Data Statement for RuBia

1 Header

Dataset Title: RuBia (A **Ru**ssian Language **Bia**s Detection Dataset)

Dataset Curator(s): Veronika Grigoreva, Anastasia Ivanova, Ilseyar Alimova, Ekaterina Artemova

Dataset Version: 1.0

Dataset Citation:

Data Statement Authors: Veronika Grigoreva, Anastasia Ivanova, Ilseyar Alimova, Ekaterina Artemova

Data Statement Version: 1.0

Data Statement Citation and DOI:

Links to versions of this data statement in other languages:

2 Executive Summary

RuBia is a bias detection dataset for the Russian language. It consists of sentence pairs with the first sentence reinforcing a potentially harmful stereotype and the second contradicting it. There are 4 domains of social bias presented in RuBia (gender, nationality, socio-economic status, LGBTQ+), each of the them is further divided into multiple fine-grained subdomains. Sentence pairs were first written by volunteers and then validated by Russian speaking crowdsourcing workers (except for the LGBTQ+ domain). Overall, there are 2019 unique sentence pairs spread over 19 subdomains in RuBia.

3 Curation Rationale

Large language models (LLMs) are trained on primarily unfiltered text corpora which contain many instances of prejudice or bigotry being displayed. This dataset's purpose is to measure an extent to which Russian language LLMs are affected by the biases present in the training data.

Structure RuBia comprises sets of examples, where each example is made up of two sentences. The first one is always reinforcing a particular harmful social trope or stereotype (pro-trope sentence), while the second one contradicts it (anti-trope sentence). Each example belongs to one of the four domains (gender, nationality, socio-economic status, LGBTQ+). The two sentences differ only by the subject social group; although this difference may be expressed in multiple words because of a rich morphology of Russian language.

Each bias domain is further subdivided into subdomains, which either correspond to a certain way the data is collected (e.g., sentences following template "All __blank__ are __blank__") or to a certain way bias may be displayed (e.g., sentences, describing men in the first sentence of the pair and women in the second sentence of the pair in professional context). The full list of subdomains with descriptions and examples is given in the project's repository.

Data collection Each subdomain has its own specific crowdsourcing task (or tasks) to collect examples. Each task consists of two messages: with the first one asking a user to come up with a pro-trope sentence and the second one asking them to change some aspect of it (pronouns, subject's profession, etc.) to make an anti-trope sentence.

We have distributed the crowdsourcing tasks via the popular in Russia Telegram messenger, warning potential respondents that (i) we are conducting research, (ii) the questions may contain sensitive or triggering material, (iii) participation is voluntary, unpaid, and anonymous, (iv) collected responses would be processed and made publicly accessible. Thus, our method relies on people motivated to complete the task without financial reward. In our opinion and experience, this leads to people approaching the sentence-writing task creatively and allows for a wider coverage of different manifestations of bias. On the other hand, we sacrifice the overall reliability of data collected on this step, since it may contain irrelevant, grammatically incorrect or malicios answers.

The collected sentences were then validated by paid crowdsourcing workers. Each subdomain has its own specific crowdsourcing task (or tasks) to annotate examples, and the validation questions covered multiple ways in which examples can be unsuitable (as opposed to binary "good example" or "bad example") and carefully constructed to minimally rely on users' knowledge of social terminology.

To summarize, our pipeline focused on maximal variablity and coverage during the collection step and removing low-quality examples (and thus, increasing the overall quality of data) on the validation step.

4 Documentation for Source Datasets

No source datasets were used.

5 Language Varieties

The dataset consists of sentences written in Russian using the Cyrillic script as used in Russia (ru-Cyrl-RU according to BCP-47).

6 Speaker Demographic

The data was collected via bot in Telegram messenger, which was sent in multiple group chats and channels (for example, student community chats) and asked several people to share it further. In order to make speakers feel free to write potentially harmful stereotypes and ensure their anonymity, demographic data was not collected. However, speakers had to read instructions and write sentences in Russian, so all answers collected via bot were written by the speakers of Russian language.

