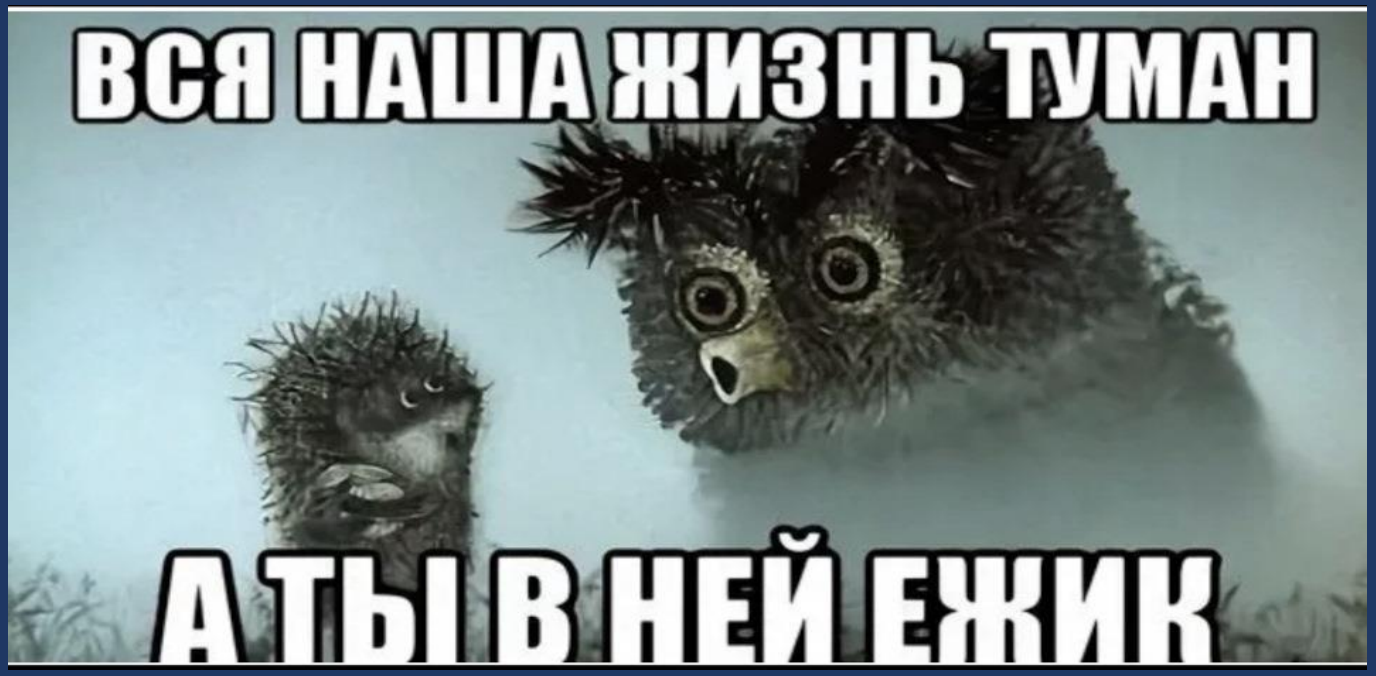
Григорий Кожанов Представляет…



И Вы представьте…

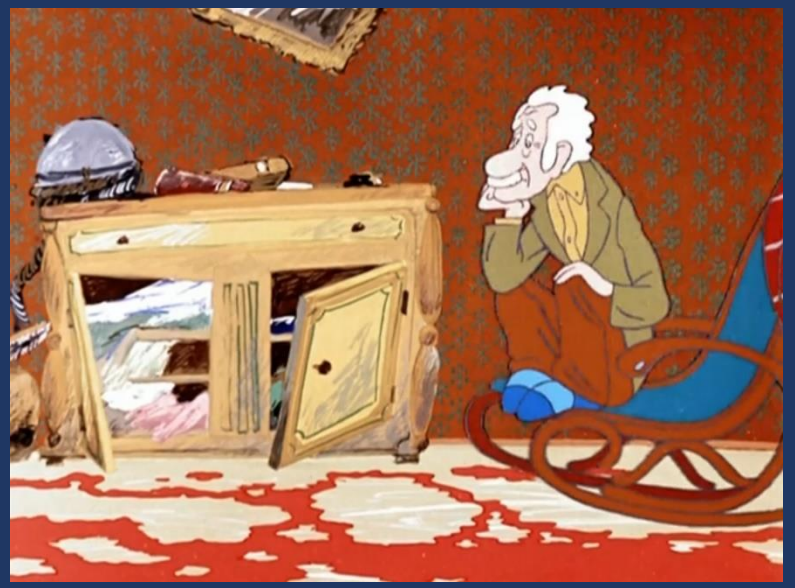
Основные задачи / сложности:

