Цифровое качество жизни населения по данным социальной сети “ВКонтакте”

Школа прикладного анализа данных для исследователей КМНС


Дунаева Дарья Олеговна

Концепция качества жизни

Цель исследования:
Оценить субъективное благополучие населения в регионах РФ, основываясь на цифровых данных – рассчитать индекс проблем регионов

Индекс проблем
– состояние институтов
1

Индекс среды
– условия окружающей среды
2

Индекс жизненно-стилевых форм
– деятельностное своеобразие жизни
3

2019-2021 гг

Индекс актуальности темы – о чем в сообществах региона пишут чаще всего + какие темы вызывают наибольшие реакции у пользователей?

Индекс субъективного не(благополучия) региона – какие институты в регионе вызывают отрицательные и положительные оценки пользователей?

Методология

Национальный исследовательский Томский государственный университет

I. Реестр региональных сообществ → II. Выгрузка контента региональных сообществ → III. Классификация контента по категориям и тональностям → IV. Расчёт индексов по регионам → V. Аналитика: сравнение регионов, онлайн и офлайн данных

Результат

- Анализ контента региональных сообществ
- Индекс проблем региона

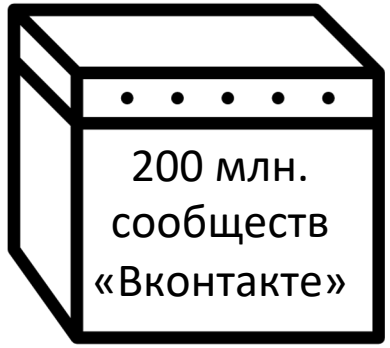
Данные

- 7 тысяч региональных сообществ
- 57 млн активных пользователей
- 10 млн сообщений

Методы

- Машинное обучение:
  Обучающая выборка = ручная разметка 100 тысяч сообщений
- API и WEB-crawling

Реестр региональных сообществ



Критерий №1 – численность подписчиков сообщества не должна превышать 2 млн.

2 млн. - 491 сообщество

Критерий №2 – минимум 50% подписчиков сообщества указали в информации на своей странице населенный пункт, относящийся к одному региону (для Мск и Спб – 80%).

1,5 млн сообществ

Критерий №3 - численность подписчиков сообщества должна быть не ниже порогового значения равного нижнему квартилю от общего числа подписчиков сообществ в каждом регионе.

715 тысяч сообществ

Критерий №4 – сообщества не должны содержать в названии мусорные стоп-слова

52 тысячи сообществ

Критерий №5 – сообщество должно достаточно часто публиковать посты.

25 тысяч сообществ

7692 региональных сообществ

Ручная разметка

Региональные и мусорные сообщества

Национальный исследовательский Томский государственный университет

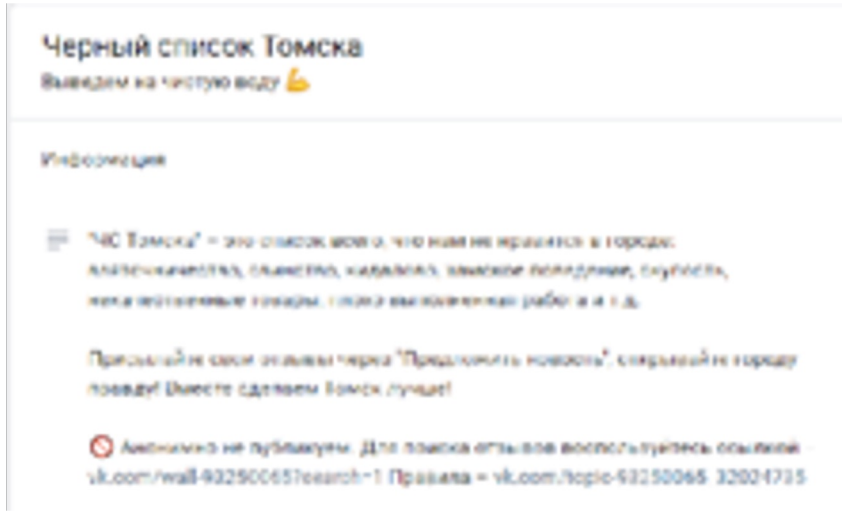
Критерии отнесения сообщества к региональному

