

Хакатон МИФИ

"Классификация отзывов" о компании Норси-Транс

«Норси-Транс» с 1995 года разрабатывает информационно-аналитические системы, предлагая клиентам отечественные аппаратно-программные решения. В продукцию компании входят: системы информационной безопасности, мониторинга сети, аналитические комплексы и бизнес-платформы.



by Anna Perova & Lab_Story



Хакатон МИФИ, команда 🍁 Lab_Story



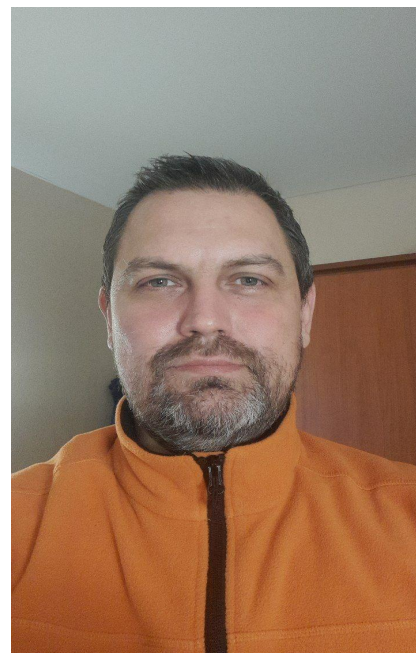
Перова
Анна

Team Lead
ML researcher
Data Engineer
Full-Stack



Иванов
Фёдор

ML researcher
Data Engineer
Backend



Григорьев
Ярослав

ML researcher
Data Engineer
Backend



Лобас
Фанис

ML researcher
Data Engineer
QA

Описание задачи

Создать систему, которая определяет вероятность принадлежности текста к одной или нескольким тематикам из заданного списка. Это задача классификации с пересекающимися классами

Детали:

- Тематики: спорт, юмор, реклама, соцсети, политика, личная жизнь

Критерии:

- PEP8-оформление, docstrings;
- Оформление проекта по шаблону (например, [cookiecutter](#));
- Документация всех этапов проекта;
- Возможность запуска через CLI и предпочтительно Docker;
- requirements.txt с версиями библиотек;
- Документация по сбору своего датасета;
- Метрики (доля пропусков, ложных срабатываний по каждому классу и др.) с интерпретацией;
- Покрытие кода модульными тестами;
- Полнота экспериментов по обучению и валидации;
- Качество предсказаний на закрытых тестовых данных.

Описание задачи

Задача:

Классифицировать короткие тексты (2–30 слов) по тематикам: спорт, юмор, реклама, соцсети, политика, личная жизнь. Текст может относиться к нескольким темам или ни к одной. Это задача многоклассовой и мультилейбловой классификации.

Предобработка текста и EDA

Создание датасета из направленных файлов + синтетические данные

Дано:

1. **6 файлов, без разметки**, по разным темам, темы пересекаются, в тексте опечатки (“эйфелева башня”->”айфилфа”, текст зашумлён, содержит ссылки, смайлы) - необходима качественная чистка данных, чтобы далее использовать для обучения.
Формат .csv
 - данные разделены на 3 столбца doc2text, image2text, voice2text. Большинство материалов - doc2text.
2. **Есть небольшой файл - в формате .doc - примеры ответов** решавших данную задачу - в прошлом файл содержит небольшие отзывы + ответ - к какой теме относится, с наибольшей вероятностью, есть пересечение тем

Решение о работе с датасетом:

1. Взять файл с готовым решением, использовать его за эталонное решение, создать на его основе датасет
2. Взять 6 файлов .csv, представленных в задании, объединить в единый файл, разметить вручную (файлы совсем небольшие, и это несложно)
3. Объединить все файлы вместе

В данном датасете есть данные из doc, image, voice recognition

В 1 версии решения принято решение работать только с текстами, поэтому все остальные данные анализируются как текст

В будущем возможны следующие версии

1. **распознавание изображений** - распознаёт текст с помощью **OCR** (например, Tesseract, EasyOCR)
2. **распознавание голоса** - распознаёт речь через **ASR** (например, Whisper от OpenAI, Vosk, Silero)

