

Университет ИТМО. Факультет Программной инженерии и
компьютерной техники

**Лабораторная работа № 1 “Проектирование системы
интеллектуального анализа данных ” по дисциплине
“Интеллектуальный анализ данных”**

**Тема работы: “ Классификация депрессивного состояния на
основе текстовой информации ”**

Выполнили:

Рогаленко Никита Р42142
Коков Алексей Р42141
Шишкин Никита Р42143
Рустамов Акмал Р42142
Давыдов Иван Р42143
Авраменко Антон Р42141

Преподаватель:
Платонов Алексей Владимирович

Санкт-Петербург 2022

1. Формальная постановка решаемой задачи

Решаемая задача

Разрабатываемая система должна выявлять депрессивные состояния на основе текстовой информации, предоставляемой пользователем в виде текстовых сообщений или описания мировосприятия человека.

Разрабатываемая система может быть использована как вспомогательная система для частной психологической практики или же может быть использована спецслужбами для привлечения внимания органов к личностям, которые могут представлять опасность в будущем или нуждаться в психологической помощи. Система может быть использована для выявления признаков психологического неблагополучия для его своевременной диагностики и профилактики; снижения расходов на психологическую диагностику; получения доступа к группам населения, которые не обращаются за психологической помощью по разным причинам.

Решаемую задачу можно определить как задачу бинарной классификации текста: на вход подается текст, на выходе формируется класс текста - имеет депрессивный оттенок / не имеет.

При работе с системой предполагается использование текстов на русском языке. В дополнение к анализу сообщений предполагается также использование информации о половой принадлежности и возрасте автора. Данная информация будет использована для формирования и предоставления статистики частоты возникновения депрессивных настроений у разных групп населения.

Задача для краудсорсинга

Примером депрессивного сообщения/текста является: "Нет желания жить, все очень плохо. С 10 лет начал курить, пить, в 13 наркотики. 5 лет подряд наркотики + алкоголь+ легкие наркотики и так каждый день. С трудом закончил школу. Поступил в ПТУ которое было не далеко от точки где брал ""дела""(выбирал

по этому принципу). Раз так 5 откачивали...", – нужно классифицировать как депрессивное.

Следующий текст/сообщение не носит в себе депрессивных оттенков: “Как погода сегодня? Я смотрел в yandex тепло! Всегда бы так.”, – не следует классифицировать как депрессивное.

В качестве задачи для внешнего ассессора после ознакомления с предоставленными примерами сообщений предлагается классифицировать набор других текстовых сообщений на предмет того, являются они депрессивными или нет.

Критерии оценки качества системы

Для оценки качества разработанной системы будут использоваться стандартные метрики, используемые для оценки модели бинарного классификатора: precision, recall, accuracy, F-мера. Предполагается использование стандартной матрицы ошибок:

TP (сообщение предсказано как депрессивное и является таковым)	FP (сообщение предсказано как депрессивное, но не является таковым)
FN (сообщение предсказано как не депрессивное, но является таковым)	TN (сообщение предсказано как не депрессивное и не является таковым)

К итоговой модели предъявляются следующие критерии качества:

- 1) При оценке модели на тестовом датасете доля действительно депрессивных сообщений относительно всех сообщений, определенных как депрессивные (precision) должна составлять не менее 60%.
- 2) При оценке модели на тестовом датасете доля найденных депрессивных сообщений относительно всех депрессивных сообщений в выборке (recall) должна составлять не менее 60%.
- 3) При оценке модели на тестовом датасете доля верно классифицированных сообщений (accuracy) должна составлять не менее 60%.
- 4) При оценке модели на тестовом датасете F-мера должна составлять не менее 60%.

