Проектный практикум

Задача Beeline: качество транскрибации

Команда «DataSorcerers»:

Вяткин Роман
Баймлер Ярослав
Ихматуллаев Даврон
Косачев Дмитрий
Назаров Михаил
Новиков Валентин
Яськова Марина

май 2024

Как без выполнения разметки данных определить, хорошо ли ASR-модель транскрибировала аудио?

- Бинарная классификация
- Целевая метрика — ROC-AUC

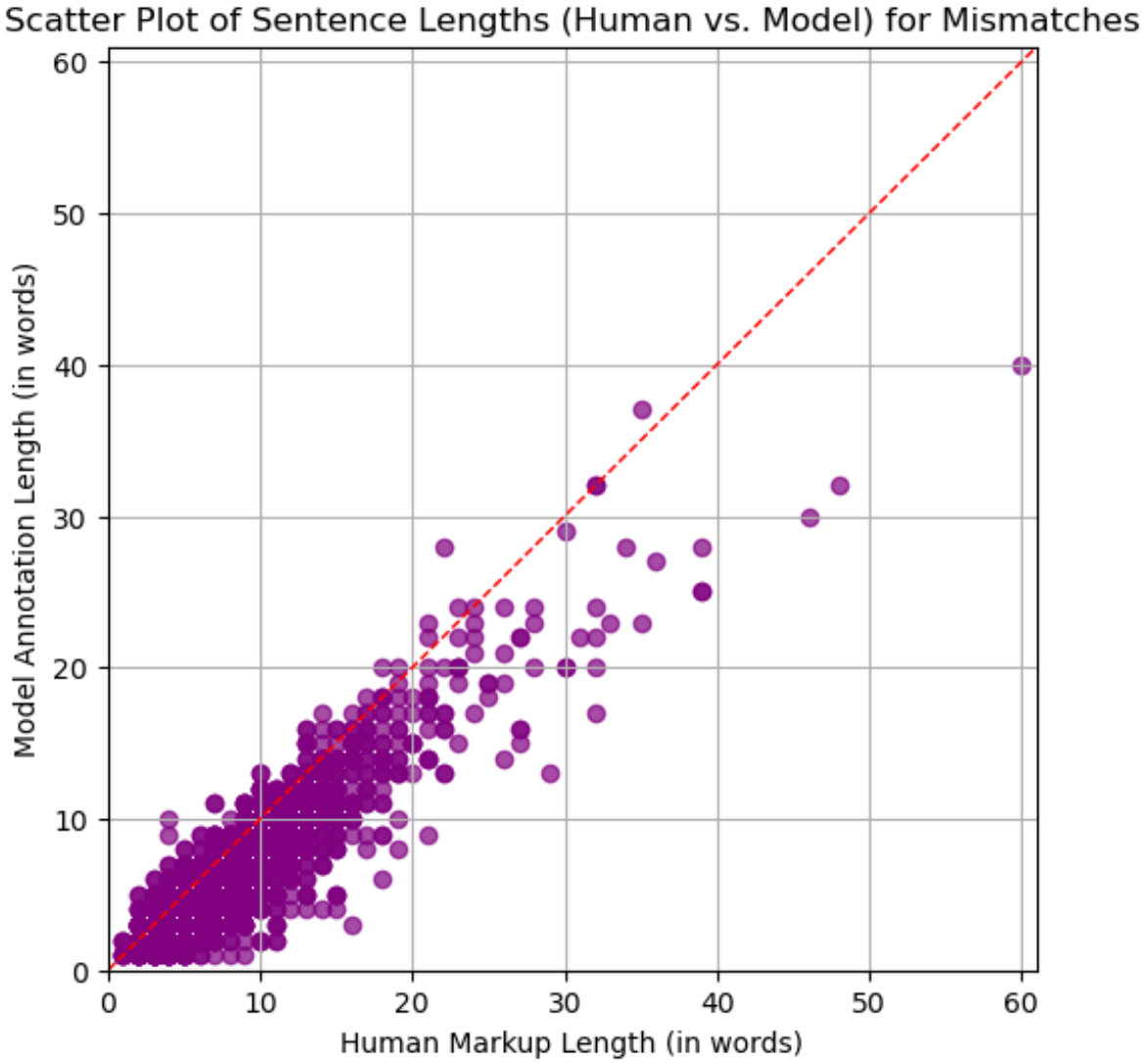
Датасет

6508 записей

Столбцы:

- Аннотация модели
- Ручная разметка
- Путь к аудио
- Метка класса (совпадает или нет)