Что сделано для подготовки данных

1. Создан эталонный датасет из результатов исследования

- 10 строк - ответы из файла doc
- 210 строк данные датасета, заполненные вручную

2. 6000+ строк неразмеченного датасета

данные из 6 объединённых блоков текста для обучения модели

данные из 4 столбцов объединены в 1 столбец объединённые удалён, так как он избыточен

текст	спорт	реклама	политика	юмор	личная жизнь	соцсети
Отлично провели время. Билеты дарила мужу в подарок на десятилетнюю годовщину свадьбы. Он у меня поклонник стендапа. Но и я получила огромное удовольствие. Все выступления были отличные. Два часа непрерывного смеха. По дороге домой в машине и дома вспоминали и хохотали. Одновременно с нами ходила моя коллега с родственницей, они тоже в восторге. Насмеялись от души. Обязательно пойдём ещё, если будет такое мероприятие.	0	0	0	1	1	0
Стендап классный. Всё понравилось, возможно не все шутки смешные, но это "на вкус и цвет"... Девушки замечательные, открытые, простые, приятные и просто восхитительные и красивые ☺ Молодцы! Шоу, оставило только положительные эмоции. Был с супругой и ей однозначно понравилось, возможно немного больше чем мне	0	0	0	1	1	0
Недавно мы с женой ходили на концерт группы «32nd to Mars» и «Джарда Лета». Билеты на этот концерт я подарил жене на день рождения. Впервые в жизни угадал с подарком.	0	0	0	0	1	0
Успел в Париж только на второй день игр. Токо что паел краусан с кофе. Думаю прошвырнутся по магазинчикам, куплю модных вещей	0	0	0	0	0	0
Решили прогуляться с семьей по торговому центру. Обожаю такие семейные посиделки и все такое. Когда проголодались, вспомнили про Вкус-Вилла. В нем всегда все свежее и полезное	0	0	0	0	1	0

Создан датасет для **многоклассовой тематической классификации текстов**

Датасет

- Использовали **размеченный мини-датасет (210 строк)**.
- Добавили **аугментацию с синонимами** для увеличения выборки (в 2 раза хотя бы).
- Масштабирование данных - 6000 данных из неразмеченного датасета

Модели, которые сравнили

1. **TF-IDF + Logistic Regression (OneVsRestClassifier)**
2. **RuBERT** ([DeepPavlov/rubert-base-cased](#))



RuBERT



Качество модели на мини датасете (210) :

	precision	recall	f1-score	support
спорт	0.0000000	0.0000000	0.0000000	37.0
реклама	1.0000000	0.2222222	0.3636363	54.0
политика	0.0000000	0.0000000	0.0000000	11.0
юмор	0.0000000	0.0000000	0.0000000	26.0
личная жизнь	1.0000000	0.047619	0.090909	42.0
соцсети	0.0000000	0.0000000	0.0000000	35.0
micro avg	1.0000000	0.068293	0.127854	205.0
macro avg	0.3333333	0.044974	0.075758	205.0
weighted avg	0.468293	0.068293	0.114412	205.0
samples avg	0.065728	0.056338	0.059468	205.0

