Сравнение эффективности русскоязычных и мультиязычных моделей в решении задачи анализа мнений

Данил Алексеев, Варвара Тютюнникова & Всеволод Маслюков

07.04.2024

О чем этот проект

- Мы исследовали анализ мнений в текстах на тему ковида-19
- с помощью промптинга больших моделей (YandexGPT, Mistral)
  - zero-shot / few-shot
- и дообучения моделей (BERT, Qwen) на тематических текстах.

Исследовательские вопросы:

- даст ли использование мультиязычной большой модели прирост по сравнению со специализированно русской моделью?
- продуктивно ли «натаскивать» модели на определённую тематику?

Корпус 1: NoraAlt/Mawaqif 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

Корпус 2: RuArg-2022

- Общее количество текстов — 9550
- Каждое предложение имеет 6 меток
- Метки позиции: за/против/другое/неактуально
- Метки довода: за/против/нет аргумента/неактуально)
- 3 темы: маски, карантин, вакцины

Корпус 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

Предобработка и перевод данных I

Все датасеты были приведены к единообразному виду:

- Список из словарей,
- в которых по ключам ‘arb’, ‘eng_tr’, ‘rus’ и ‘label’ лежат переводы на язык X
- и метки мнения, соответственно: 0 — ‘против’, 1 — ‘нейтральная позиция’ и 2 — ‘за’.

Предобработка и перевод данных 2

- *Из RuArg* были взяты только такие тексты, в которых выражена позиция по ровно одной из трех тем,
- т. к. в остальных корпусах тексты имеют только одну метку без разделения по подтемам.
- Арабский датасет был переведён на английский и русский моделью *ModelSpace/GemmaX2-28-2B-v0.1*. Из него мы убрали имена пользователей, которые в исходном датасете были заменены на Mention, что улучшило качество машинного перевода.
- *RuArg* был переведён на английский, а *TweetStance* — на русский при помощи функции *=GOOGLETRANSLATE* в *Google Sheets*.

Обучающая и тестовая выборки

- 10% от объема каждого из трех корпусов были использованы в качестве отдельных тестовых выборок.
- Оставшиеся 90% были объединены.
- Данные в аугментированном корпусе были случайно упорядочены с помощью функции *shuffle* из библиотеки *random*.
- Аугментированный корпус использовался для дообучения моделей BERT и Qwen с соотношением 8:1 объема обучающей и валидационной выборок.

Интересный факт от Соника

В датасете Supakrit65/stance-general-json мы убрали хештеги, но заметили, что метрики инструктивных моделей ухудшились относительно наших предварительных результатов на подвыборке.

Метрики по BERT* с дообучением — baseline

	NoraAlt/Mawaqif Stance-Detection		RuArg-2022		Supakrit65/stance-general-json	
	рус.	англ.	рус.	англ.	рус.	англ.
F1-score	0.69	0.72	0.63	0.66	0.65	0.68

* На русских данных дообучался DeepPavlov/rubert-base-cased, на английских — google-bert/bert-base-cased

Метрики по Yandex GPT (промптинг)

	NoraAlt/Mawaqif Stance-Detection		RuArg-2022		Supakrit65/stance-general-json	
	рус.	англ.	рус.	англ.	рус.	англ.
few-shot F1-score	0.55	0.40	0.37	0.21	0.29	0.16
zero-shot F1-score	0.51	0.50	0.34	0.36	0.28	0.38

Метрики по Mistral (промптинг)

	NoraAlt/Mawaqif Stance-Detection		RuArg-2022		Supakrit65/stance-general-json	
	рус.	англ.	рус.	англ.	рус.	англ.
few-shot F1-score	0.52	0.45	0.42	0.27	0.43	0.23
zero-shot F1-score	0.51	0.58	0.46	0.5	0.35	0.48

Какие промпты мы использовали: английский

<s>[INST]You are a knowledgeable AI model who is an expert on COVID-19. Please examine the statement in the context below after the word "STATEMENT:". Output "STANCE: 2" if the author of the statement has a positive stance towards measures put in place by governments to combat COVID-19, such as lockdowns, mask mandates, and vaccination campaigns. If this statement is neutral in sentiment, output "STANCE: 1". Output "STANCE: 0" if the statement is critical of these measures. Output only "STANCE: " and then a number. Do NOT output anything else.
<опциональный блок с примерами для few-shot>
STATEMENT: <пример>[/INST]</s>

Какие промпты мы использовали: русский

<s>[INST]Ты — модель ИИ, которая является экспертом по теме COVID-19. Пожалуйста, изучи утверждение в контексте ниже. Выведи "ПОЗИЦИЯ: 2", если автор утверждения положительно относится к мерам, принимаемым правительствами для борьбы с COVID-19, будь то карантин, требование к ношению масок и вакцинация. Если данное утверждение является нейтральным, выведи "ПОЗИЦИЯ: 1". Если же в утверждении содержится критическая по отношении к этом мерам позиция, выведи "ПОЗИЦИЯ: 0". Выводи только "ПОЗИЦИЯ:" и число. НЕ выводи ничего другого.
<опциональный блок с примерами для few-shot>
УТВЕРЖДЕНИЕ: <пример>[/INST]</s>

Метрики по Qwen-1.5B-Instruct (дообучение)

	NoraAlt/Mawaqif Stance-Detection	RuArg-2022	Supakrit65/stance-general-json
F1-score	0.36	0.4	0.36

- Параметры обучения: iter = 1000, число дообучаемых слоёв — 4, размер батча — 2.

Наблюдения над результатами

- Во многих случаях при промптинге *few-shot* способствует улучшению качества только для русских данных.
- Мультиязычная модель (Mistral) даёт лучший результат на английском, русскоязычная (YandexGPT) ведёт себя примерно одинаково на обоих языках.

Как результаты соотносятся с предшественниками

- Для анализа мнений дообучение BERT-а оказалось более продуктивным, чем промптинг инструктивных моделей.
- Поэтому надо делать так, как наши предшественники.

Участник	Базовая модель Трансформер	Дополнительные данные	$F_{1stance}$-мера	$F_{1premise}$-мера
camalibi	covid-twitter-bert-v2	+	0.70	0.74
sevastyanm	ruRoBERTa-large	+	0.68	0.72
iamdenay	ruRoBERTa-large	+	0.67	0.66
ursdth	Conversational ruBERT	-	0.66	0.71