Презентация хакатона Data Hack:

Задача классификации новостных текстов с присвоением тегов

команда **Fit-predict**

https://github.com/MichaelYashchenko/fit-predict-datahack

Команда

Никита Ятченко

Team Lead

Татьяна Яковлева

Data Scientist

Мария Швецова

Data Analyst

Михаил Ященко

Developer

Мишан Алиев

Data Scientist

Основные гипотезы

Гипотеза 1

Использование BERT Classifier vs Embedding + Boost для тэгирования новостей

Гипотеза 2

LLM как инструмент аугментации и генерации новостей

Гипотеза 3

BERTopic может быть эффективным инструментом для выделения ключевых слов

Гипотеза 4

NER как бейзлайн для выделения ключевых сущностей

Работа с данными

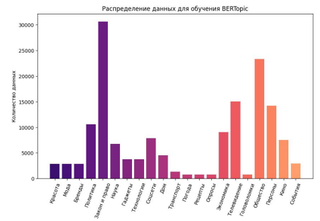
Парсинг

1. Расширение текущего словаря
2. Обогащение контекста
3. Улучшение точности

lenta.ru **mail.ru**

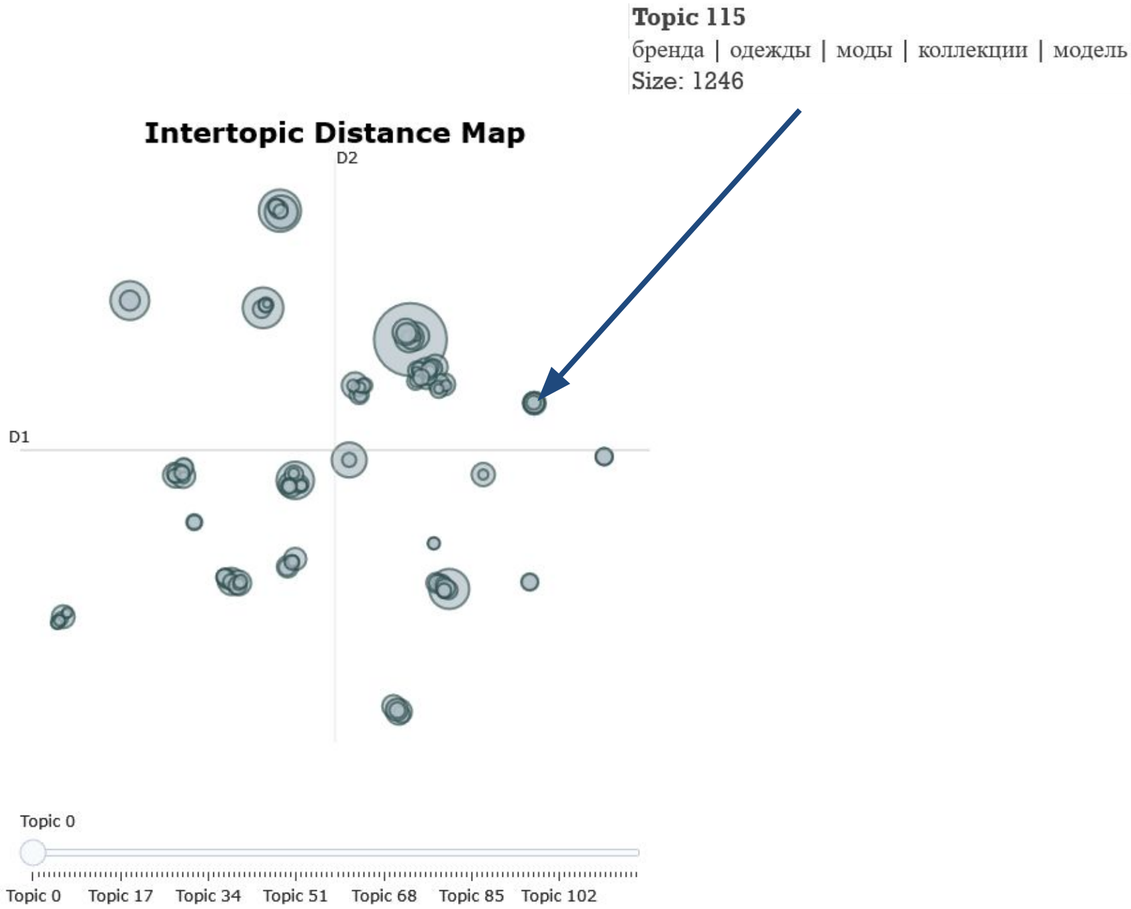
Генерация

1. Снижение дисбаланса классов (генерация LLM)
2. Аугментация данных
3. Снижение затрат на сбор данных