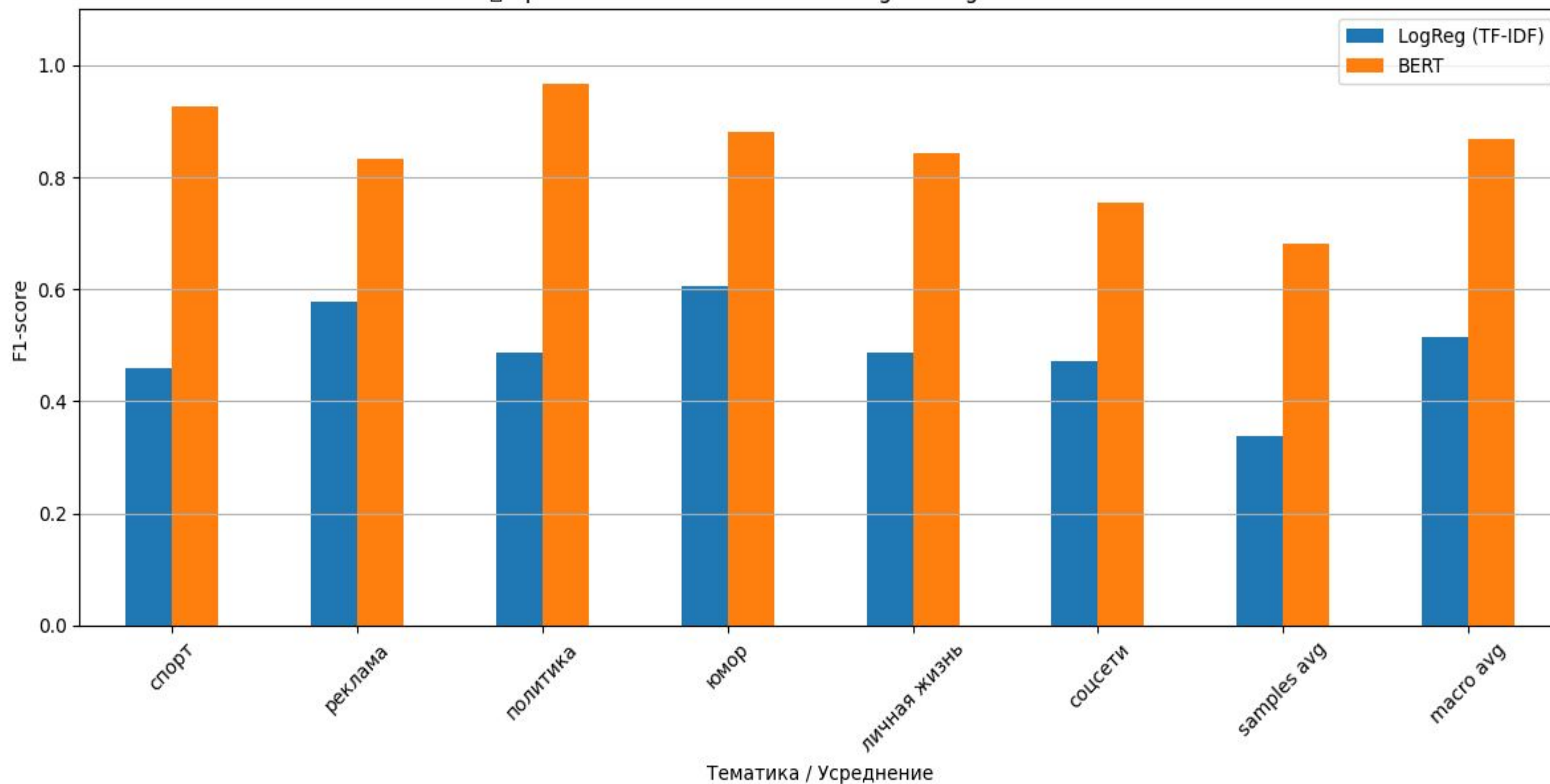
RuBERT



Качество модели на синтетически расширенном датасете:

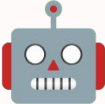
	precision	recall	f1-score	support
спорт	1.000000	0.298246	0.459459	57.0
реклама	1.000000	0.405405	0.576923	74.0
политика	1.000000	0.322581	0.487805	31.0
юмор	1.000000	0.434783	0.606061	46.0
личная жизнь	1.000000	0.322581	0.487805	62.0
соцсети	1.000000	0.309091	0.472222	55.0
micro avg	1.000000	0.350769	0.519362	325.0
macro avg	1.000000	0.348781	0.515046	325.0
weighted avg	1.000000	0.350769	0.517226	325.0
samples avg	0.342342	0.336336	0.338338	325.0