- Содержит информацию о социальной, политической или экономической жизни региона;
- У подписчиков есть возможность публиковать посты или оставлять комментарии к постам;

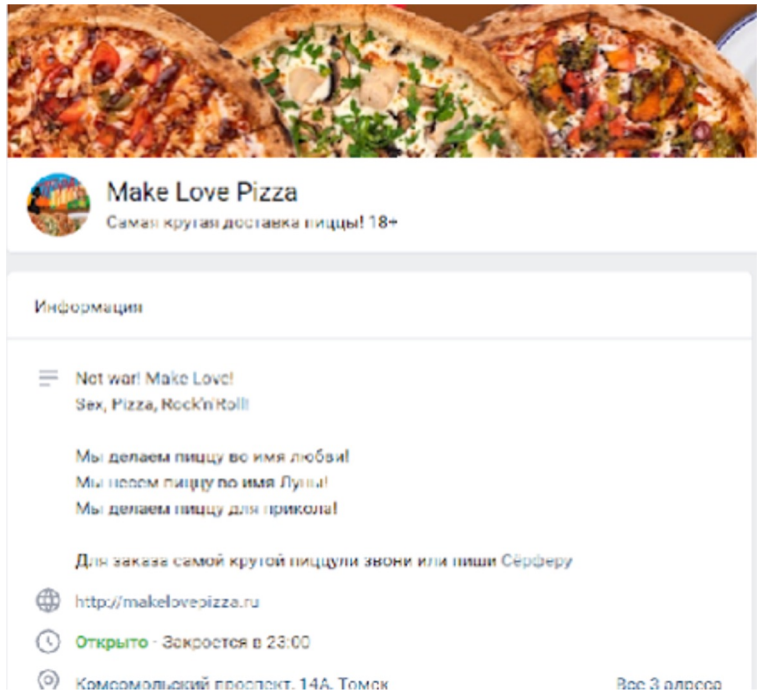
К мусорным были отнесены следующие типы сообществ

- Тематические группы по интересам
- Коммерческие интернет-страницы
- События (культурные, спортивные, жизнь знаменитостей)
- Обмен и б/у продажи от пользователей
- Истории, рассказы, вопросы людей
- Досуг
- Доставка еды
- Здоровье
- Знакомства
- Работа
- Развлекательный контент
- Образовательный контент
- Совместные поездки
- Группы помощи

Пример регионального сообщества:



Пример мусорного сообщества:



Категории качества жизни населения

Национальный исследовательский Томский государственный университет

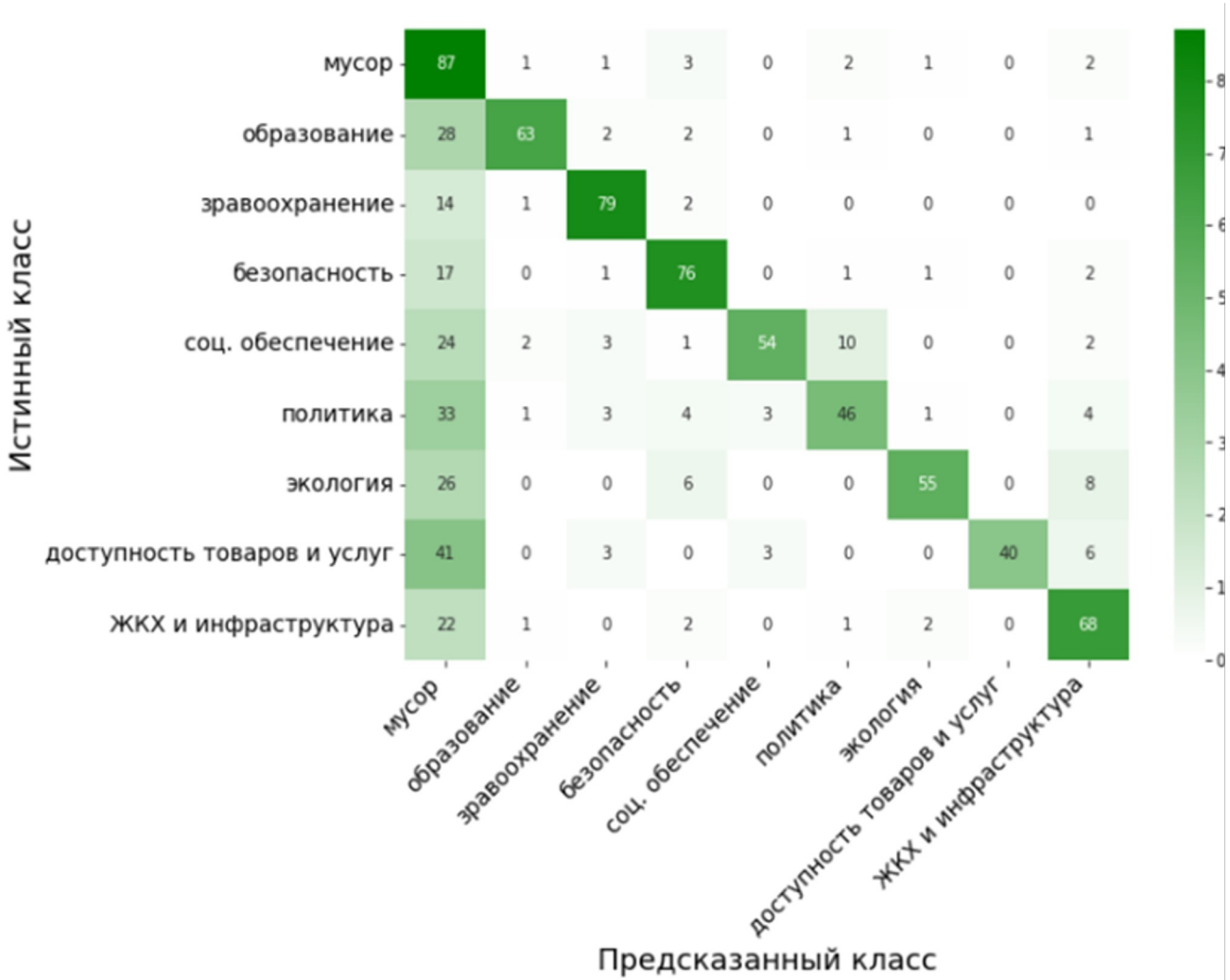
19 категорий → 8 категорий

