Т-АВИАС

«ЦИФРОПОЛ-2025»

Нейросетевые модели мониторинга социальных медиа для выявления деструктивного контента

Кузьмина К.А.
Инженер

Научно-практическая специальная конференция

Проблемы и задачи

Сложность отслеживания распространения деструктивного контента в социальных сетях и мессенджерах

Ручной анализ направленности и классификация вредоносных материалов требуют значительных временных и трудовых затрат

!

Мониторинг социальных медиа с применением нейросетевых моделей для выявления деструктивного контента

Актуальность

ОБЪЕМ АУДИТОРИИ

Число активных авторов (авторы постов, репостов и комментариев) в социальных медиа в России составило 74,9 млн по состоянию на октябрь 2024

ДОСТУПНОСТЬ И АНОНИМНОСТЬ

Анонимность в социальных сетях снижает чувство ответственности у преступников, облегчая совершение противоправных действий

СОЦИАЛЬНЫЕ ПОСЛЕДСТВИЯ

Усиление радикальных настроений, разжигание ненависти, дестабилизация политической обстановки

ИНСТРУМЕНТЫ КИБЕРПРЕСТУПНИКОВ

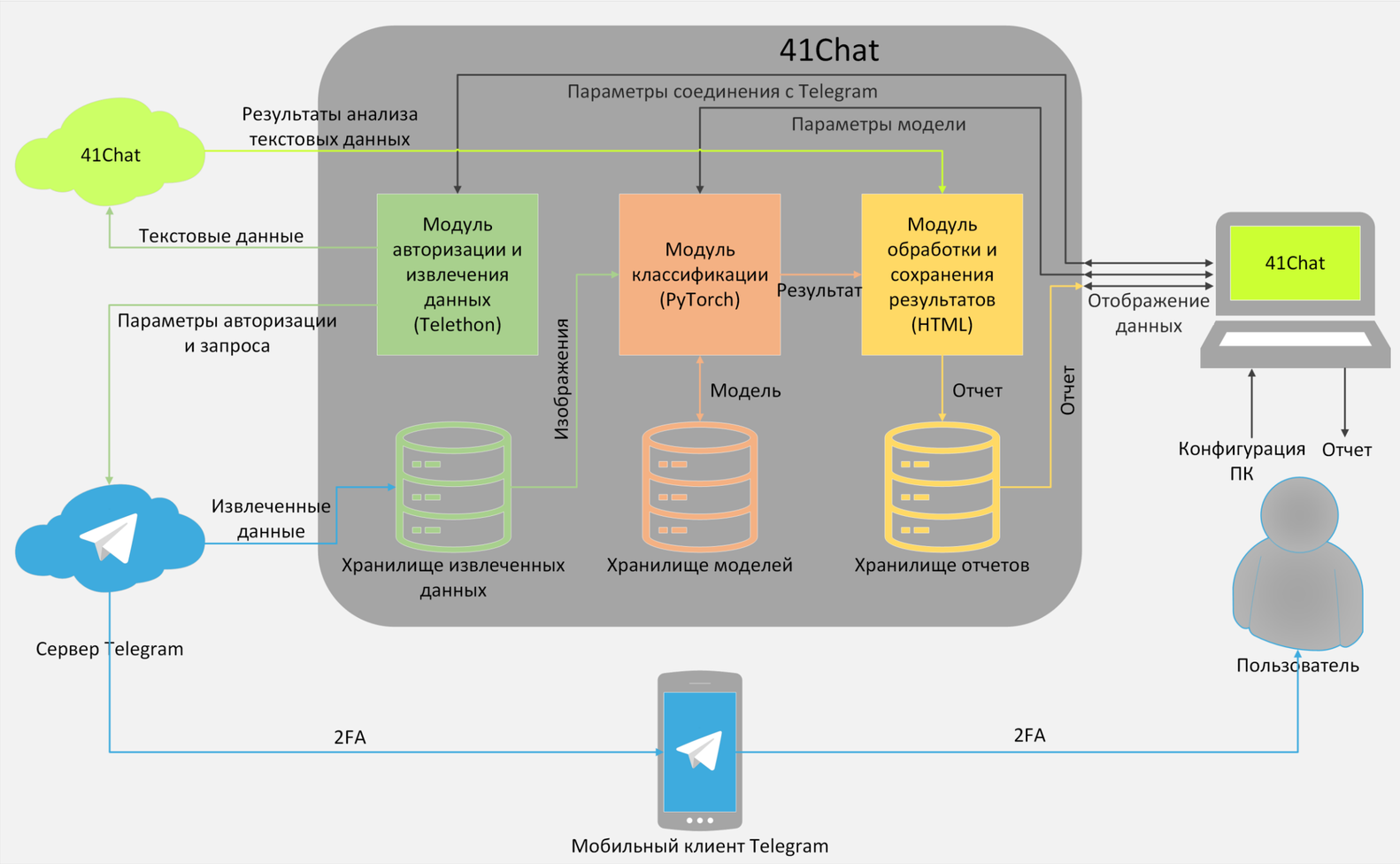
Распространение идеологии терроризма, распространение порнографических материалов, нарушение тайны переписки, незаконный оборот специальных технических средств и запрещенных веществ и многое другое

