Исследование массива медицинских данных для создания ПРЕДСКАЗАТЕЛЬНОГО СЕРВИСА

C + V team 2023

Наша команда

Ланских Святослав
ML

Динмухаметов Данис
Team lead, ML

Богдан Кристиан
DS

Демидов Григорий
ML

Анализ области

какие рекомендательные системы уже существует в мире

Для пациентов

- Рекомендации по поддержанию текущего состояния здоровья [1]
- Постановка диагноза по симптомам [2]
- Рекомендации пациентам с хроническими заболеваниями [3]
- Многие другие

какие рекомендательные системы уже существует в мире? [4]

Для врачей

- Рекомендация лекарств, которые были выписаны пациенту с похожей историей болезни
- Рекомендация историй болезней похожих пациентов
- Рекомендавать пользователям проходить опросы и делиться данными для улучшения работы сервиса

какие рекомендательные системы уже существует в мире? [4]

Для Фармацевтических компаний и клиник

- На основе большого массива данных о закупках клиник, рекомендовать производить больше востребованных препаратов
- Рекомендации больницам закупать лекарства, которые пользуются спросом в других схожих больницах

Наша рекомендательная система

на основе представленных данных

Проблема: не всегда при первых сиптомах болезни человек может понять к кому ему стоит обратиться, какие препараты принять и как себя вести, чтобы не усугубить болезнь

Решение: рекомендуем пациенту специальность врача, к которой ему стоит обратиться, препараты, основываясь на рекомендации других врачей*, а так же первичные действия для устранения заболевания.

* здесь стоит отметить, что нужно обязательно прописать условия о том, что клиника не несёт ответственности за рекомендованные препараты, т.к. у каждого пациента свои противопоказания и всегда лучше проконсультироваться с живым специалистом

Цель работы

Определим цель работы

Наша главная цель - **помочь людям**, которые не могут по симптомам своей болезни определиться что им надо делать, а это в свою очередь может привести к ухудшению состояния пациента

Постановка задач

- Изучить датасет и выявить его характерные признаки
- Придумать наилучший технический подход для достижения поставленной цели
- Реализовать этот подход и сравнить с другими, возможно похожими, объяснить почему мы выбрали именно его

Нужны ещё данные)

Для решения поставленных задач нам понадобятся дополнительные данные, а именно:

- База данных, в которой болезням сопоставляются первичные рекомендации
- База данных, в которой симптомам соответствует список лекарств, которые следует принимать
- База данных, в которой симптомам сопоставляются болезни [5]

Поэтому к нашим задачам прибавляется задача поиска соответствующих данных

Процесс исследования предоставленных данных

Первым делом разобьём наш датасет на категориальные и числовые признаки

```
Числовые признаки: MedicalRecordKey, Возраст, MedicalRecordDate
```

- *MedicalRecordKey* **-** идентификатор медицинской записи (никак исследовать это признак нельзя)
- *MedicalRecordDate* **-** дата и время мед. записи, можно посмотреть на распределение записей по времени суток:

```
утро:  4772
обед:  5655
вечер:  1352
```

- *Возраст* - посмотрим стандартные характеристики данного признака:

```
mean          36.468237
std           16.480664
min            0.016438
max  93.34246575342466
```

Категореальные признаки

```
PatientKey, Пол, СпециальностьВрача, Жалобы, ПеренесенныеЗаболевания, ПеренесенныеОперации, ПринимаемыеПрепараты
```

Для признаков “Пол”, “Специальность Врача” мы изучили и выделили классы, посмотрели на их распределение.

Пол:

```
1.0    7901
0.0    4863
```

На основе этих данных можно сделать вывод, что женщин, которые посещают клинику больше примерно в 1,5 раза, чем таких же мужчин

