами, системами поиска и другими ресурсами ставится задача решить эту потребность. Выделяют много видов информации, из которой можно извлечь ключевую часть. Например, статьи, книги, web-страницы и так далее. В данной работе рассмотрены методы извлечения ключевых фраз из медицинских документов, а именно из текстов инструментальных медицинских обследований.

В настоящее время изучено и реализовано много методов, которые достаточно хорошо справляются с извлечением ключевых фраз из текстов любого объема на английском языке. Однако, подобных методов очень мало для русскоязычных текстов.

**Цель бакалаврской работы** – разработать алгоритм и программу для формализации текстов инструментальных медицинских обследований.

Поставленная цель определила **следующие задачи**:

1. Проанализировать существующие работы, связанные с данной темой.
2. Изучить технологии NLP.
3. Изучить методы обработки текста.
4. Выбрать подходящие инструменты для обработки текста.
5. Изучить процесс создания и разметки набора данных с русскоязычными текстами.
6. Изучить методы классификации.
7. Обучить модель для классификации русскоязычных текстов.
8. Написать алгоритм для формализации текстов инструментальных медицинских обследований.

9. Написать программу для формализации текстов инструментальных медицинских обследований.

**Методологические основы формализации текстового описания** представлены в работах А. А. Миронова, Р. И. Каримова, Н. Ю. Азаренко, Е. И. Большаковой, И. В. Москалева, Э. Г. Григоряна.

**Теоретическая значимость** заключается в том, что данная работа представляет собой исследование и разработку алгоритма и программы для формализации текстов инструментальных медицинских обследований. Это позволяет улучшить процессы обработки и анализа медицинских данных, что в свою очередь способствует улучшению качества и повышению эффективности медицинских исследований.

**Практическая значимость** заключается в том, что алгоритм и программа могут быть использованы в качестве основы для дальнейших исследований в области компьютерной обработки и анализа медицинских данных.

**Структура и объём работы.**

Бакалаврская работа состоит из введения, 5 разделов, заключения, списка использованных источников и 3 приложений. Общий объем работы – 69 страниц, из них 47 страниц – основное содержание, включая 18 рисунков и 2 таблицы, цифровой носитель в качестве приложения, список использованных источников информации – 21 наименование.

## КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

**Первый раздел «Обзор источников»** посвящен анализу существующих работ, связанных с извлечением ключевой части информации из текстовых корпусов. Было выявлено четыре типа методов отбора ключевых фраз: простые статистические методы, лингвистические методы, методы, основанные на машинном обучении, и их комбинации. У каждого автора свой подход к решению задач обработки текста, что приводит к отсутствию единого мнения по созданию алгоритмов и методов

извлечения ключевых фраз и анализа текстовых корпусов. Также, большинство работ предоставляют исследования ключевых фраз из английских текстов. Исследования, которые связаны с созданием программы для извлечения ключевых фраз на русском языке, в большинстве случаев не имеют открытого доступа.

**Второй раздел «Выбор инструментов»** посвящен выбору инструментов для обработки текста, создания модели-классификатора, программы для формализации и оценки качества обучения. Методы обработки естественного языка (NLP) являются важным инструментом для обработки текста, так как они позволяют анализировать естественный язык и извлекать информацию из текстовых данных. Язык программирования Python широко используется в NLP и машинном обучении. С его помощью был создан классификатор для формализации текстовых корпусов. С помощью технологии Windows Forms был создан графический интерфейс программы. Были рассмотрены способы обработки текста:

- векторное представление слов;
- токенизация;
- лемматизация;
- стемминг;
- стоп-слова.

Обработка текста происходит перед использованием методов машинного обучения. Были рассмотрены различные методы оценки качества обучения, включая кросс-валидацию и метрики оценки, такие как точность, полнота и F-мера.

**Третий раздел «Создание размеченного датасета»** посвящен созданию размеченного набора данных для обучения классификатора. В данной работе использовались два датасета: с размеченными данными и не размеченными данными. Оба датасета содержат медицинские тексты, содержащие в себе описание травматических повреждений. Была проведена очистка датасетов от лишних символов. К лишним символам относятся:

пунктуация, специальные символы и повторяющиеся пробелы. Был описан процесс создания обучающего и тестового набора данных. Обучающий датасет представляет из себя таблицу, имеющую ключевую часть информации, а также столбцы-классы, в которых она должна храниться. Тренировочным набором данных является набор текстовых корпусов, в которые содержат описание травматических повреждений.