Применение

- Модерация контента в социальных медиа
- Мониторинг школьных чатов
- Родительский контроль
- Правоохранительные органы
- Репутационный менеджмент

Состав программного комплекса

- Модуль авторизации и извлечения данных
- Модуль классификации
- Модуль обработки и сохранения результатов
- GUI



Состав ПК

- **Модуль авторизации и извлечения данных**
- Модуль классификации
- Модуль обработки и сохранения результатов
- GUI

Состав ПК

- **Модуль авторизации и извлечения данных**
- Модуль классификации
- Модуль обработки и сохранения результатов
- GUI

Фотографии

- Относятся к визуальному контенту
- Обладают максимальной информативностью

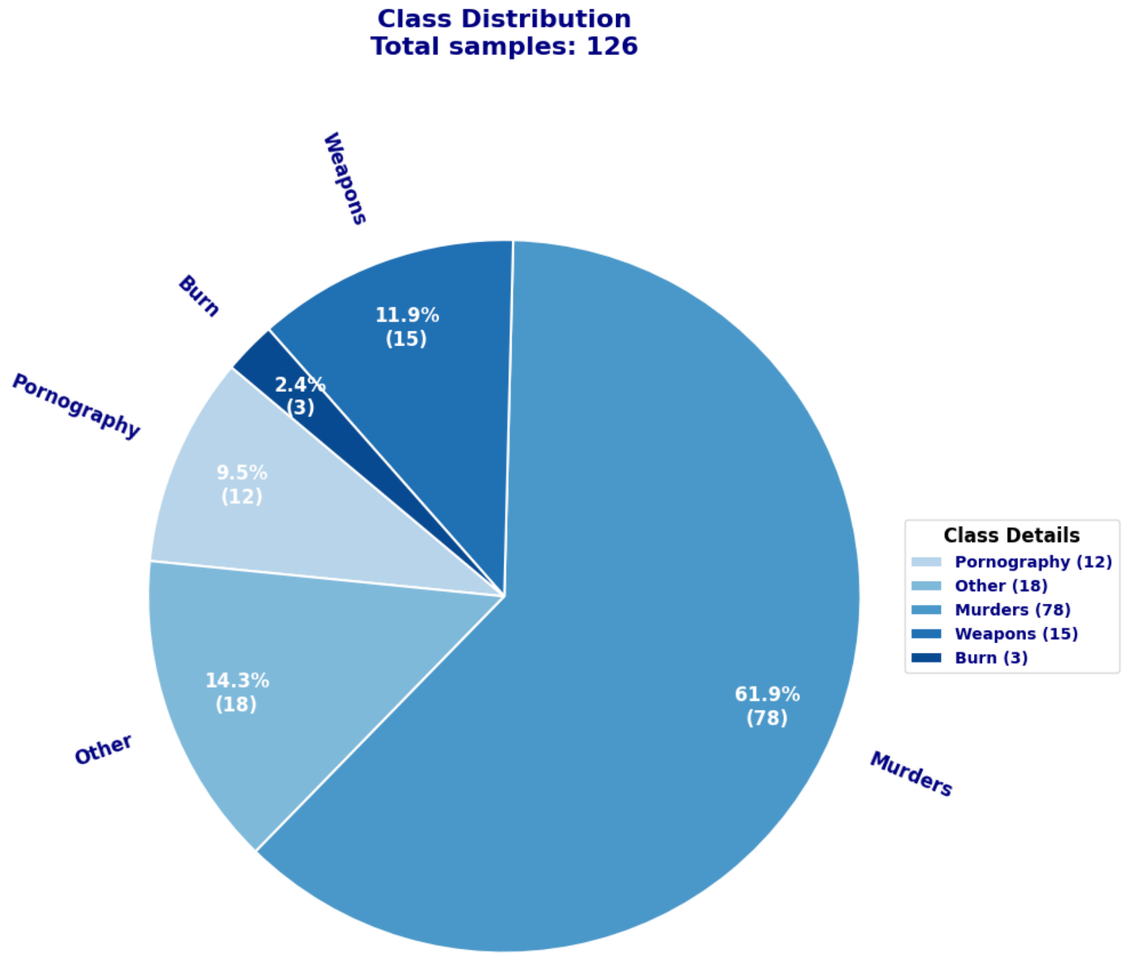
Состав ПК

- Модуль авторизации и извлечения данных
- **Модуль классификации**
- Модуль обработки и сохранения результатов
- GUI

Выделенные классы	Насильственные действия со смертельным исходом
	Контент, демонстрирующий поджоги или неконтролируемые пожары
	Контент сексуального характера
	Демонстрация средств вооружения
	Прочее