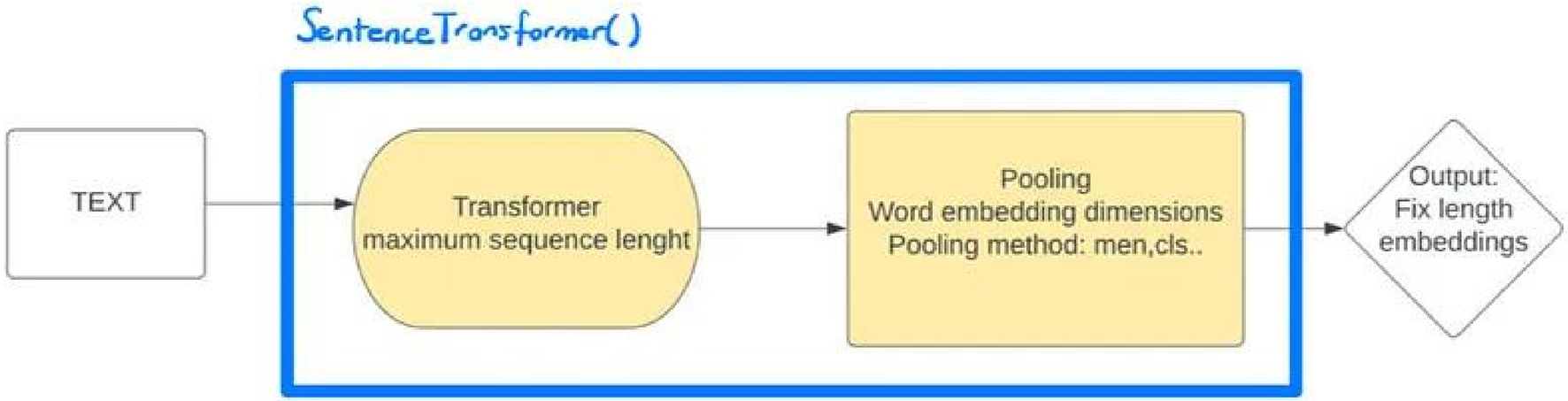
Специальность врача: 64 классов. Тут мы объединили схожих врачей, например таких, как “Уролог“ и “Заведующий отделением Уролог“ мы объединили в один класс “Уролог”, но оставили разделение на детских и не детских врачей. Однако, мы заметили очень сильный дисбаланс в классах (10 классов имеют только 1 обьект). Для нашей рекомендательной системы это довольно большая проблема. Так что бороться с ней мы будем несколькими способами:

- Генерация новых данных
- Эксперементы с функцией потерь [6]
- Искать новые данные (например попросить у Медси🙌)

Категореальные признаки. 2 часть

```
PatientKey, Пол, СпециальностьВрача, Жалобы, ПеренесенныеЗаболевания, ПеренесенныеОперации, ПринимаемыеПрепараты
```

В нашей рекомендательной системе главным и связующим признаком является признак **“Жалобы”**. Так как нам важны именно уникальные симптомы, а их довольно сложно выделить, потому что признак представляет собой строку, в которой описаны жалобы пациентов, мы решили, что будем кодировать эти признаки с помощью sentence transformer [7].



Полученные признаки

какие данные мы будем использовать для обучения

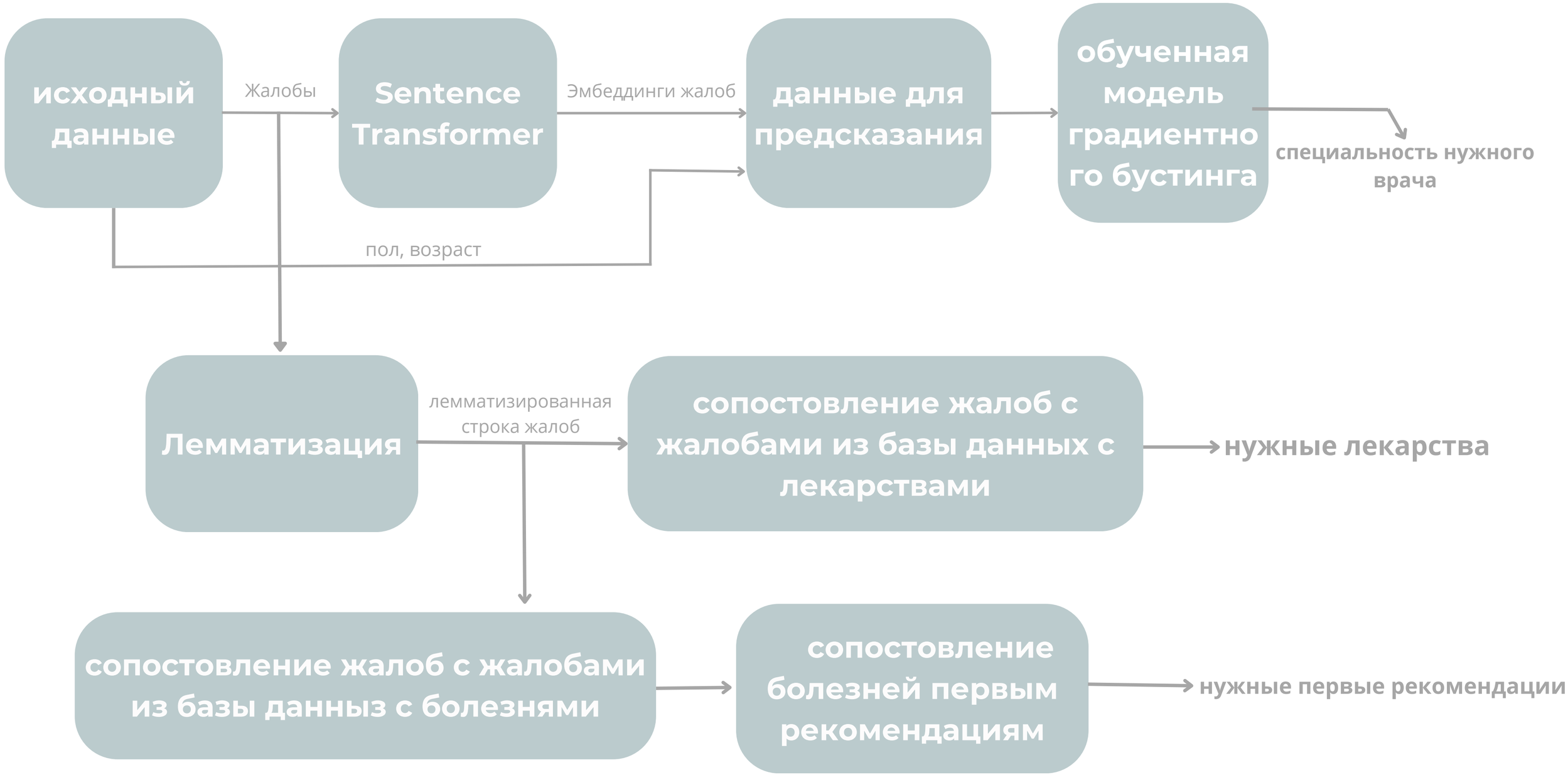
Итак, в качестве данных, на которых мы планируем обучать модели для частичного решения нашей задачи, а именно: рекомендация специальности врача, мы взяли датасет, в котором содержаться:

- Пол
- Возраст
- Закодированный признак “Жалобы”, который представляет собой эмбеддинг. В датасет мы помещаем этот эмбеддинг поэлементно.

Итого, датасет выглядит примерно так:

761	762	763	764	765	766	767	Возраст	Пол	Специальность_Врача
-1.567272	-0.020576	0.3589	0.04909	0.618903	-0.910807	0.648208	9764	1.0	0

Алгоритм рекомендаций. Промежуточный пайплайн



Какие модели мы используем?

Итак, мы сказали про общие технологии, которые мы используем. Теперь скажем про конкретные модели, которые планируем применять или уже применили

*В качестве **Sentence Transformer** мы применили **RuBert** из билиотеки DeepPavlov, т.к Bert - это наиболее популярный и качетсвенный автоэнкодер, а RuBert специализируется именно на русском языке, что нам и нужно.*
Так же мы попробуем некоторые другие модели с сайта hugging face.

*В качестве модели градиентого бустинга был использован **CatBoostClassifier**, основываясь на опыте, он даёт хорошие результаты. Так же в планах попробовать модель **HistGradientBoostingClassifier**, т.к она показывает хорошие результаты именно на числовых признаках, которыми наши данные и являются.*
Так же мы обязательно попробуем другие методы обработки табличных данных и сравним их между собой.

Первый тест(RuBert + CatBoost) без перебора гиперпараметров и с высоким дисбалансом:

```
Accuracy: 0.5998433829287392
```

Планы по улучшению решения

- обязательно сравнить разные комбинации моделей
- избавиться от сильного дисбаланса классов
- применить другие способы решения задачи для получения первичных рекомендаций
- Придумать как можно связать рекомендацию лекарств, врача и первичных действий воедино
- Задействовать оставшиеся признаки из данного нам датасета

Полезные материалы

- [1] https://www.apple.com/ios/health/
- [2] https://today.duke.edu/2015/10/autismbeyond
- [3] https://www.broadcastmed.com/neurology/5225/news/epiwatch-app records-seizure-data-using-apple-watch
- [4] https://www.mathnet.ru/links/335f3429b3ff5cba8dc0433b02d20d97/ubs1024.pdf
- [5] https://github.com/rahul15197/Disease-Detection-based-on-Symptoms
- [6] https://medium.com/visionwizard/understanding-focal-loss-a-quick-read-b914422913e7
- [7] https://medium.com/@gulsum.budakoglu/from-sentencetransformer-transformer-and-pooling-components-7d9ad4fcd70f

Наш репозиторий

GitHub

C + V team