МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Национальный исследовательский ядерный университет «МИФИ»

ИНСТИТУТ ИНТЕЛЛЕКТУАЛЬНЫХ КИБЕРНЕТИЧЕСКИХ СИСТЕМ

КАФЕДРА КИБЕРНЕТИКИ (№ 22)

НАПРАВЛЕНИЕ ПОДГОТОВКИ **09.03.04 “Программная инженерия”**

Учебно-исследовательская работа на тему:

Разработка и сравнительный анализ моделей машинного обучения для ответа на связанные с медициной контекстно-зависимые вопросы

Студент: Плотников В. И.

Группа: Б21-514

Научный руководитель: Сбоев А. Г.

Москва, 2019

Реферат

Общий объем основного текста, без учета приложений - 21 страницы, с учетом приложений 29.
Количество использованных источников 11.
Количество приложений 2.
Количество рисунков 7.
Количество таблиц 2.

Ключевые слова: NLP, text classification, BERT, transfer learning.

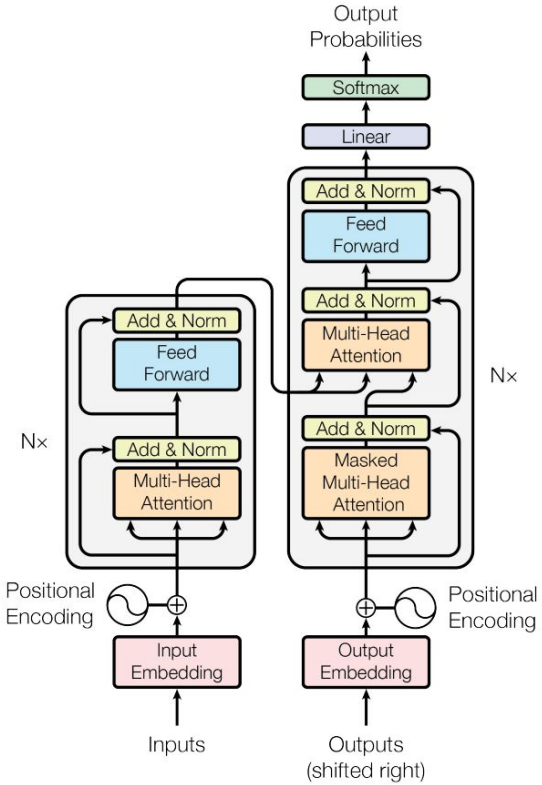
Актуальность работы

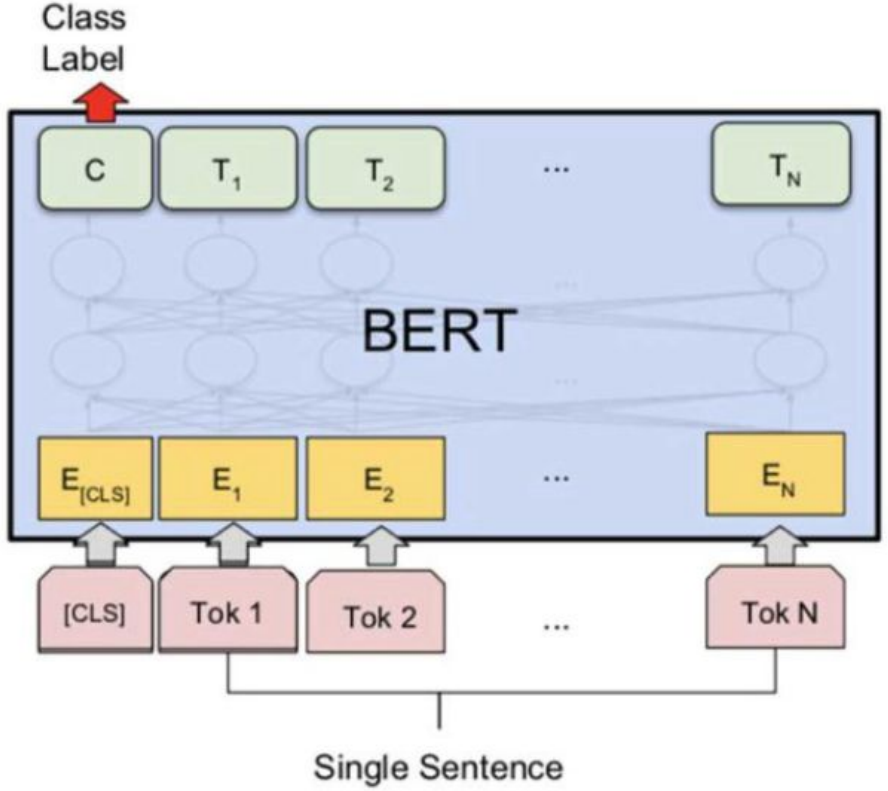
С ростом объема медицинских данных возникает необходимость в эффективных методах их обработки и анализа. Одним из направлений в этой области является задача оценки способности моделей "извлекать информацию" из текста и правильно отвечать на уточняющие вопросы, в том числе связанные с медициной.