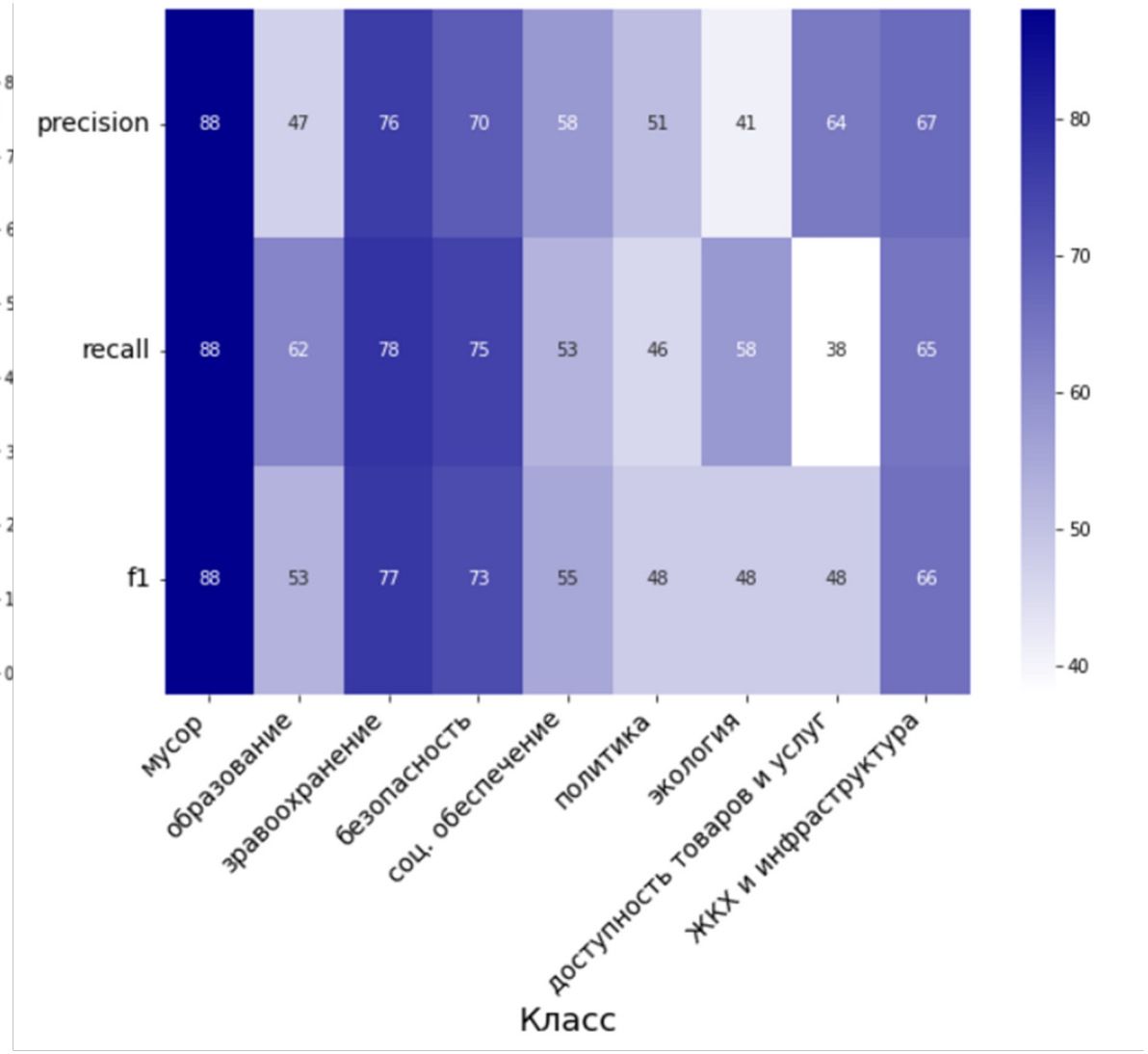
Обоснование укрупнения:

- Категории пересекаются
- Категории собирают мало сообщений
- Упрощение процедуры разметки