Сравнение F1-score: TF-IDF + LogisticRegression vs BERT



Выводы

 **TF-IDF переобучается и боится предсказывать редкие классы (recall низкий, precision высокий).**

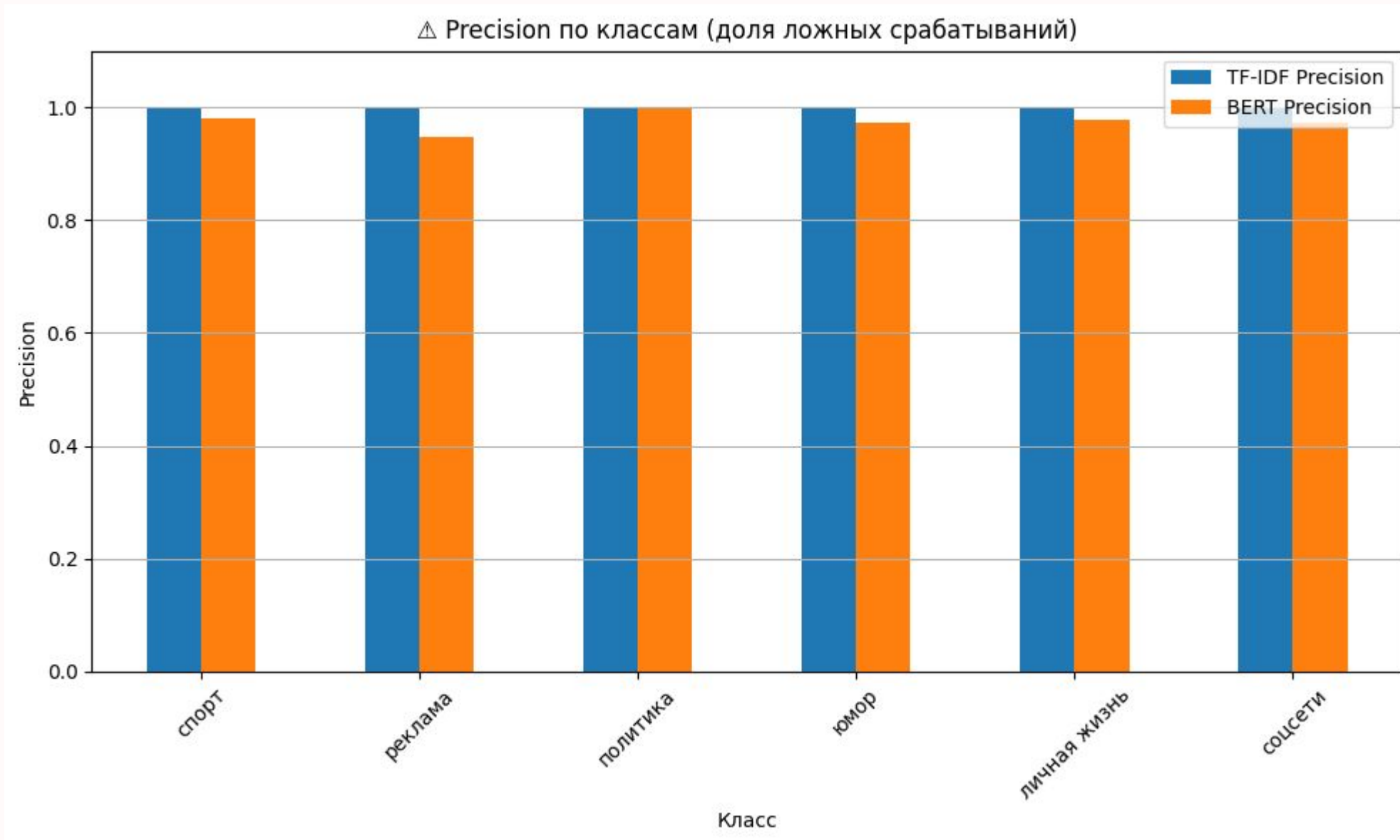
 **BERT значительно лучше обобщает, уверенно ловит больше примеров при приемлемом уровне ошибок.**

 **F1-score — лучший индикатор, и он выше у BERT во всех темах.**

- BERT выигрывает по всем классам.
- Особенно сильно BERT улучшил результаты для "юмор", "политика", "соцсети".