6000 дополнительно сгенерированных наблюдений по 6 различным категориям

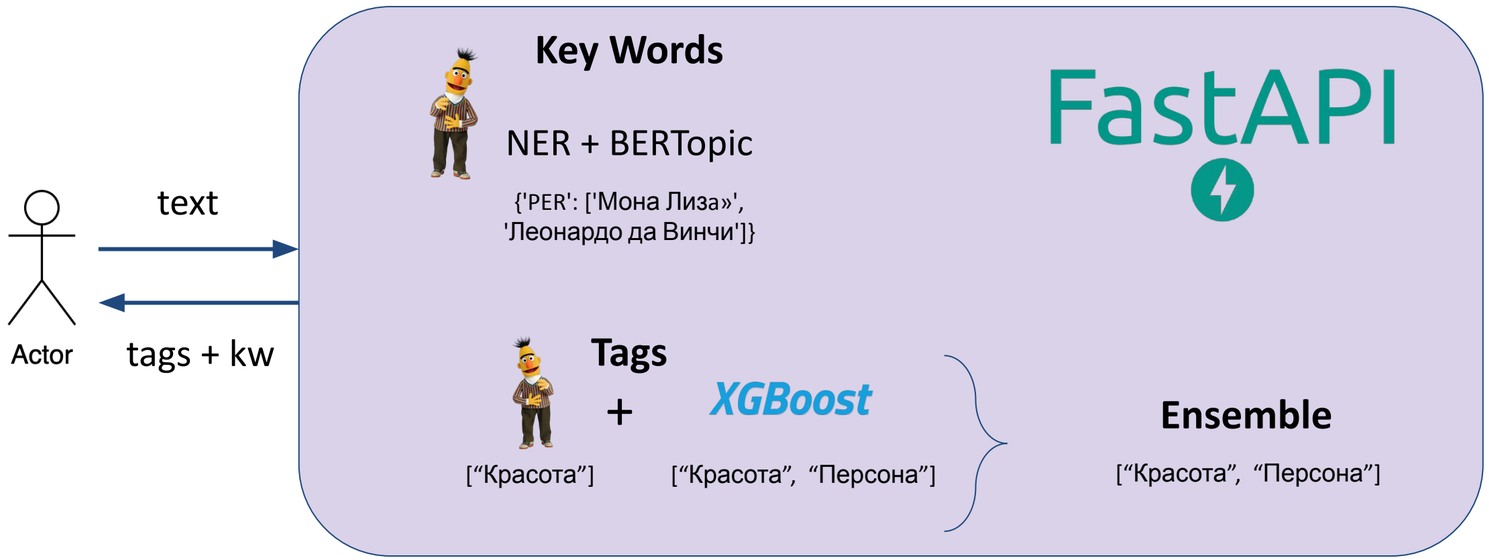
Topic modeling

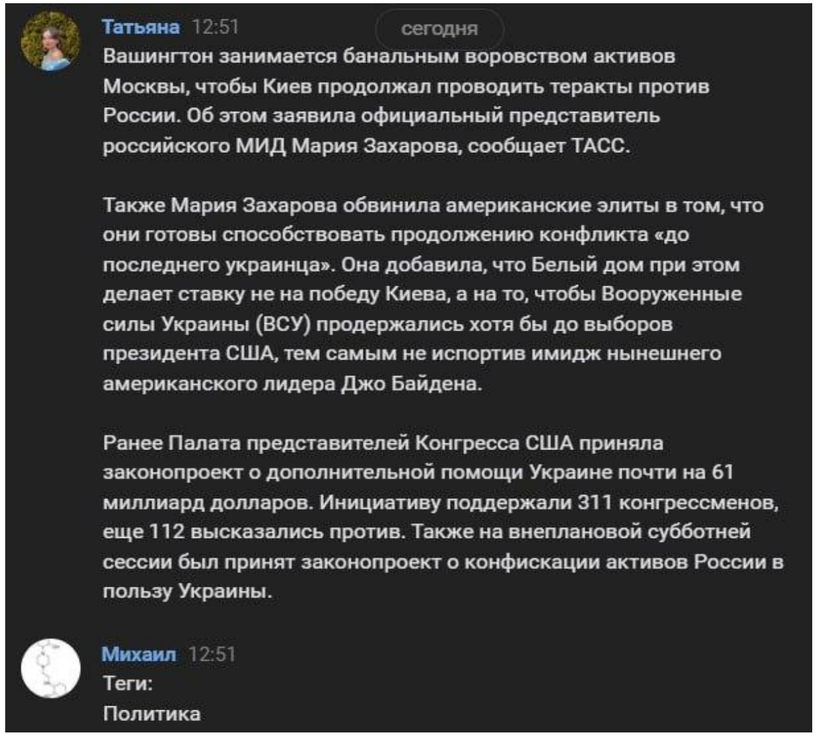