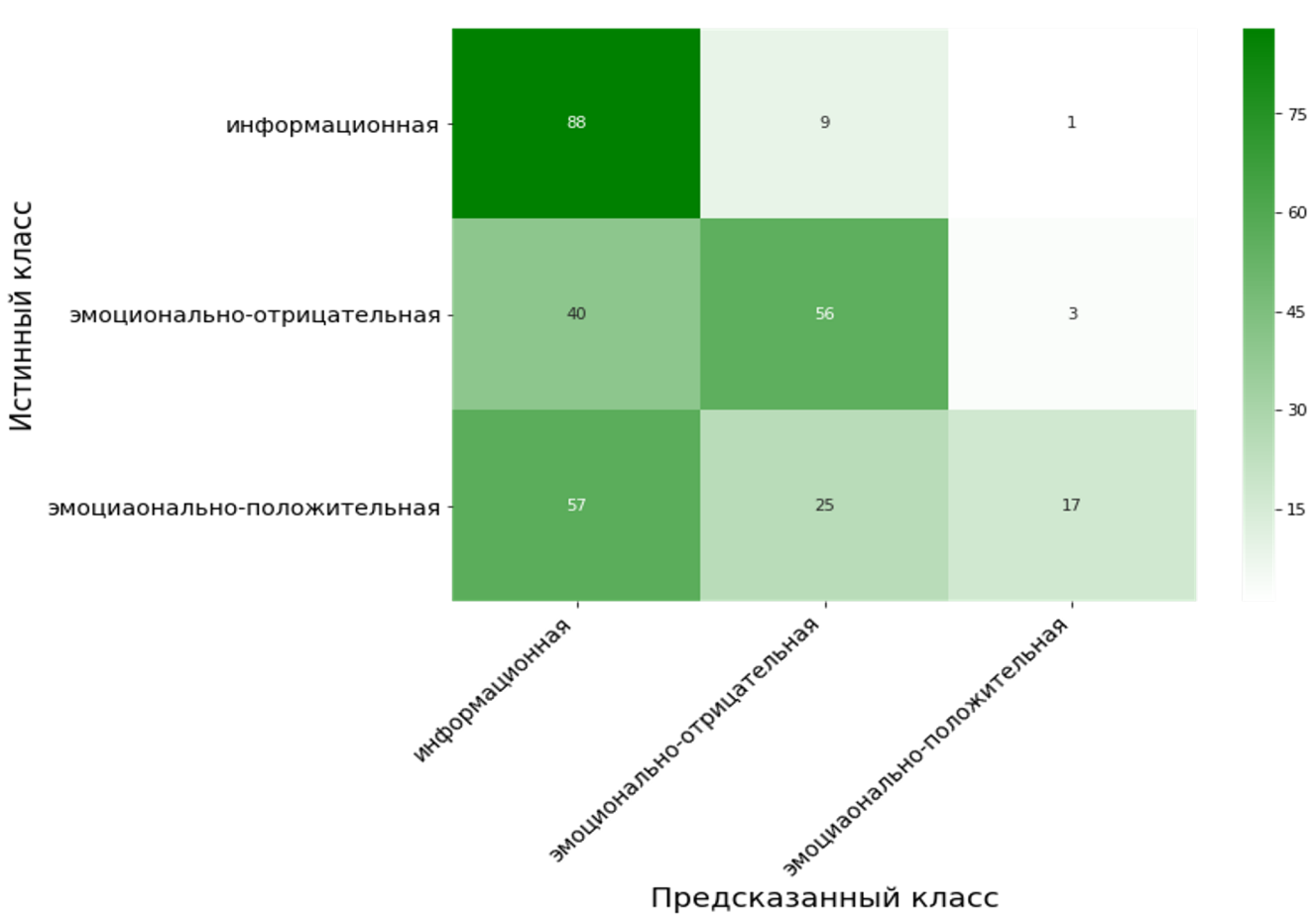
Старая категория (2018-2019 гг)	Новая категория (2020-2021 гг)
1 (Образование)	1. Образование
3 (Медицина)	2. Здравоохранение
5 (Безопасность)	3. Безопасность
13 (Социальная поддержка от государства)	4. Социальное обеспечение
14-19 (Свобода СМИ, Протестный потенциал, Свобода выборов, Отношение к власти, Политические решения, Внутренняя политика)	5. Политика
6 (Экология)	6. Экология
10 (Магазины)	7. Доступность товаров и услуг
2 (ЖКХ), 4 (Инфраструктура)	8. ЖКХ и инфраструктура