- Обработка текстов с ошибками и на разных языках

- Классификация роликов без ручной разметки данных

- Создание гибкой модели, способной обучаться на текстах и числовых данных

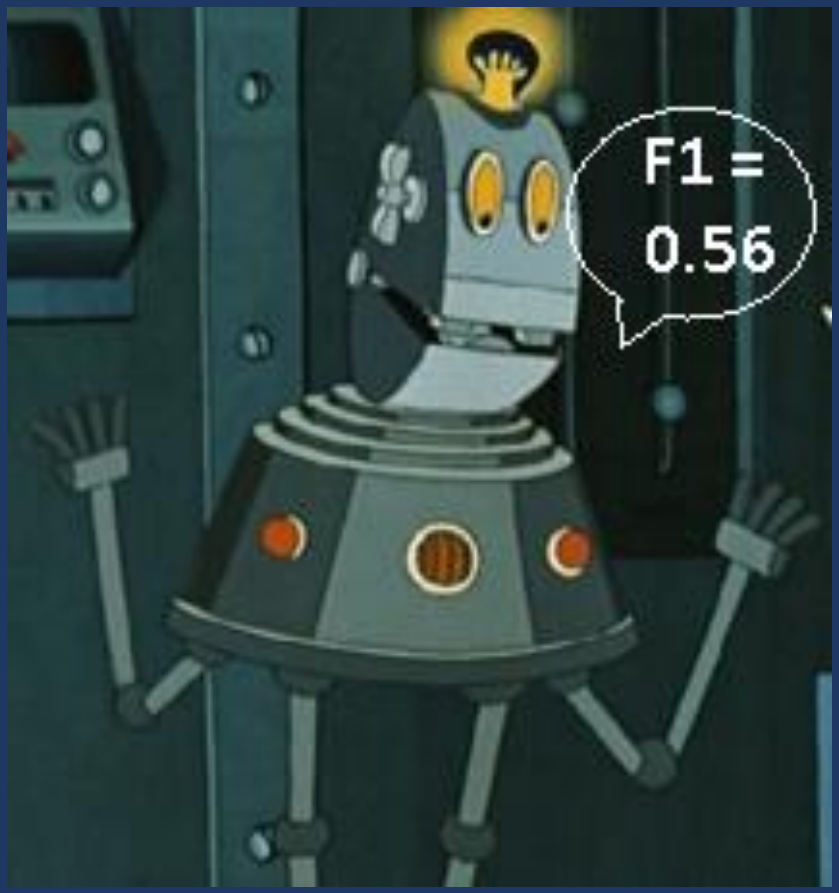
- Создать автоматизированное решение для классификации



Подход и основные этапы работы

Использование ML

- DistilBERT + Dense Layers: для обработки текстовых данных с параллельной обработкой числовых данных через полносвязные слои. Гибридная архитектура: объединение выходов трансформера и Dense слоев для финальной классификации.
- Обучение на сниппетах данных и на всем датасете.
- Максимальный F1-score – 0.56
- Ограничения: технические – для загрузки более сложных моделей и более «богатых» параметров нужно более мощное оборудование
- Результат- импортируемый модуль



В поисках киллер-фичи.

- Наставник намекнул, я поверил))

- Были проверены разные условия расщепления

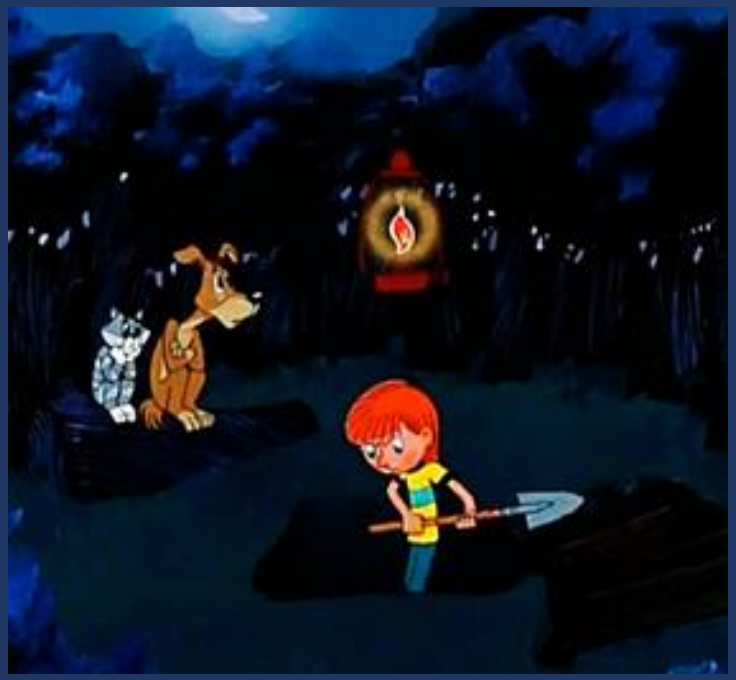
- 34 класса выделяются по минимальному паттерну

- Для остальных классов есть индивидуальные доп. условия.