Все однородно и без пропусков



Подготовка данных: векторизация

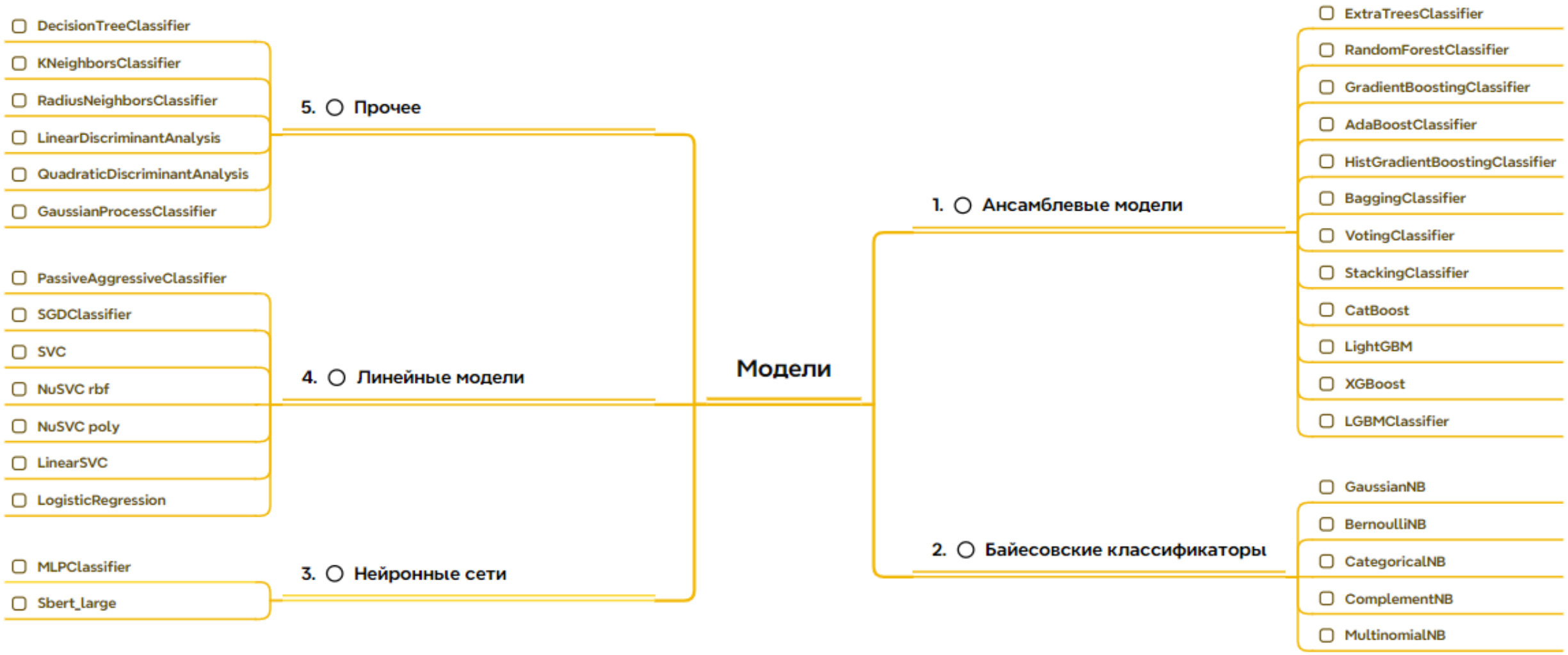
BERT
sbert_large_mt_nlu_ru

Scikit-learn

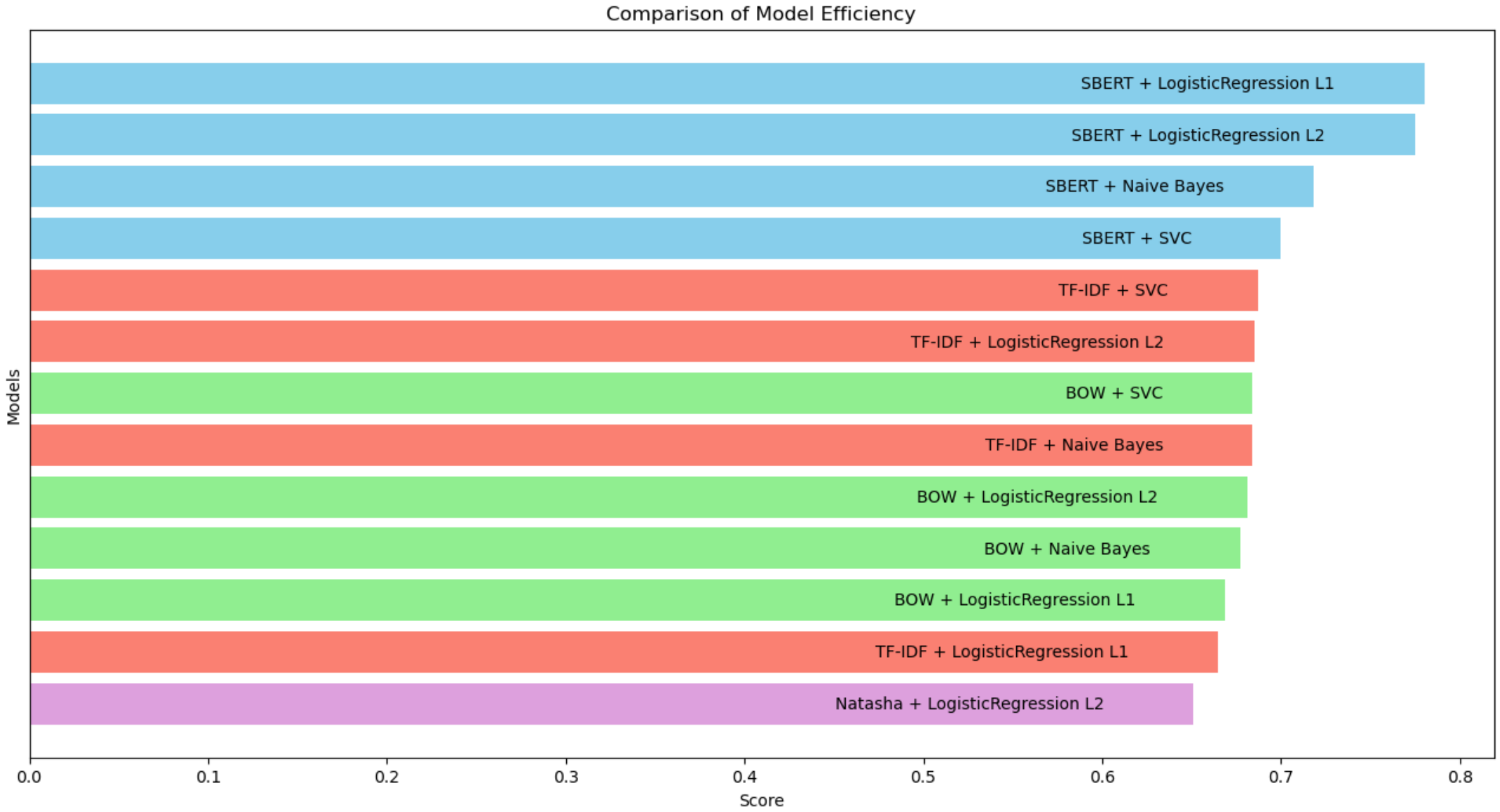
- Bag of words: CountVectorizer()
- TF-IDF with unigrams and bigrams

Natasha
Navec

Протестированные модели



Векторизация



Аугментация данных

при помощи модели Whisper (от OpenAI):

mitchelldehaven/whisper-medium-ru

на данных:

https://github.com/snakers4/open_stt

\+ **15k данных**

0.80
accuracy при сравнении с Human Markup

0.771
ROC-AUC на тестовом датасете*

* sbert + LogisticRegression

Дообучение модели

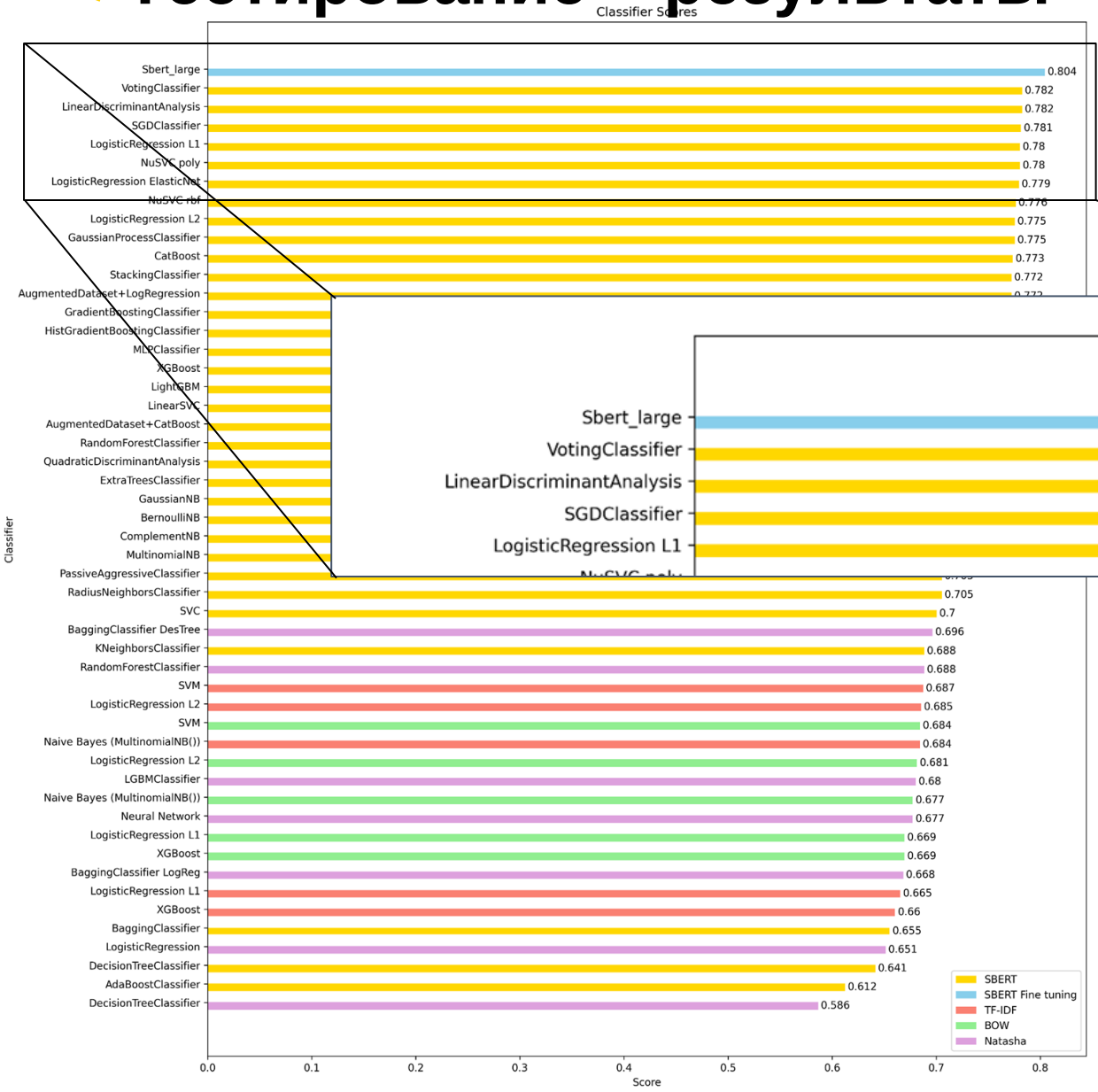
sbert_large_mt_nlu_ru

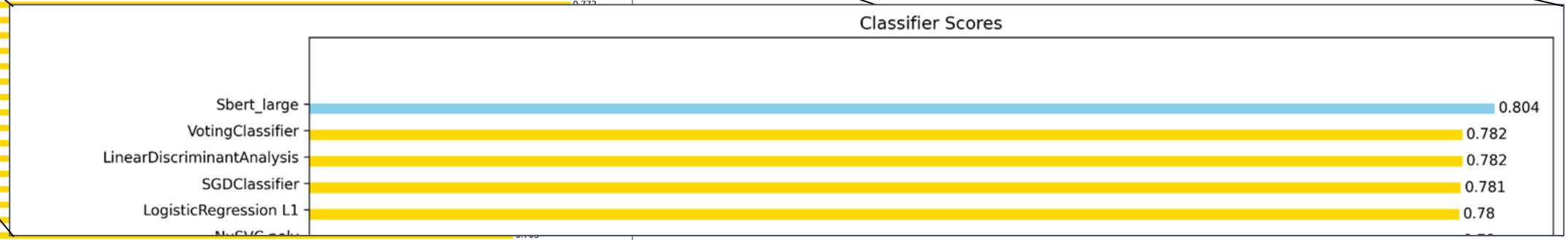
0.804
ROC-AUC

“Я тебе не звонил..” → 8x [Encoder** … Encoder** … Encoder] → Linear Layer** → Label

** обучаемые слои

Тестирование + результаты «финалистов»





0.804

ROC-AUC

/Выводы

- Достигнут ROC-AUC 0.804
  SBERT fine-tuning + Linear layer
- Метод векторизации имеет ключевое значение
- Нужно больше данных

kosatchev/ClarityAnalyzer

nazarovmichail/sbert_large_transcription_classification