Метрики

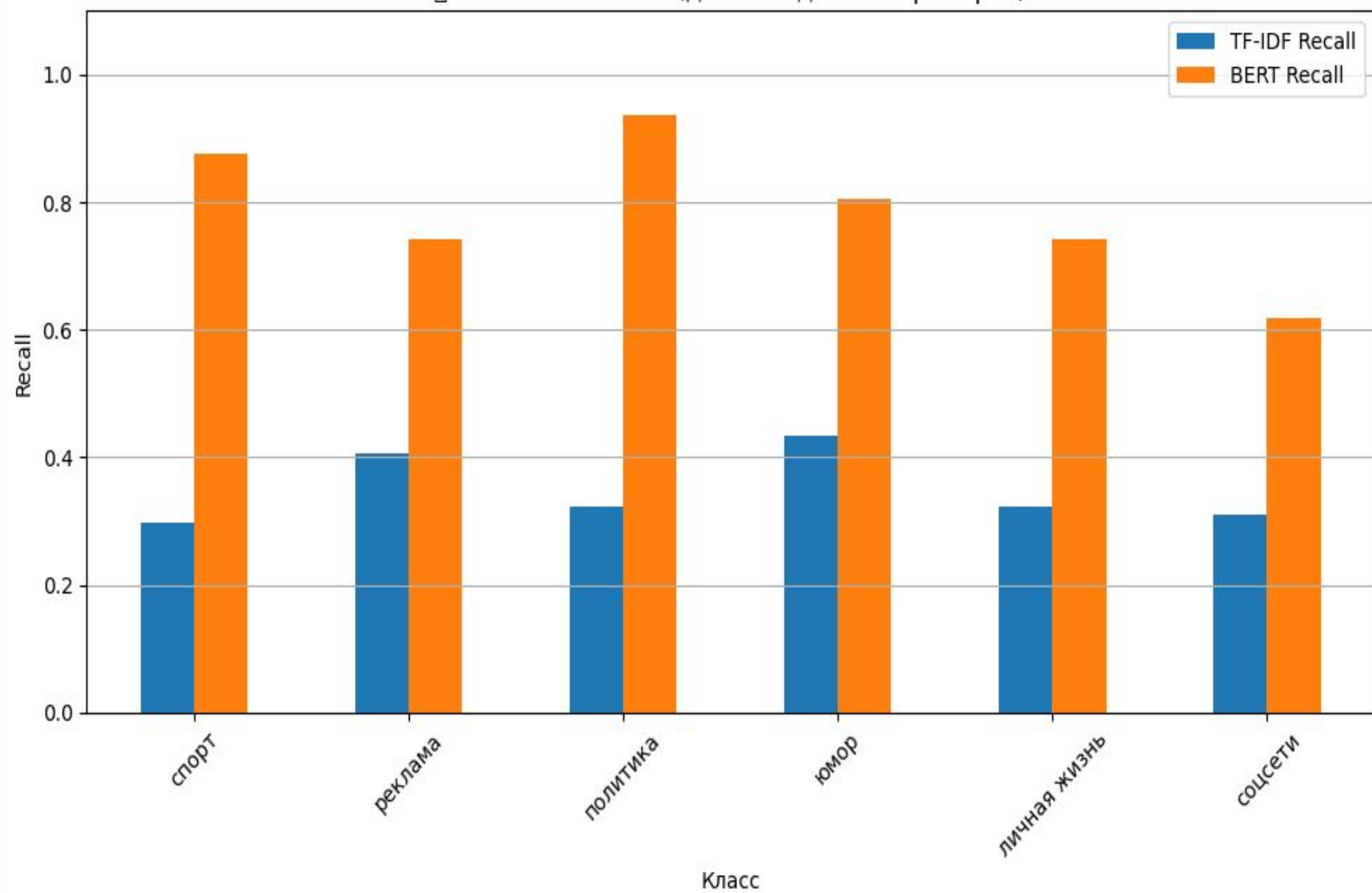
- **Accuracy** (по каждому классу)
- **Precision, Recall, F1** (macro / micro / per label)
- **Classification report**