- Созданы словари для классов и для паттернов (инструкций)

- Создана функция для классификации

- Максимальный F1-score – 0.96



- Ограничения: функциональные- не автоматизировано
- Результат: функция.

План автоматизации:

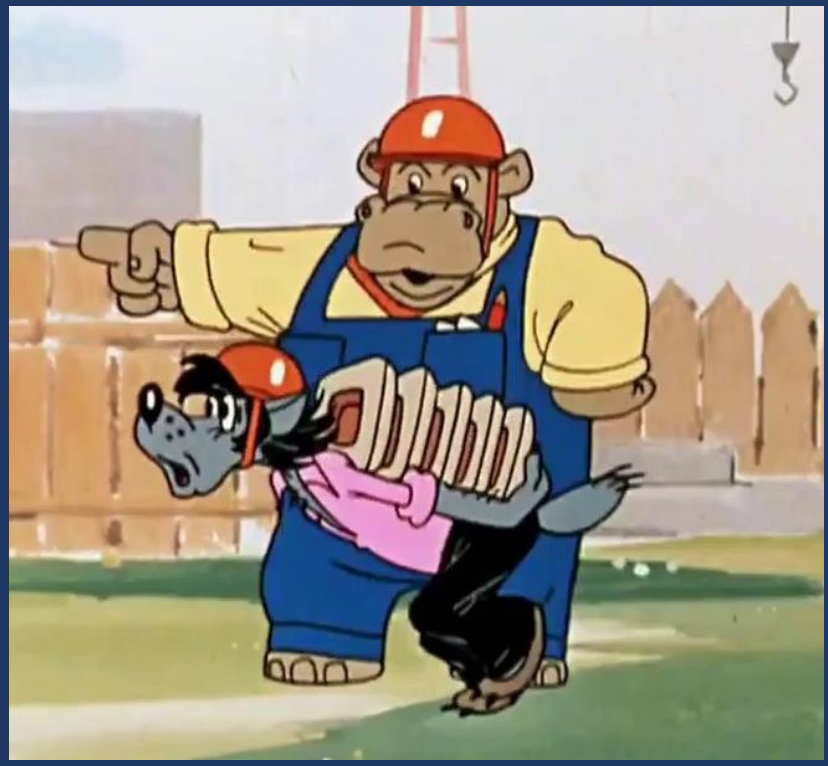
- каскадная классификации: от однозначно определяемых классов к сложным