Архитектура решения

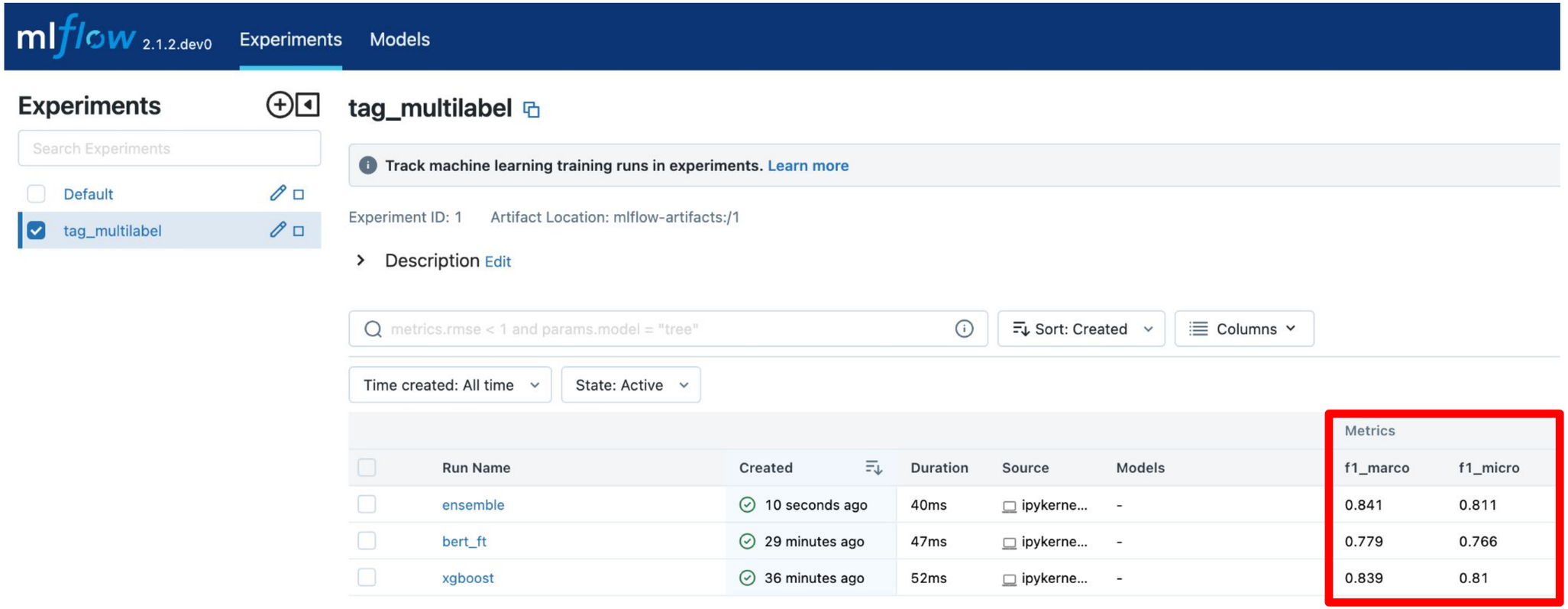
Реализовано:

- VK chat-bot & mini app
- Сервис на FastAPI
  - Хостинг ансамбля моделей по tag и модели NER с BERTopic