**Четвертый раздел «Обучение модели классификации»** посвящен постановке задачи классификации. В данной работе использовалась мультиклассификация экстремально коротких текстов. Это задача классификации текстов, которые состоят из очень небольшого количества слов или символов. Для подготовки к решению использовать методы обработки текста, которые были описаны во втором разделе, а также специальные методы векторизации текста TF-IDF и CountVectorizer. Tf-IDF присваивает вес каждому слову в документе, основываясь на частоте его вхождения в документ и редкости его вхождения во всей коллекции документов. CountVectorizer преобразует каждый текстовый документ в вектор, который содержит количество вхождений каждого слова в документе. Другими словами, он создает "мешок слов" (bag of words), где каждый словообразующий элемент рассматривается как отдельный признак. Каждый элемент вектора представляет собой количество вхождений соответствующего слова в документе. Для обучения модели классификации было принято решение сравнить классификацию с помощью классических и ансамблевых методов машинного обучения. Представителем классических методов являлась логистическая регрессия. это метод классификации в машинном обучении, который используется для предсказания категориальных (дискретных) значений на основе набора непрерывных признаков. Он относится к методам обучения с учителем, где каждый объект в обучающей выборке имеет метку класса. Представителем ансамблевых методов являлся CatBoost Classifier. CatBoost – это алгоритм градиентного бустинга, который используется в машинном обучении для решения задач

классификации и регрессии. Он был разработан компанией Yandex и отличается высокой скоростью работы и хорошей точностью предсказания. Основная идея CatBoost заключается в том, что он учитывает категориальные признаки и использует градиентный бустинг для построения модели.

**Пятый раздел «Практическая часть»** посвящен созданию алгоритма и программы для формализации текстов инструментальных медицинских обследований. Был проведен анализ исходных данных и результат формализации, который нужно получить. Был показан процесс нормализации текста с помощью библиотеки Pymorphy2. Произведено разбиение данных на обучающую и тестовую выборки с помощью модуля sklearn.model_selection из библиотеки sklearn. Описан процесс обучения двух классификаторов с помощью логистической регрессии и CatBoost Classifier. Для обучения модели логистической регрессии были установлены следующие параметры:

- C – обратная сила регуляризации. Принимает значение по умолчанию равное 1;
- solver – алгоритм оптимизации параметров модели. Принимает значение «saga», который оптимизирует параметры для задачи многоклассовой классификации;
- multi_class – стратегия обучения для многоклассовой задачи. Указывает, что модель обучается с использованием мультиномиальной логистической регрессии, которая пытается минимизировать мультиномиальную потерю, а не бинарную потерю.

Для настройки CayBoost Classifier были выбраны следующие параметры, чтобы добиться лучшей производительности:

- task_type – задает тип задачи, который будет решаться на процессоре (CPU);
- iterations – задает количество итераций градиентного бустинга, то есть количество деревьев, которые будут построены. Чем больше iterations, тем более точная будет модель, но это может

привести к переобучению. В данном случае, выбрано достаточно большое число итераций, чтобы модель могла выучить более сложные паттерны в данных;

- eval_metric – задает метрику, которая будет использоваться для оценки качества модели во время обучения. TotalF1 – это метрика, которая измеряет среднее значение F1-меры по всем классам;
- od_type – задает стратегию, которая будет использоваться для остановки обучения, если качество модели не улучшается. В данном случае, используется стратегия остановки по количеству итераций (Iter). Это означает, что обучение остановится после заданного количества итераций;
- od_wait – задает количество итераций, которые необходимо преодолеть, прежде чем остановить обучение, если качество модели не улучшается. В данном случае ожидается до 500 итераций, прежде чем остановить обучение.

После оценки результатов работы классификаторов с помощью метрики F1-score было выявлено, что точность у CatBoost лучше, чем у логистической регрессии. Также, был проведен сравнительный анализ двух классификаторов, чтобы выявить причины улучшения точности. Был показан процесс получения результатов классификации, которые представляют собой массивы с метками классов, и их записи в файл для дальнейшего использования программой для формализации. Был описан алгоритм формализации текстового описания, который состоит из следующих этапов:

1. Считать корпус текстов.
2. Отправить корпус текстов в модель.
3. Процесс классификации.
4. Получение результатов классификации.
5. Формализация текстового описания.
6. Создание новой таблицы с результатами формализации.

Программа была написана на языке C# с использованием Windows Forms. Также, были использованы библиотеки ExcelDataReader и Microsoft.Office.Interop.Excel для чтения и создания новых Excel таблиц.

Windows Forms – это библиотека классов .NET Framework, которая предоставляет набор элементов управления и функциональность для создания графического интерфейса пользователя. Данная технология позволит создать приложение, которое будет понятно обычному пользователю, а также предоставит удобство в работе. Графический интерфейс имеет две кнопки – «открыть» и «сохранить», а также поле DataGridView, для вывода и работы с табличной структурой.

