AI-DRIVEN РАССТАНОВКА ЗНАКОВ ПУНКТУАЦИИ

ПОДГОТОВИЛИ: СУВОРОВ М.Д., РАКИТЯНСКИЙ В.М., ПЕТРОВ Р.А.

ГРУППА: М8О-111М-21 И М8О-105М-21

ПРЕПОДАВАТЕЛИ: СУДАКОВ В.А., ПАНОВСКИЙ В.Н.

Этапы работы

Сбор и подготовка датасета

Выбор и обучение модели

Сбор веб-приложения на основе Streamlit

Использованные технологии

Глубокое обучение

- torch
- keras

Токенизаторы

- AutoTokenizer
- DeepPavlov /rubert-base-cased-sentence

Сбор датасета

Был взят датасет Lenta2 из библиотеки corus (проект Наташа)

Датасет Lenta включает в себя текст и заголовки статей с сайта lenta.ru в количестве 800975 штук, весом почти в 2Гб, в каждом тексте примерно 5-10 предложений.

Подготовка датасета

Пример предложения:

Казнить нельзя, помиловать.

Пример вывода:

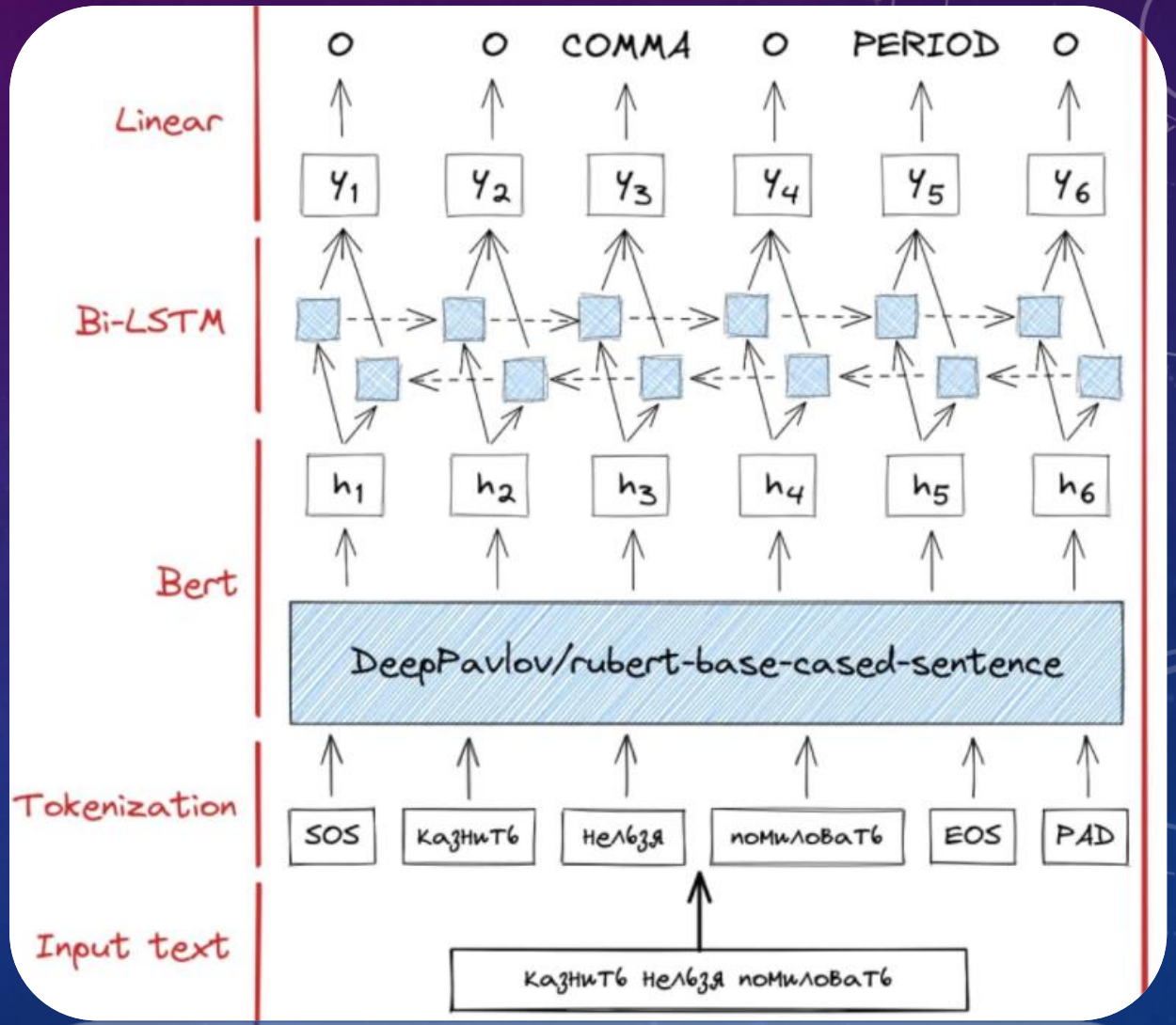
казнить	O
нельзя	COMMA
помиловать	PERIOD