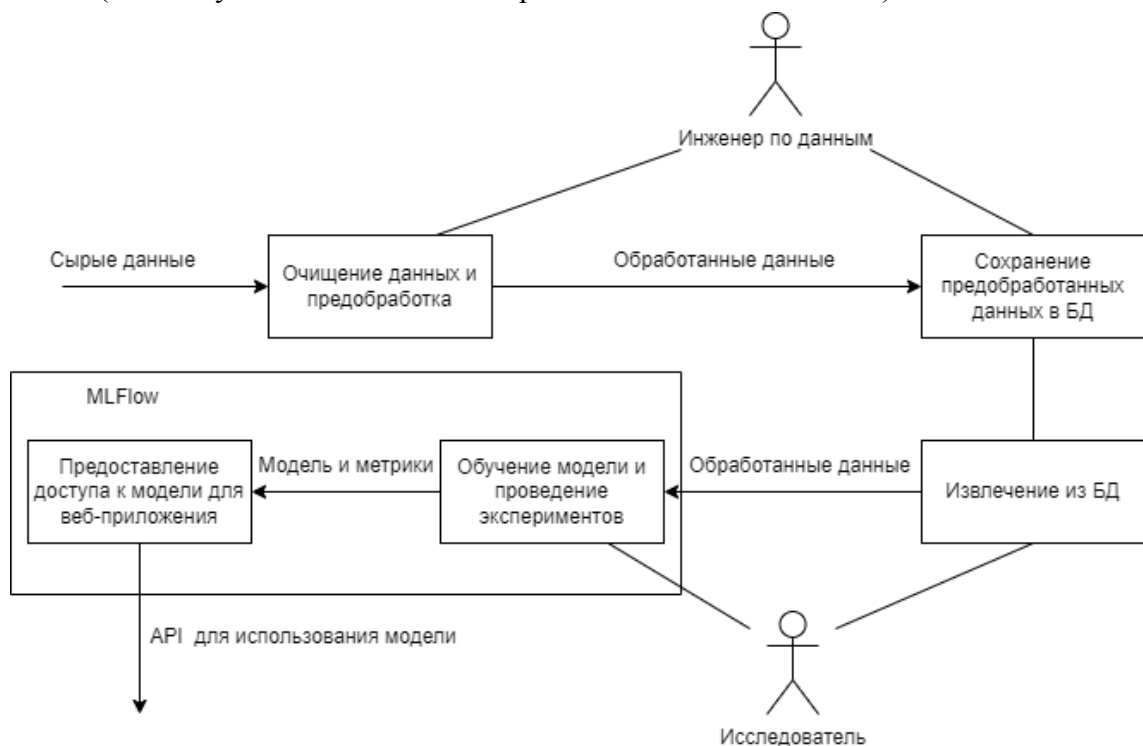
2. Проектирование систем хранения и предобработки данных

Для хранения предобработанных данных предполагается использование хранилища данных, позволяющего хранить в пригодном для обработки формате как сами текстовые сообщения, так и ассоциированную с ними информацию, такую как пол или возраст автора. В качестве базы данных предварительно решено использовать MongoDB. В качестве документа будет использоваться JSON, содержащий текст, пол, возраст в разных полях.

Для версирования экспериментов и полученных моделей, метрик предполагается использование MLFlow. Будет использоваться ML Flow с Tracking Server (<https://mlflow.org/docs/latest/tracking.html#id30>).

Пайплайн выстраивается подобным образом:

- 1) Инженер по данным готовит данные для системы. Происходит предобработка данных: очищение от сторонней информации, удаление стоп-слов, исправление опечаток и т.д.
- 2) Инженер по данным загружает предобработанные данные в базу данных.
- 3) Исследователь берет данные из базы (посредством подготовленного программного инструмента), производит эксперименты, обучает модель. Результаты экспериментов фиксируются MLFlow.
- 4) Удовлетворяющая модель используется backend-частью веб-приложения (используется mlflow serve с предоставлением REST API).




3. Проектирование системы, использующей модель

В качестве системы использующей модель предлагается выбрать Веб-приложение, которое будет взаимодействовать с API MLFlow.

Стек технологий

- Django 4.x.x
- HTML + CSS с использованием Django templates и сторонних JS-библиотек
- MongoDB

Прототипы пользовательских интерфейсов

Распознаватель депрессии


Введите текст для проверки

Type here

Дата рождения: 10/06/2022

Пол: ☒ Мужчина ☐ Женщина ☐ Другой

ПроверитьОтчистить

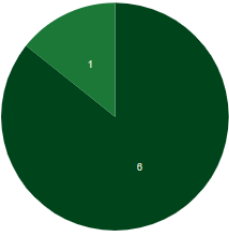


Депрессивные признаки не обнаружены

Статистика по возрасту:

■ Без депрессии

■ С расстройством



Категория	Количество
Без депрессии	6
С расстройством	1

Распознаватель депрессии

Введите текст для проверки

Type here

Дата рождения:

10/06/2022

Пол:

☒ Мужчина

☒ Женщина

☒ Другой

Проверить

Отчистить

X

Имеются депрессивные признаки

[Искать помощи](#)

Статистика по возрасту:

■ Без депрессии

■ С расстройством

1

8

Взаимодействие конечного пользователя с системой организуется посредством описанных веб-интерфейсов:

- 1) Пользователь вводит текстовое сообщение для проверки на наличие депрессивного оттенка;
- 2) Пользователь вводит дополнительную информацию - пол и возраст;
- 3) Пользователь отправляет форму системе;
- 4) Пользователь видит результат проверки текста на депрессивность и обновленную статистику.

Также возможен второй вариант:

- 1) Пользователь вводит только свой пол и возраст;
- 2) Система отображает ему статистику

Диаграммы деятельности при получении запросов от пользователя системы

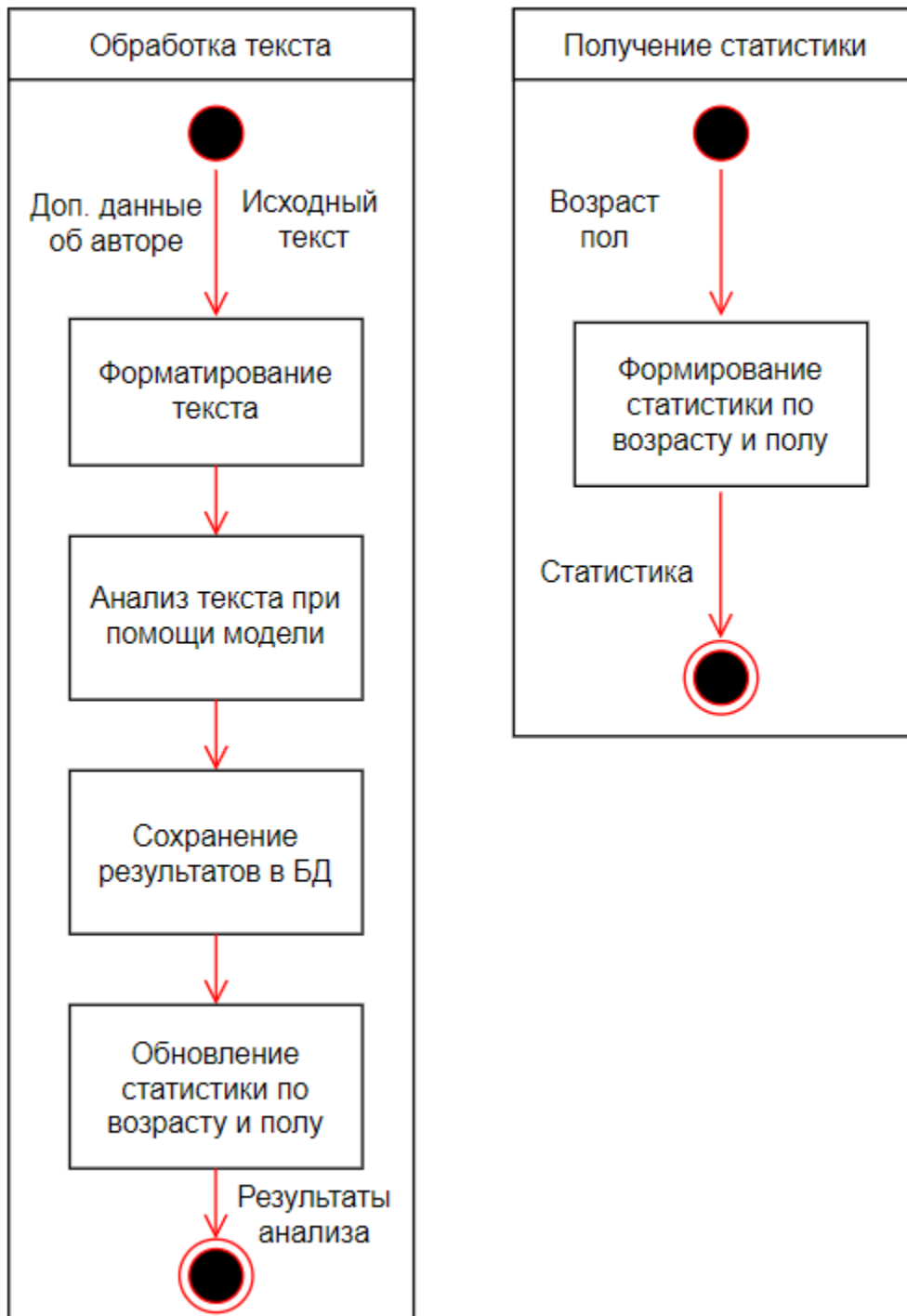
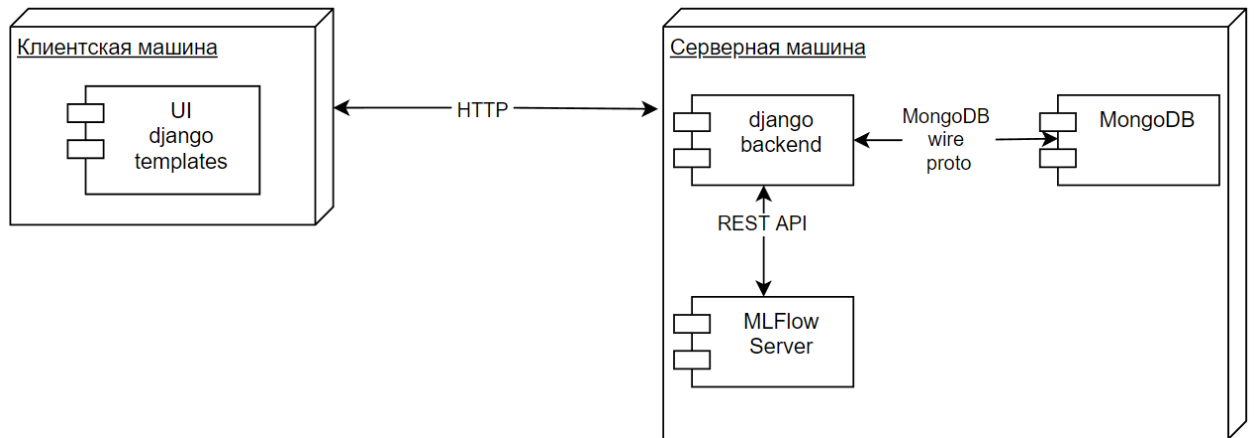