Пользователь может выбрать файл с корпусом текстов, нажав на кнопку «Открыть». Далее программа использует библиотеку ExcelDataReader для считывания содержимого Excel таблиц и передачи результатов в DataGridView, где создаются новые столбцы для формализации.

После чтения файла и создания новых столбцов, исходные корпусы текстов передаются в модель для классификации. Программа также обрабатывает исходные корпусы текстов, чтобы очистить их от лишних знаков пунктуации и пробелов. После процесса классификации, программа получает файл с массивами для каждого текстового корпуса и переносит ключевую информацию в соответствующие столбцы. Затем происходит поиск логических и числовых значений с помощью регулярных выражений. Результат работы программы отображается пользователю в DataGridView.

Пользователь может сохранить табличную структуру из DataGridView в новый .xlsx файл, нажав на кнопку "Сохранить". При нажатии на эту кнопку, программа генерирует новую таблицу Excel, содержащую исходный корпус текстов и полученный результат формализации. Пользователь также может вносить изменения в любой участок табличной структуры, кликнув на интересующий элемент и внести необходимые изменения. Если пользователь сохранит файл после внесения изменений, то они отобразятся в новой Excel таблице.

# ЗАКЛЮЧЕНИЕ

В ходе работы был произведен анализ существующих работ, связанных с данной темой. Благодаря этому были получены необходимые знания для решения задач, связанных с обработкой текстов. Были изучены технологии NLP и методы предварительной обработки текста, с помощью которых можно подготовить корпусы текстов на русском языке для дальнейшей работы с ними. Были изучен процесс создания обучающих и тестовых датасетов. Были изучены методы классификации с помощью логистической регресси и CatBoost Classifier. Был проведен сравнительный анализ между двумя классификаторами, с целью получить лучший результат на практике. Была обучена модель для классификации корпусов текста, связанных инструментальными медицинскими обследованиями. Были созданы алгоритм и программа для формализации инструментальных медицинских обследований.

Данная программа может помочь сотрудникам медицинских учреждений оптимизировать процесс сбора и анализа данных. С её помощью не придется тратить большое количество времени на заполнение или извлечение информации с бумажных носителей.

**Отдельные части бакалаврской работы были представлены на конференции** «Фундаментальная и прикладная медицина», Саратов, 29–30 ноября 2022 года и опубликованы в сборнике тезисов:

Сидоров С. А. Алгоритм и программа для преобразования текстовых описаний диагнозов в таблицу / С. А. Сидоров, А. С. Фалькович // Фундаментальная и прикладная медицина. Материалы Всероссийской конференции молодых ученых. Саратов, 2022. С. 141.

**Основные источники информации:**

1. Миронов А. А. Использование Join-layer neural networks для решения задачи извлечения ключевых фраз из постов социальной сети

«Твиттер» / А. А. Миронов, Я. П. Горожанкин, А. О. Иванов, С. О. Целикова, Я. В. Ахремчик // Молодой ученый. 2019. № 26. С. 37-41.

2. Каримов Р. И. NLP - обработка естественного языка. лингвистический метод обработки текста / Р. И. Каримов // Мавлютовские чтения. Уфа : изд-во УГАТУ, 2021. № 3. С. 250-254.

3. Азаренко Н. Ю. NLP в задачах анализа научного текста / Н. Ю. Азаренко, О. Д. Казаков // Глобальная нестабильность и цифровые технологии: реалии XXI века. М.: РУДН, 2020. № 2. С. 273-276.

4. Большакова Е. И. Автоматическая обработка текстов на естественном языке и анализ данных (Методы хранения словарей, Инструментальные системы для извлечения информации, Извлечение терминологической информации) / Е. И. Большакова, К. В. Воронцов, Н. Э. Ефремова, Э. С. Клышинский, Н. В. Лукашевич, А. С. Сапин // Высокопроизводительные вычислительные системы и технологии. 2020. № 1. С. 144 – 150.

5. Москалев И. В. Автоматизация процесса извлечения структурированных данных из неструктурированных медицинских выписок с применением технологий интеллектуального анализа текстов / И. В. Москалев, О. С. Кротова, Л. А. Хворова // Высокопроизводительные вычислительные системы и технологии. 2020. № 1. С. 163 – 167.

6. Григорян Э. Г. Методы NLP для предобработки текстовых данных и выделения признаков / Э. Г. Григорян, М. Н. Паршин // Бизнес и общество. 2021. № 3. С. 213 – 224.