Перевод в числовое представление:

tokens	['[SOS]', 'казнить', 'нельзя', 'помил', '##овать', '[EOS]', '[PAD]', '[PAD]']
x	[101, 65272, 18960, 34994, 6123, 102, 0, 0]
attn_mask	[1, 1, 1, 1, 1, 1, 0, 0]
y	[0, 0, 1, 0, 2, 0, 0, 0]
y_mask	[1, 1, 1, 0, 1, 1, 0, 0]

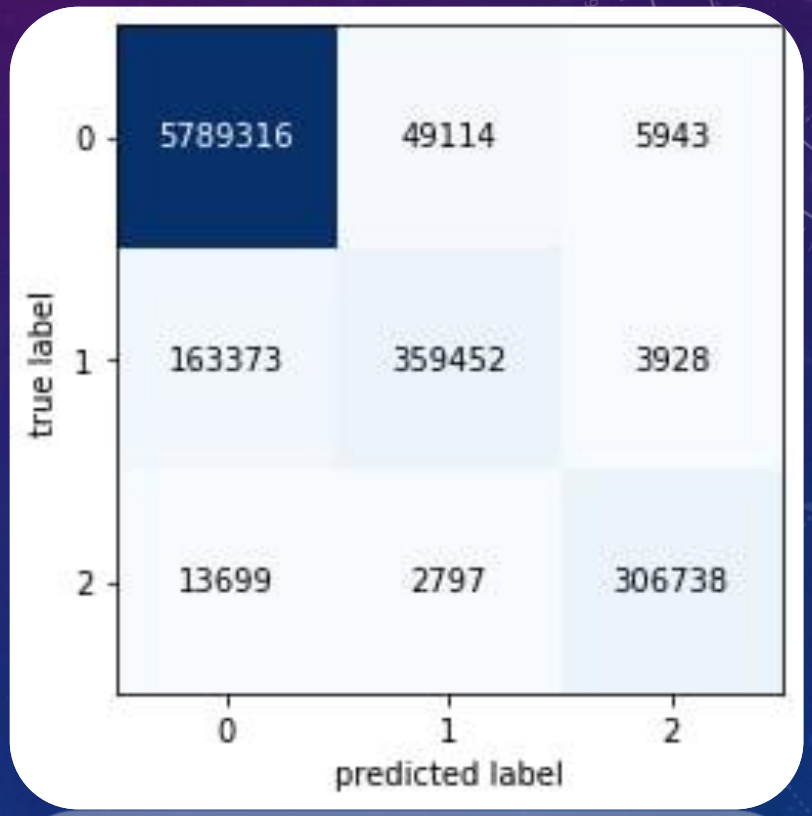
Выбор и обучение модели

- Предобученная модель-трансформер (DeepPavlov/rubert-base-cased-sentence)
- Два слоя LSTM
- Слой Linear



Результаты

Precision (0.97, 0.87, 0.96, 0.91)
Recall (0.99, 0.67, 0.94, 0.78)
F1 (0.98, 0.76, 0.95, 0.84)
Accuracy 0.96



cpu

Text to analyze

Это классическое определение Из него следует что автор берет какое-то значительное в жизни общества явление Далее он пытается его осмыслить и проанализировать Сделать на основе каких-то ситуаций например информационного повода обобщение и выводы Выдвинуть идею концепцию и рассмотреть ее Это основная цель статьи Неслучайно когда защищаем кандидатскую диссертацию то перед защитой готовим несколько публикаций Здесь мы пытаемся открыть и проанализировать нечто новое Это аналитический жанр Он не столько информационный сколько аналитический Тут осмысливаем события приводим какие-то примеры исследуем явления

Result: Это классическое определение. Из него следует, что автор берет какое-то значительное в жизни общества явление. Далее он пытается его осмыслить и проанализировать. Сделать на основе каких-то ситуаций, например информационного повода, обобщение и выводы. Выдвинуть идею концепцию и рассмотреть ее. Это основная цель статьи. Неслучайно когда защищаем кандидатскую диссертацию, то перед защитой готовим несколько публикаций. Здесь мы пытаемся открыть и проанализировать нечто новое. Это аналитический жанр Он не столько информационный, сколько аналитический. Тут осмысливаем события, приводим какие-то примеры исследуем явления

Применение:

- в задачах генерации текста
- в улучшении правописания текстов
- подзадача в распознавании речи

Точки роста:

- адаптация к другим языкам
- большее количество экспериментов и лучшее железо для качественного обучения и расстановки других типов знаков препинания