Диаграмма развертывания



4. Исследование вариантов решения задачи

1. <https://psy-journal.hse.ru/2020-17-1/359611339.html>

Данная статья имеет внушительную теоретическую базу исследования, состоящую из множества статей, подтверждающих связь между поведением пользователей в социальных сетях и наличием у них признаков депрессивных расстройств.

Отличительными особенностями могут являться так и кажущаяся логичной более высокая частота употребления слов с негативным оттенком (“устал”, “надоело”, “ненавижу”, “депрессия”), так и более неочевидные признаки: например, преобладание существительных в речи, при сниженном количестве глаголов и местоимений. В целом, люди с признаками депрессии отправляют сообщения реже, но как правило они содержат больше слов.

Поскольку в рамках данного исследования были использованы социальные сети (а именно “ВКонтакте”), то для классификации использовался не только психолингвистический анализ сообщений, но и другие следы активности пользователя: число друзей, подписок, групп, аудиозаписей, фотографий, видео, подарков, интересных страниц и т.д.

Подопытные были разбиты на группы в зависимости от результата прохождения опросника депрессии Бека. Выгрузка данных осуществлялась автоматически при помощи специализированного ПО и API системы

Среди методов машинного обучения для анализа данных были задействованы:

- Метод опорных векторов (SVM)
- Алгоритм случайного леса (Random forest)

В качестве признаков были использованы:

- Количественные (число друзей, подписок, групп, аудиозаписей, фотографий, видео и т.д.)
- Бинарные (факты наличия опциональной информации в профиле)
- Фиксированные ответы

В психолингвистические маркеры вошли:

- Среднее количество слов в предложении
- Среднее количество символов в слове
- Соотношение знаков препинания и количества слов
- Доля уникальных слов в лексике
- Средняя глубина синтаксического дерева
- Соотношение различных частей речи
- И другие, не упомянутые в статье

Итог:

Данное исследование только частично можно соотнести с нашей работой. Мы не располагаем доступом к данным кроме сообщений и возраста автора, поэтому их анализ попросту невозможен. С другой стороны, полезно обратить внимание на задействованные психолингвистические маркеры, поскольку данное исследование показывает эффективность их использования.