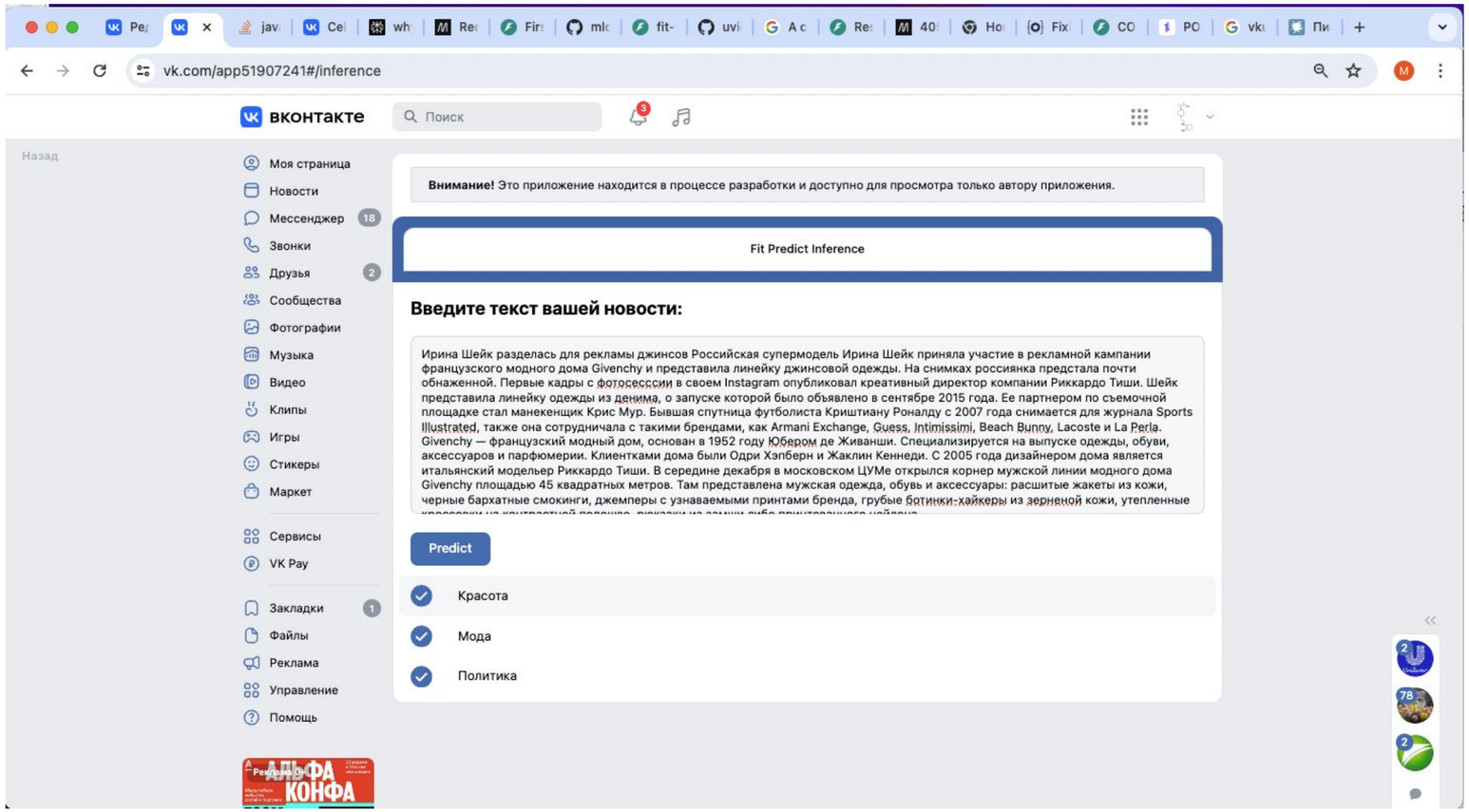




Метрики


mlflow 2.1.2.dev0
Experiments
Models
Experiments
Search Experiments
Default
tag_multilabel
tag_multilabel
Track machine learning training runs in experiments. Learn more
Experiment ID: 1
Artifact Location: mlflow-artifacts:/1
Description Edit
metrics.rmse < 1 and params.model = "tree"
Sort: Created
Columns
Time created: All time
State: Active
Metrics
Run Name
Created
Duration
Source
Models
f1_marco
f1_micro
ensemble
10 seconds ago
40ms
ipykerne...
-
0.841
0.811
bert_ft
29 minutes ago
47ms
ipykerne...
-
0.779
0.766
xgboost
36 minutes ago
52ms
ipykerne...
-
0.839
0.81

Мы не забыли про Персика



Спасибо за внимание!