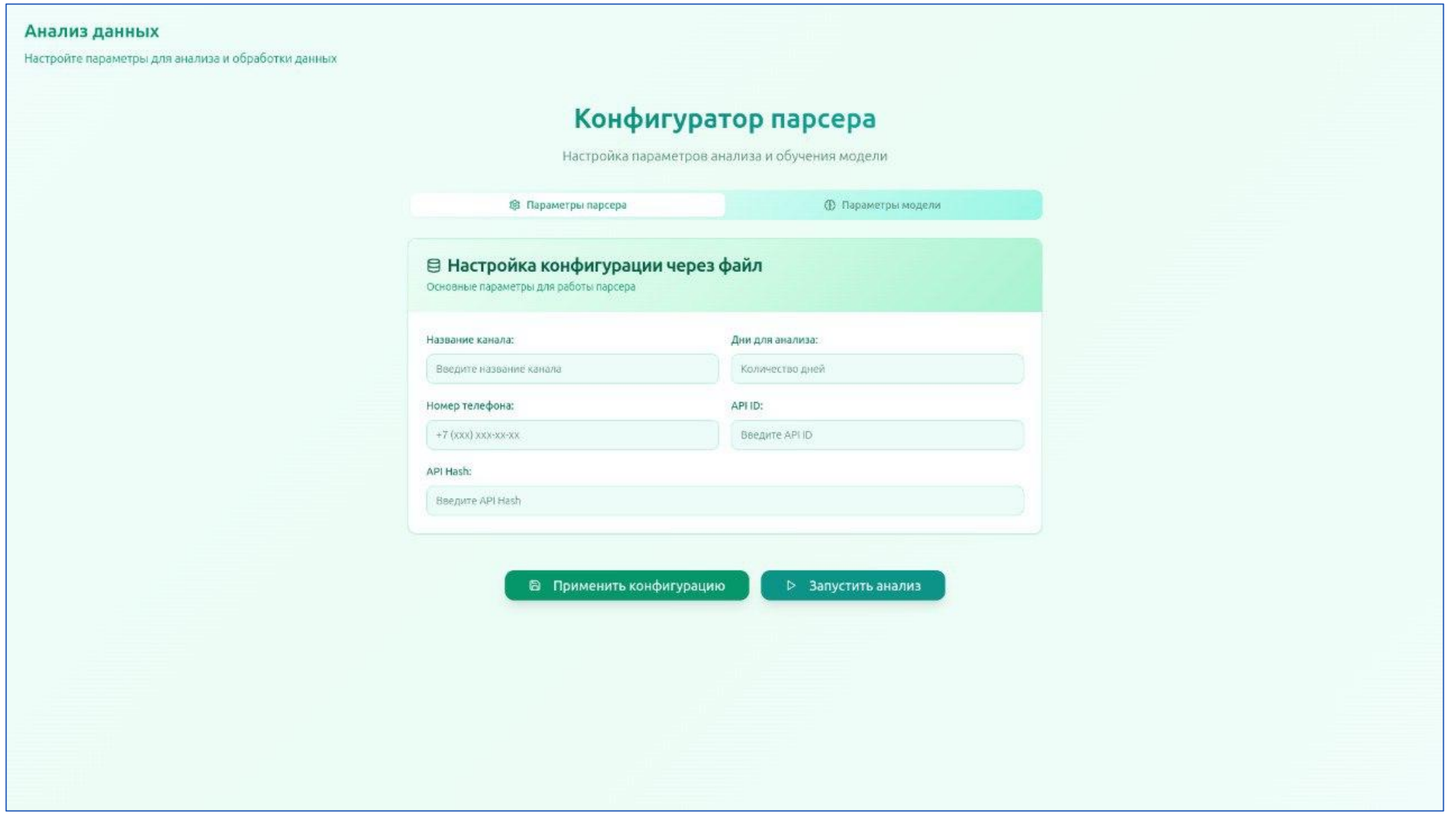
Состав ПК

- Модуль авторизации и извлечения данных
- Модуль классификации
- **Модуль обработки и сохранения результатов**
- GUI



Состав ПК