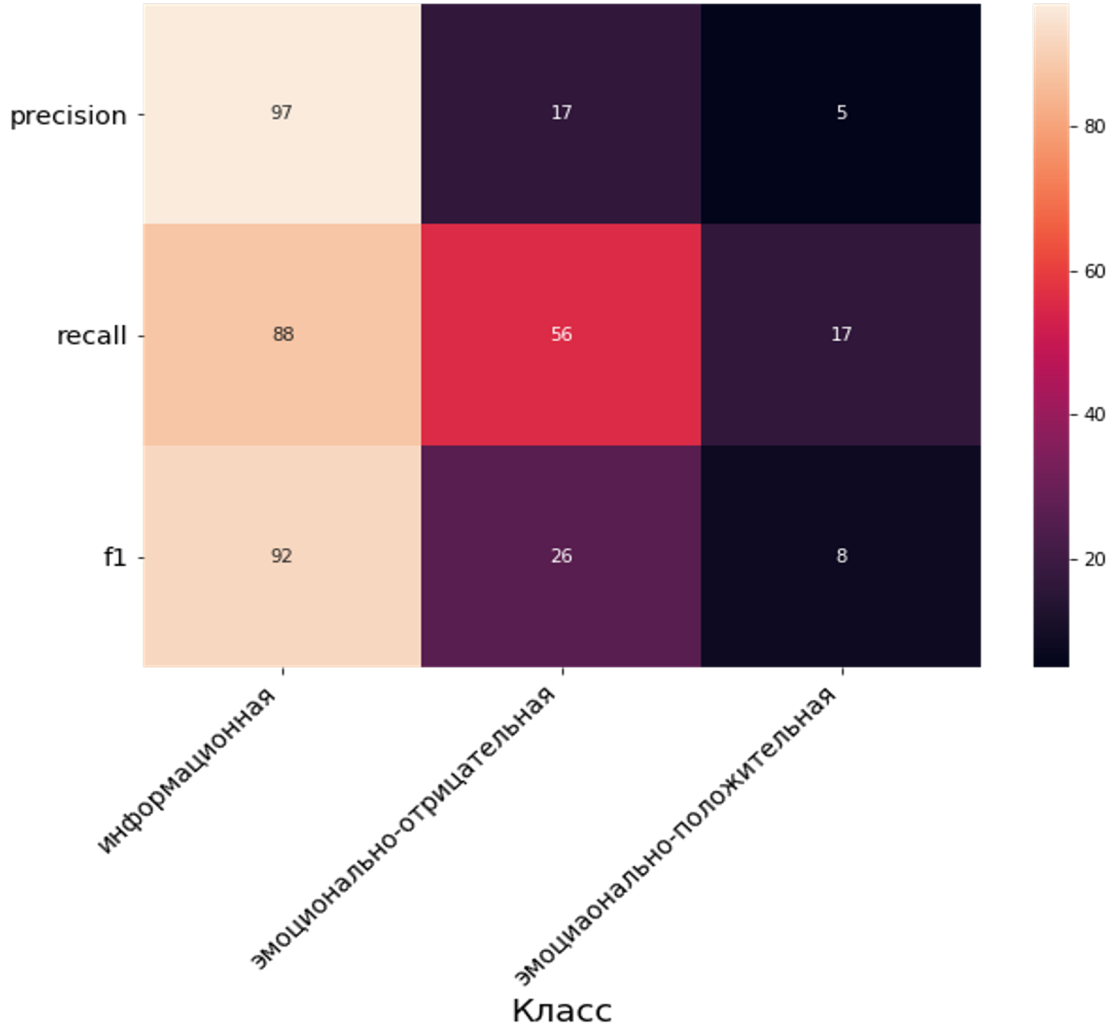
Классификатор rubert-tiny: оценка работы – категории