2. <http://proceedings.spiiras.nw.ru/index.php/sp/article/view/14930>

Признаки депрессии, получаемые из лексического содержания
<ul style="list-style-type: none">• Запинки в речи• Использование более коротких предложений• Частое употребление местоимения первого лица единственного числа, больший фокус на себе• Использование слов-абсолютов, негативно окрашенных слов, упоминание фармацевтического лечения• Редкое использование местоимения первого лица множественного числа• Частое использование лексики с аффективной семантикой• Употребление лексики протестного поведения

3. <http://www.nauteh-journal.ru/files/3d67cd8c-d9b6-47c8-b7b6-776026337ad5>

В данной статье приводится краткое описание ряда работ направленных на диагностирование психических заболеваний. Можно выделить следующие способы выявления депрессии при помощи методов машинного обучения:

- анализ электроэнцефалограммы
- распознавание по произвольной речи без контекста
- по звучанию голоса
- распознавание эмоций по мимике человека
- классификация по данным мониторинга физической активности
- несколько примеров моделей по выявлению депрессии у пользователей Вк и фейсбук по параметрам активности и психолингвистическим маркерам их текстов

4. <https://habr.com/ru/post/421775/>

Из данной статьи (а точнее - её резюме) можно вынести следующее:

- Практически гораздо большую значимость имеет возможность анализа произвольных текстов, а не ответов на определенные вопросы. Это даёт возможность отслеживать состояние исследуемых в режиме реального времени.
- Данная система не ищет паттерны в речи конкретного человека. Она ищет закономерности, указывающие на депрессию, а затем накладывает их на конкретного индивида.
- Для обучения нейросети использовалась техника под названием «моделирование последовательностей» (sequence modeling. Модель обучается на последовательностях текстовых и звуковых данных из вопросов и ответов от людей с депрессией и без неё. Постепенно она выявляет общие закономерности. В конечном счете, модель сама определяет, есть в речи признаки депрессии или нет.

Итог:

Данная статья ещё раз подтверждает наличие взаимосвязи между определенными маркерами в речи. Однако речь содержит кроме типичных для текстовых данных маркеров дополнительные уникальные (тон, интонация, тембр и т.д.), которые также оказывают влияние на итоговый результат. В нашей работе к такого рода данным опять же нет доступа, что снижает практическую ценность статьи. Однако в целом конкретно текстовая часть речи исследуется аналогично предыдущим описанным методам - при помощи различных типов маркеров.

5. Предварительный план тестирования

Предполагается разработка автоматизированных тестов с использованием тестового фреймворка pytest. Для e2e тестирования также дополнительно будет использован инструмент Selenium WebDriver. Для тестирования производительности под нагрузкой будет использован инструмент Apache JMeter.

В качестве набора тестовых данных будет использоваться разработанный тестовый датасет с входным набором данных, согласно которому будет определяться корректное поведение системы.

Тестирование будет производиться в несколько этапов:

1. Модульное тестирование

В случае предполагаемой системы тестируемой функциональностью будет, например, класс обработки текста, который содержит функциональность:

- проверки текста на предмет депрессивных оттенков;
- ведения статистики по полу и возрасту;

2. E2E (функциональное) тестирование

При e2e тестировании будут применяться тестовые сценарии, шаги которых будут соответствовать тем, которые будет осуществлять конечный пользователь.

В случае предполагаемой системы предварительно можно выделить следующий сценарий:

№	Описание шага	Ожидаемый результат
1	Зайти на главную страницу “Распознавателя депрессии”	Открыта главная страница Отображены все необходимые элементы: поле ввода текста, поле ввода пола, поле ввода даты рождения и кнопка отправки формы
2	Заполнить форму	Внесенные изменения отображаются на странице
3	Отправить форму	На странице отображается результат анализа текста и статистика по возрасту

3. Тестирование производительности

При тестировании производительности необходимо проверить соответствие нефункциональным требованиям. Примерами таких требований могут послужить:

- Передача на проверку больших текстов (>10k символов);

- Скорость обработки текстов под нагрузкой (запросы JMeter).