- Модуль авторизации и извлечения данных
- Модуль классификации
- Модуль обработки и сохранения результатов
- GUI



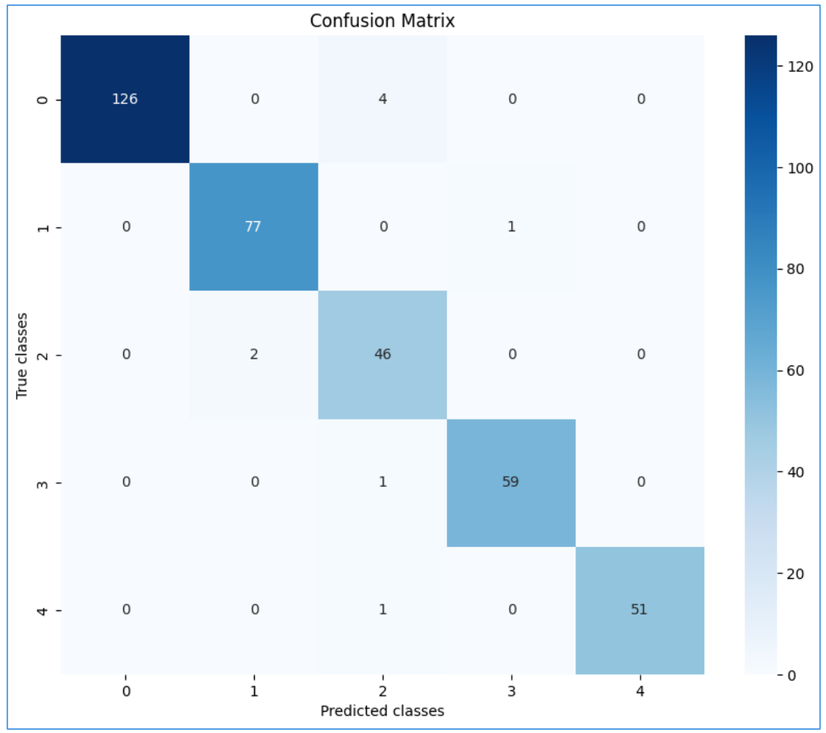
Входные данные: содержание конфигурационного файла