Классификатор rubert-tiny: оценка работы - тональность

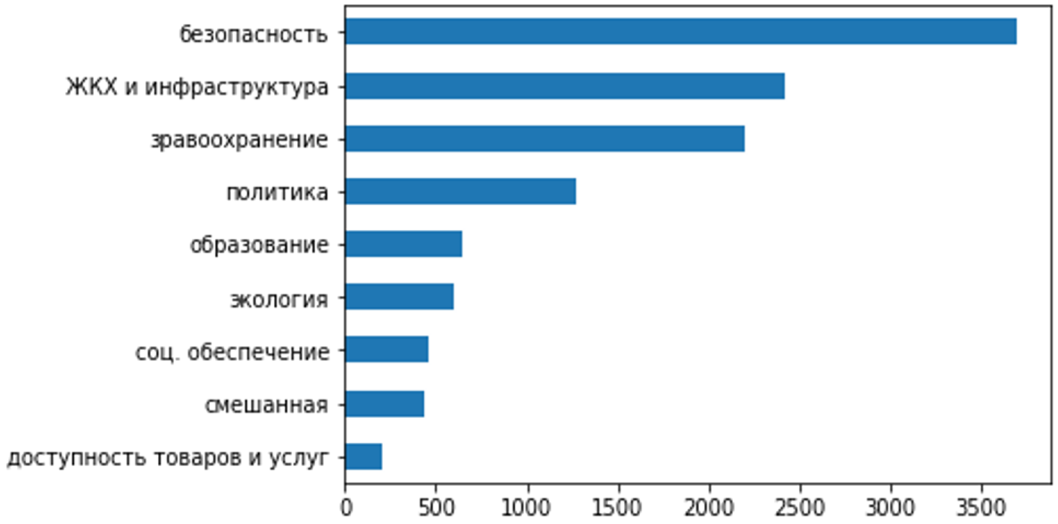




Проблемы

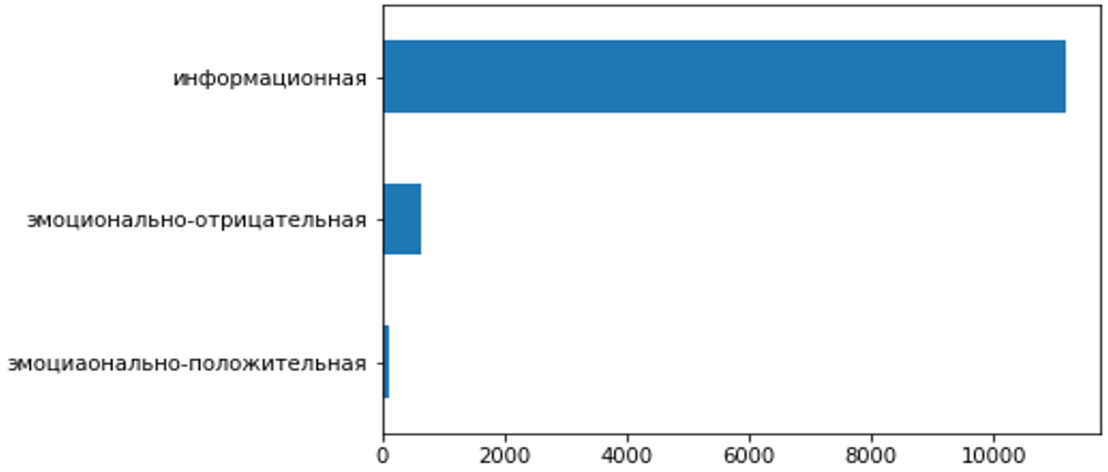
Национальный исследовательский Томский государственный университет

