анализ тональности для азербайджанского.

Григорьев Севастьян
4 курс ОТиПЛ

Постановка цели и задачи

Анализ тональности.

Датасет:https://huggingface.co/datasets/DGurgurov/azerbaijani_sa/viewer/default/test?views%5B%5D=test

Данные были собраны из различных источников, таких как социальные сети, обзоры.

его набор данных содержит набор данных для анализа настроений из Local Doc (2024)

Enhancing GloVe Embeddings for Low-resource Languages with Graph Knowledge

Данные были использованы для проекта по: https://github.com/pyRis/retrofitting-embeddings-lrls?tab=readme-ov-file

Этот проект направлен на усовершенствование эмбеддингов GloVe для языков с ограниченными ресурсами за счет использования знаний о графах, а также на создание централизованного хранилища с предварительно подготовленными статическими эмбеддингамидля различных языков.

ФОРМАТ ДАННЫХ

text: user review / comment

labels: sentiment label

В нынешнем датасете 2 класса: 1: положительный

0: отрицательный

В датасете Local Doc (2024) 3 класса: положительные

отрицательные

нейтральные

Пример данных

text: Dunya seyaheti etmek ucun limit-siz bilet ve

мир Путешествие.ACC делать чтобы лимитный-без билет
и

pul

деньги

Неограниченное количество билетов и денег, чтобы путешествовать по миру.

label: 1

Анализ данных

train_data – тренировачная выборка (19600)

validation_data – валидационная выборка (4200)

test_data – тестовая выборка (4200)

Планируемая модель:

DGurgurov/xlm-r_azerbaijani_sentiment:
https://huggingface.co/DGurgurov/xlm-r_azerbaijani_sentiment

МЕТРИКИ ДЛЯ ОЦЕНКИ

точность — 0,79381;
макро F1 — 0,79378;
микро F1 — 0,79381.

Детали обучения

- Epochs: 20
- Batch Size: 32 (train), 64 (eval)
- Optimizer: AdamW
- Learning Rate: 5e-5

Модель вообще обучена, поэтому можно попробовать прогнать её на зеро-шот или фью-шот промптинге (какой именно, пока не знаю)

Все метрики и детали обучения оставлю такими же.

В случае дисбаланса скорее всего воспользуюсь модификацией функции потерь. Например, можно добавить штраф за ошибки в классификации класса меньшинства, чтобы минимизировать ошибки в этом классе.