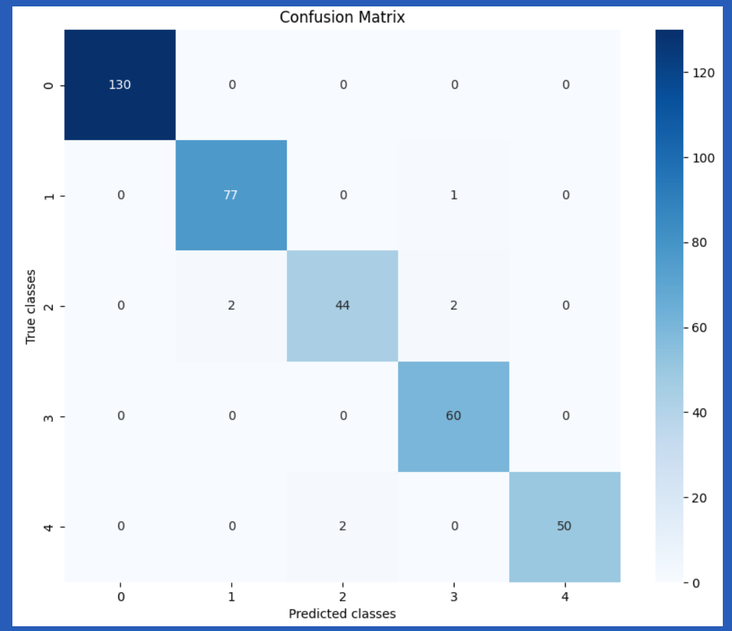
data_path	= путь_до_папки_с_обучающими_данными
model_path	= путь_до_папки_с_сериализованными_.pt_моделями
random_state	= параметр_для_контроля_случайности
data_mean	= средние_значения_для_тренировочных_данных
data_std	= СКО_для_тренировочных_данных
valid_part	= доля_данных_которая будет_выделена_под_валидационную_выборку
batch_size	= размер_батча
api_id	= api_id_для_подключения_к_аккаунту_telegram
api_hash	= api_hash_для_подключения_к_аккаунту_telegram
phone_number	= номер_телефона_к_которому_привязан_аккаунт
channel_name	= название_канала_который_необходимо_проанализировать
output_folder	= путь_до_папки_сохранения_фотографий_из_анализируемого_канала
days	= количество_дней_за_которое_надо_проанализировать_информацию