- использование классификаторов OVR

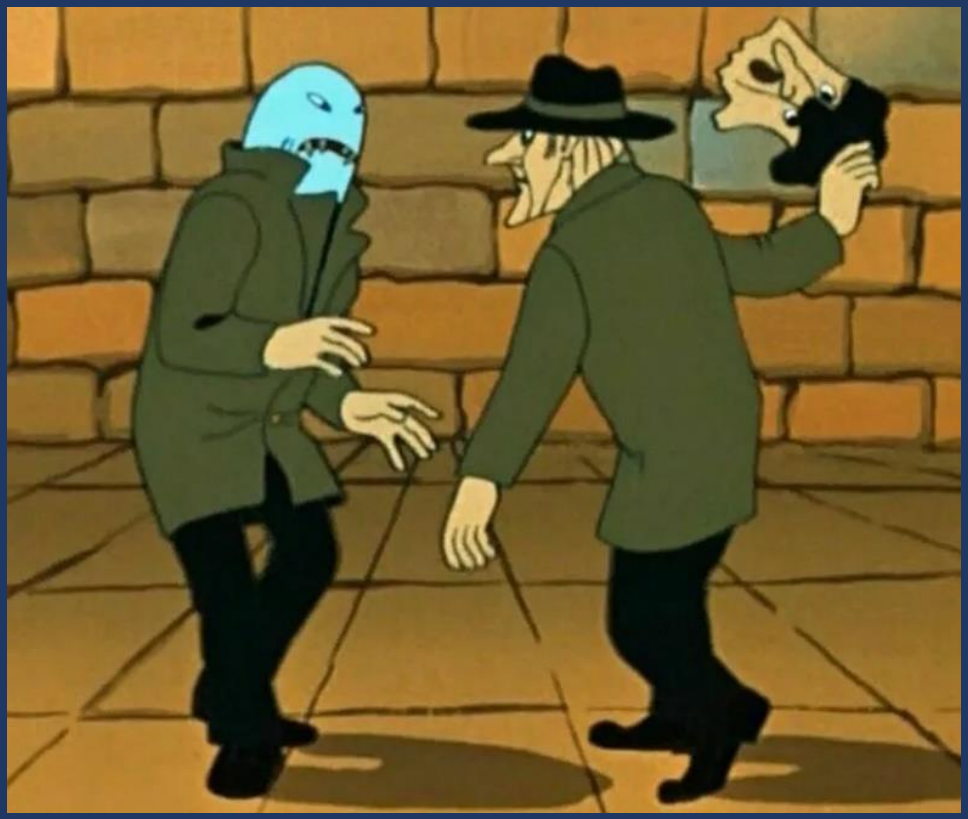
- штрафы на длину паттернов

- стоп-паттерны

- именованные сущности

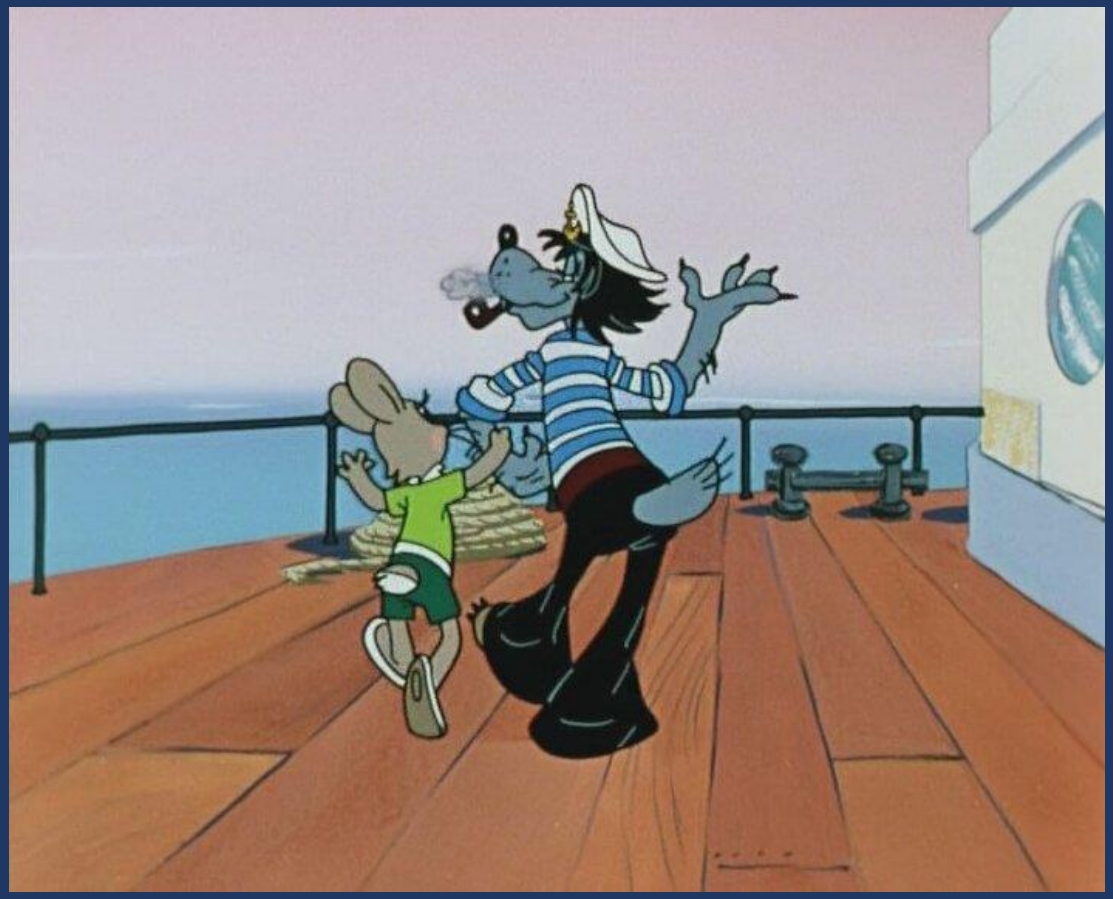


Ошибки разметки



- ЖилаБыла Царевна vs Царевны + есть прямые ошибки (мультик для детей "Пип и Альба)
- Фиксики - Непонятна логика разметки. В класс попадают как песенки с Фикисками, так и не попадают другие экземпляры- такие же песенки.
- Синий Трактор vs ЖилаБыла Царевна – пересечение по песенкам
- Малышарики vs ЖилаБыла Царевна – также
- Приключения Пети и Волка – id 38654- ошибочно размечен, 60320, 74490 – не размечены
- Крутиксы – не ошибки, но непонятна разметка на некоторых рилсах
- Чебурашка - 73732 none [Трогательные моменты дружбы между Чебурашкой и Геной Союзмультфильм shorts]
- Ну_погоди каникулы пересекается с Петей и Волком – непонятные правила
- Цветняшки – не все попадают в разметку. По-моему д.б. больше

Уходим в закат…



Спасибо за внимание!