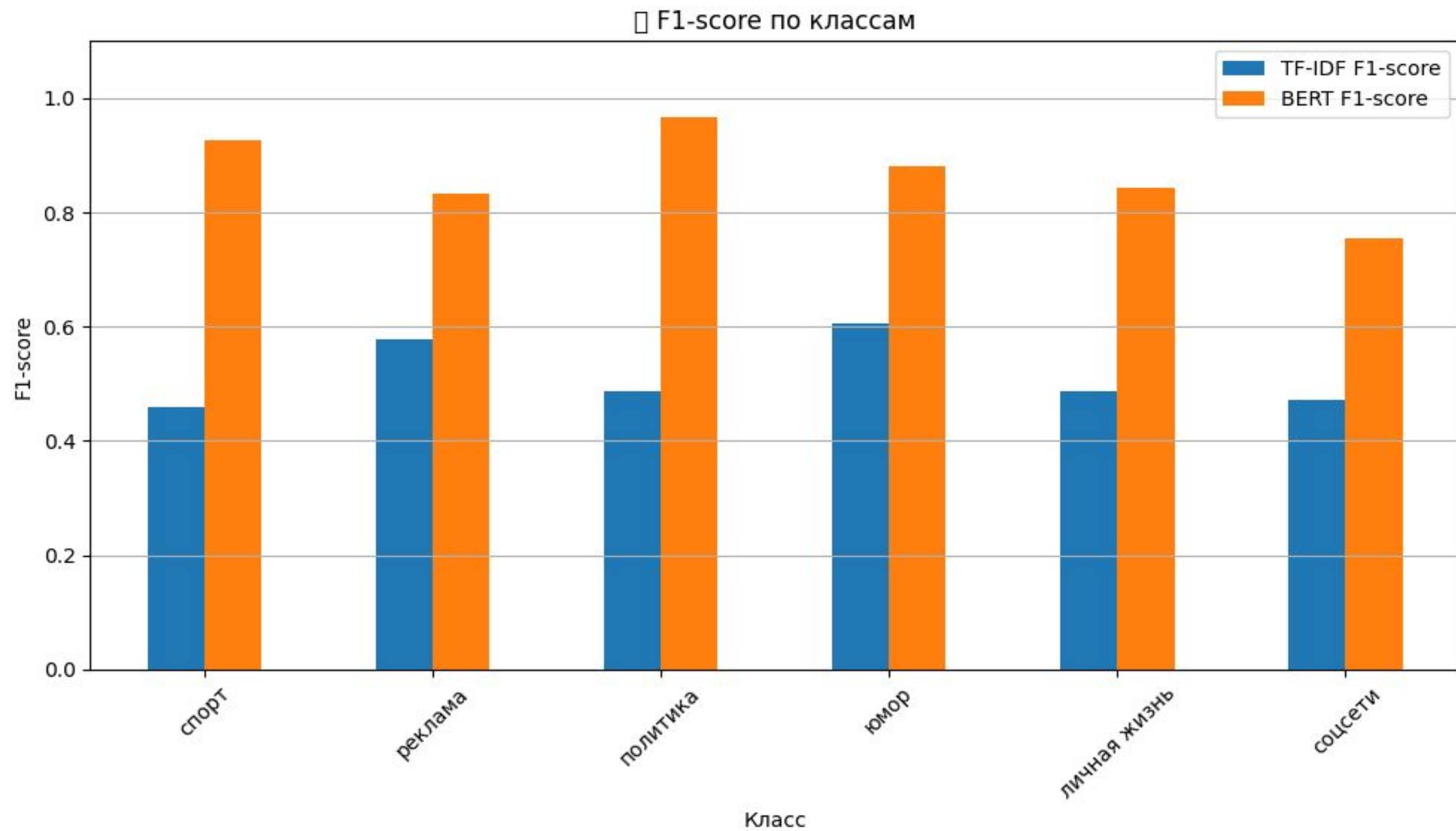


Модель на TF-IDF показывает искусственно высокую точность (precision = 1.0), но это связано с тем, что она почти не делает предсказаний.

Напротив, модель на основе BERT предсказывает активнее (при чуть более низком precision), но обеспечивает более широкий охват и лучшее общее качество.

□ Recall по классам (доля найденных примеров)





Спасибо за внимание!