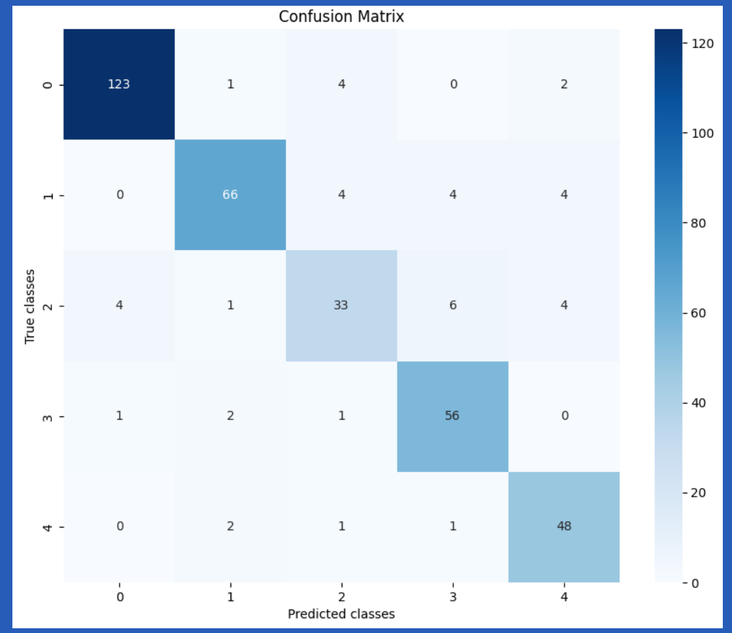
Базовые модели

Сравнение метрик качества обучения

Архитектура модели	F1
ResNet50	0.976
DenseNet161	0.981
VGG-16	0.886





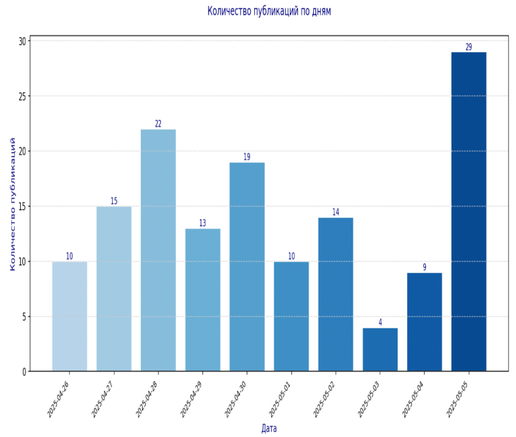


Аналитический отчет

Отчет по результатам анализа Telegram канала "КБ"

Дата: 06.05.2025 05:05:55

Временной анализ активности канала



Аналитический отчет содержит

Идентифицирующие сведения о канале социальных медиа, условиях проведения проверки, определение характера деструктивных данных

Методику проведения проверки и особенности учтенные аналитиком. Детали проведения анализа

Выводы о наличии деструктивного контента.
Предложения по дальнейшим действиям, подготовленные ИИ

Применение ИИ

Аналитический отчет формируется с применением ИИ

41ChAT →

- Попарный анализ изображений и их подписей обеспечивает глубокое понимание контента
- Текстовый анализ для глубокого понимания контента: выявление ключевых тем и важной информации
- Обобщение результатов анализа изображений и текстов, принятие решения о вредоносности рассматриваемого канала

Адаптация для заказчиков

Анализ разных типов контента

- Текст
- Видео
- Аудио
- Ссылки
- Файлы прочих форматов

Улучшение процесса анализа

- Применение различных предобученных моделей
- Добавление моделей обнаружения значимых областей входных изображений
- Добавление объяснения интерпретации результатов для повышения доверия к системе
- Обучение моделей на пользовательских данных

Гибкая настройка парсера данных

- Подключение других социальных медиа
- Добавление расширенных возможностей по настройке анализируемого временного периода
- Создание хранилища результатов для сравнения отчетов о разных запусках программного комплекса

Т-АВИАС

Российский разработчик программного обеспечения беспилотных авиационных систем

Кузьмина Ксения Александровна

Тел.: +7(926)924-1027

E-mail: ksenon2512@gmail.com

Резидент АНО «ФЦ БАС»

ФОНД НТИ Фонд поддержки проектов Национальной технологической инициативы

Резидент АНО «НПЦ БАС САМАРА»

НПЦ САМАРА БЕСПИЛОТНЫЕ АВИАЦИОННЫЕ СИСТЕМЫ

Партнеры

НИЦ ИНСТИТУТ ИМЕНИ Н.Е. ЖУКОВСКОГО

ИВК ИНФОРМАЦИОННАЯ ВНЕДРЕНЧЕСКАЯ КОМПАНИЯ

МЧС РОССИИ ФГБУ ВНИИ ГОЧС (ФЦ)