Цель УИР

- Разработка моделей для решения задачи ответа на связанный с медициной контекстно-зависимый вопрос

Анализ подходов решения задач классификации в условиях недостатка данных и вычислительных мощностей





Задачи УИР

1) Нахождение модели, натренированной на похожих данных
2) Нахождение размеченных данных для дообучения модели
3) Обучение на новых данных

Моделирование алгоритма для дообучения сети

Для реализации необходимы следующие функции:

- Функция векторизации вопроса и контекста для подачи в модель
- Функция начальной инициализации (загрузки) модели
- Функция дообучения модели
- Функция для оценки результатов дообучения
- Функционал для взаимодействия пользователя с дообученной моделью

Список системных и пользовательских требований

Системные требования: в процессе обучения сети должно использоваться не более 16 ГБ GPU (объем памяти GPU P100)

Пользовательские требования: пользователь должен иметь возможность ввести контекст с вопросом, затем получить ответ

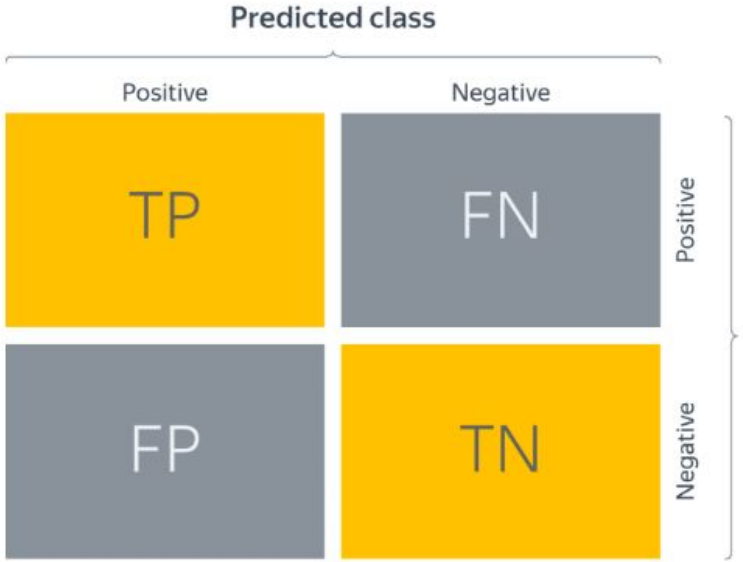
Инструментальные средства реализации

В качестве языка программирования был выбран Python из-за обилия библиотек , связанных с машинным обучением, и в частности NLP.

Модели брались из huggingface.co/models. Выбор обусловлен большим количеством доступных для дообучения моделей и возможностью загрузки напрямую с сервера

Сами вычисления производились в Kaggle notebook. Выбор обусловлен наличием возможности использования GPU P100

Метрики



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

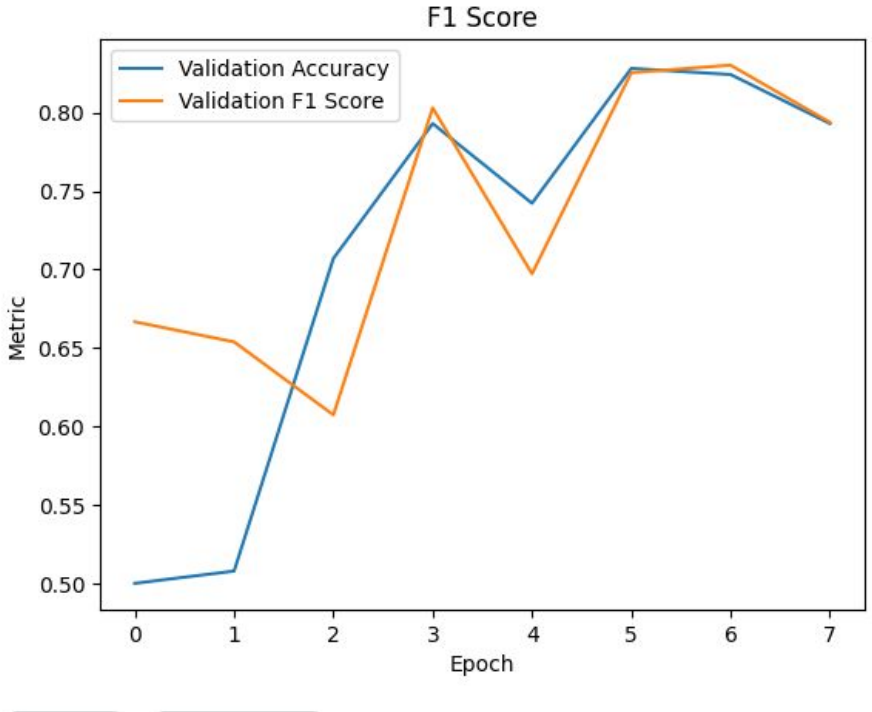
$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