Несбалансированность классов



Решение: ручная доразметка сообщений

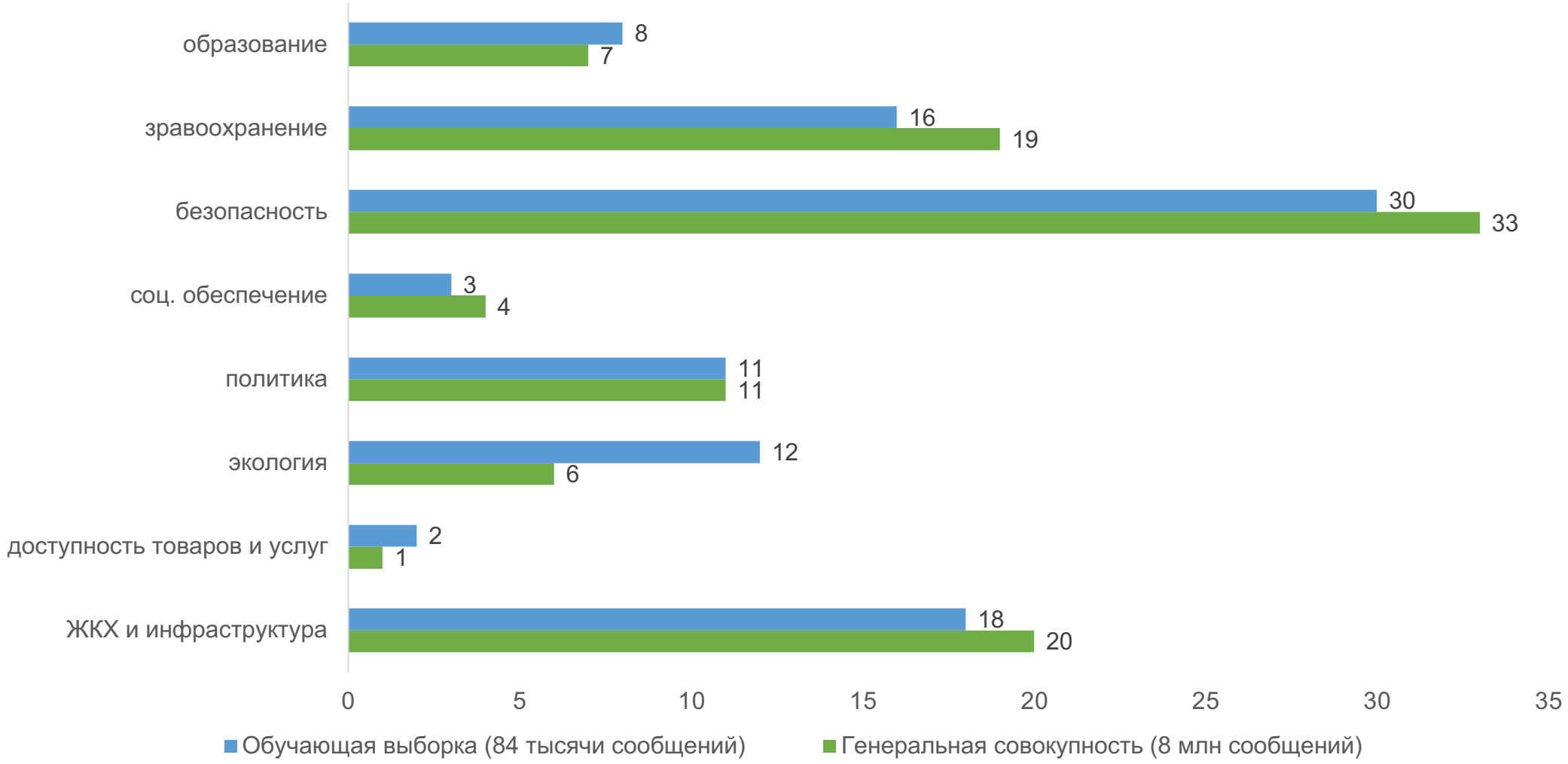
Низкая точность классификации по тональности сообщения



Решение:

- Расширение выборки с помощью готовых классификаторов тональности и доразметка
- Поиск пабликов с преобладанием постов с эмоциональной окраской и доразметка
- Использовать готовую модель для тональности (русские посты твиттер)

Итоговое распределение контента сообществ по категориям



Распределение контента сообществ и цифровых следов по регионам-лидерам (топ-15, **% от общего количества**)

Национальный исследовательский Томский государственный университет

Регион	**Посты**	**Лайки**	**Комментарии**	**Репосты**
Санкт-Петербург	10%	10,30%	14,31%	11,83%
Москва	9%	9,25%	6,79%	3,92%
Башкортостан	6%	3,49%	4,05%	4,36%
Татарстан	4%	1,89%	3,01%	1,46%
Московская область	3%	2,07%	1,18%	1,16%
Челябинская область	3%	4,29%	3,18%	4,59%
Свердловская область	3%	2,90%	2,07%	3,17%
Самарская область	2%	3,79%	4,93%	5,05%
Тверская область	2%	1,43%	0,95%	1,31%
Вологодская область	2%	2,71%	1,79%	3,18%
Нижегородская область	2%	3,80%	3,51%	5,63%
Пермский край	2%	2,47%	2,34%	2,75%
Архангельская область	2%	4,20%	4,14%	4,74%
Новосибирская область	2%	3,13%	5,23%	5,06%
Красноярский край	2%	1,19%	1,37%	1,10%

Спасибо за внимание

Дунаева Дарья, ddo@data.tsu.ru