# Сравнение эффективности дообучения русскоязычных и мультиязычных моделей для решения задачи анализа мнений

Данил Алексеев, Варвара Тютюнникова & Всеволод Маслюков
23.02.2024

# Возможные подходы к решению задачи

- Промптинг больших моделей
  - Zero-shot / few-shot
- Дообучение модели на тематических текстах и на нужную задачу
- Специализированный ковидный англоязычный BERT использовался в решении, занявшем первое место на RuArg-2022
- Из доступных для дообучения декодерных-моделей:
  - Для русского: ruGPT-3.5
  - Мультиязычные модели: gpt2, LLaMA-3.1 (через QLoRA)
- Два аспекта: даст ли использование мультиязычной большой модели прирост по сравнению с специализированно русской моделью; стоит ли того натаскивание модель на определённую тематику.

# Данные RuArg-2022: разбиение и разметка

- Общее количество текстов — 9550
- Каждое предложение имеет 6 меток
- 2 подзадачи: определить позицию (за/против/другое/неактуально) и довод (за/против/нет аргумента/неактуально)
- 3 темы: маски, карантин, вакцины

# Данные RuArg-2022: распределение данных

- *За*: наименьшее количество текстов во всех тематиках
- *Против*: относительно мало текстов, особенно по теме «Карантин» (251 текст)
- *Другое / Нет аргумента*: большее количество текстов
- *Неактуально*: самый большой класс — многие тексты не содержат релевантной информации для анализа

=> набор данных имеет значительный дисбаланс, поэтому класс *неактуально* исключается из оценки (как в задачах по анализу тональности)

# Данные RuArg-2022: распределение данных

| Dataset | Total | Stance | | | Premise | | | Irrelevant |
|---|---|---|---|---|---|---|---|---|
| | | For | Other | Against | For | No argument | Against | |
| *Masks* | | | | | | | | |
| train | 6,717 | 704 | 1,832 | 594 | 339 | 2,451 | 340 | 3,587 |
| val | 1,431 | 148 | 388 | 126 | 62 | 542 | 58 | 769 |
| test | 1,402 | 147 | 401 | 123 | 63 | 523 | 85 | 731 |
| all | 9,550 | 999 | 2,621 | 843 | 464 | 3,516 | 483 | 5,087 |
| *Quarantine* | | | | | | | | |
| train | 6,717 | 587 | 1 341 | 172 | 217 | 1,756 | 127 | 4,617 |
| val | 1,431 | 125 | 290 | 39 | 46 | 369 | 39 | 977 |
| test | 1,402 | 116 | 274 | 40 | 50 | 358 | 22 | 972 |
| all | 9,550 | 828 | 1,905 | 251 | 313 | 2,483 | 188 | 6,566 |
| *Vaccines* | | | | | | | | |
| train | 6,717 | 374 | 866 | 418 | 149 | 1,238 | 271 | 5,059 |
| val | 1,431 | 78 | 183 | 92 | 24 | 282 | 47 | 1,078 |
| test | 1,402 | 75 | 181 | 81 | 21 | 262 | 54 | 1,065 |
| all | 9,550 | 527 | 1,230 | 591 | 194 | 1,782 | 372 | 7,202 |

# Корпуса 1: NoraAlt/Mawqif_Stance-Detection

- *Mawqif: A Multi-label Arabic Dataset for Target-specific Stance Detection*
- 1167 текстов с позициями по теме «вакцина от ковида» на арабском
- 3 метки для позиции: *Against*, *Favor* & *null*
- Дополнительные поля:
  - мотивация позиции: *эксплицитная*, *имплицитная* или *не ясна*
  - наличие сарказма: *да* или *нет*
  - тональность: *негативная*, *позитивная* или *нейтральная*
  - уверенность в позиции, сарказме и тональности: 0–1.0000
  - дата публикации твита

## Корпуса 2: webimmunization/COVID-19-conspiracy-theories-tweets

- Синтетический датасет: 6591 твит по теме «конспирологические теории, связанные с ковидом», сгенерированный GPT-3.5:

  Ex.: *My friend's cousin said he got sick after taking the COVID vaccine. Coincidence, or a sign of something sinister?* #VaccineReactions

- 3 метки для мнения: *deny*, *support* & *neutral*
- Дополнительные поля:
  - 6 меток для разных типов конспир. теорий

    Ex.: *CT5: The Chinese intentionally spread the virus.*

# Корпуса 3: Supakrit65/stance-general-json

- ~3MB реальных твитов на английском языке по 3 подтемам, связанным с ковидом: ‘*school closures*’, ‘*face masks*’ & ‘*stay at home orders*’:

  Ex.: *Identify the stance of tweet*: ‘@Liz_Cheney #NoMasks *needed by real men... we have NO fear... your lies ain't working...*’ on topic of ‘stay at home orders’

- 3 метки позиции: *FAVOR*, *AGAINST* & *NONE*
- Дополнительное поле с *zero-shot* подводкой

# Приведение корпусов к единому формату

- Так как в наших корпусах есть поле только для *stance* (позиция), в RuArg мы не будем учитывать метки для *argumentation* (довод).
- Мы исключим из выборки RuArg тексты с метками –1 во всех трех полях *stance*, так как в наших корпусах есть тексты с нейтральной позицией, но не тематически нерелевантные тексты.
- Мы также будем рассматривать все тексты «под зонтиком» ковида без дополнительного разделения на подтемы, так как в 2 из 3 наших датасетов подобное разделение отсутствует.
- Для русскоязычных моделей тексты наших трех корпусов будут переведены на русский, а для мультиязычных — RuArg и арабский корпус будет переведён на английский.