Выбор модели

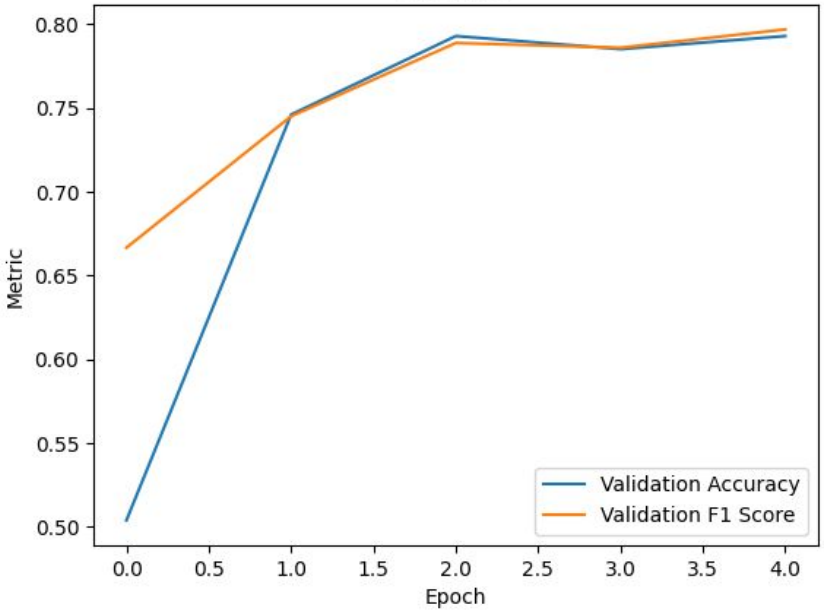
Была выбрана модель xlm-roberta-large-sag, потому что она показала лучшие значения метрик

Модель	accuracy	F1
RuBioRoBERTa	0.79	0.80
rubert-base-cased-mcn	0.59	0.59
xlm-roberta-large-sag	0.83	0.83
RuBioRoBERTa (команда SAI junior)	0.77	



Сравнение с известными аналогами

Команда SberAI использовала модель RuBioRoBERTa и дообучала на тех же данных. accuracy их модели равно 0,77. Я тоже эксперементировал с данной моделью сумел немного улучшить результат: accuracy моей модели равно 0,79



Пользовательские возможности

Пользователь может ввести вопрос с контекстом и получить ответ

```
Введите контекст:  Корь крайне опасная для здоровья и заразная болезнь
Введите вопрос:  Карантин из-за кори является крайне излишней мерой?
Предсказание: нет
```

Заключение

В ходе работы по теме “Разработка и сравнительный анализ моделей машинного обучения для ответа на связанные с медициной контекстно-зависимые вопросы ” были сделаны:

1. Обзор литературы, содержащей описание различных способов для решения задачи классификации текстов. После сравнительного анализа было принято решение использовать BERT-подобные модели.
2. Работа по подбору гиперпараметров для модели RuBioRoBERTa. При скорости обучение равной 10^(-5) были достигнуты следующие результаты: accuracy = 0,79, F1 = 0,80.
3. Сравнение различных BERT моделей. Лучший результат достигнут при использовании модели xlm-roberta-large-sag: accuracy = 0,83, F1 = 0,83.

Список литературы

1. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean / “Efficient Estimation of Word Representations in Vector Space”

2. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin / “Attention Is All You Need” 2017

3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova / “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, 2018

Спасибо за внимание!