7 Annotator Demographic

The data was annotated by the Russian speaking users of Toloka. Toloka provides a standard set of information about annotator demographic (age, gender, education level, residence place, language knowledge), which is collected and verified by the platform at the moment the user registers on Toloka website. To prove Russian language knowledge, users should have to take a Toloka Russian language exam.

- Age: 19-81 (average: 39.8, median: 39)
- Gender:
  - 416 men,
  - 344 women,
  - 1 non-binary person
- Country:
  - Russia (757 annotators),
  - Ukrain (27 annotators),
  - Belarus (17 annotators),
  - Kazakhstan (19 annotators),
  - Moldova (2 annotators),
  - Uzbekistan (2 annotators)
- Education level: basic education, higher education
- Languages:
  - only Russian (301 annotators),
  - Russian and English (97 annotators),
  - Russian, Ukranian and English (22 annotators),
  - Russian, Ukranian (12 annotators),
  - Russian, English and German (11 annotators)

  - another combination of languages (362 annotators, from 1 to 6 annotators with the same combination of languages)

- Proficiency in Russian: all annotators passed Toloka Russian language exam

- Number of different annotators represented: 805

- Relevant training: no training was needed for the annotation task

8 Speech Situation and Text Characteristics

- Time and place of linguistic activity: Telegram messenger (data collection), crowdsourcing platform Toloka (data annotation)

- Date(s) of data collection: November-January 2021

- Date(s) of data annotation: April-May 2023

- Modality: written

- Scripted/edited vs. spontaneous: spontaneously written and edited during the annotaion process

- Asynchronous interaction

- Topic: a wide range of direct and indirect social stereotypes about gender, nationality, socioeconomic status and LGBTQ+.

- Non-linguistic context: absent

9 Preprocessing and Data Formatting

Anonymization procedures. Users' text responses were first stored with corresponding chat IDs (chat session identifiers, unique for specific chat session) and no other user information was gathered. Then, before the validation step, all text responses were compiled into a dataset table and chat IDs were dropped. Moreover, during validation no responses containing private information were found. Thus, no information that can identify or reveal individual people was included in the final dataset.

Data preprocessing. Before rule-based validation and human annotation, the raw data underwent preprocessing. Punctuation marks were mostly removed, except for commas, as they can significantly alter the meaning of a sentence in the Russian language. Excessive whitespace characters were removed, the initial letters of sentences were capitalized, and dots were added at the end of every sentence. All steps of data preprocessing are implemented in this notebook.

10 Capture Quality

The data may contain types and and slang expressions that were accidentally not deleted during the dataset validation.

11 Limitations

Choice of domains and subdomains. Our choice of biases is specific to Russian social context and may be different from other cultures and language enviroments. Future works, which would like to re-use our annotation protocols, should revise the choice of domains and subdomains.

Demographics. The diversity of participants may be limited, as the experiment was advertised in select Telegram chats. The data collection protocol keeps the anonimity, so we cannot present any demographic statistics of respondees.

Shortcoming of sentence-pair format. Many bias displays are hard to measure using the pro-trope sentence/anti-trope sentence format. It is especially evident in biased contexts where the subject's group is not referenced directly, does not have a context-appropriate counterpart, or where not the prescribed attribute itself, but a reason for prescribing this attribute indicates bias. We suggest that work examining LLMs' bias on the level of discourse rather than individual phrases is needed.

12 Metadata

License: The dataset is distributed under the Creative Commons Attribution 4.0 International

(CC BY 4.0) license license.

Annotation Guidelines: Data Collection pipeline can be found in the project's repository.

Annotation Process: Data validation instruction can be found in the project's repository.

13 Disclosures and Ethical Review

Data collection. The crodwsourcing strategy used in this paper utilizes the Telegram platfrom. The respondees, who participated in the data collection, were warned about potentially sensitive nature of the task and that they would not receive any financial compensation.

Annotators' compensation Three Toloka workers annotated each RuBia example, with varying payment rates ranging from an average of \$1.4/hr to \$2.4/hr depending on the task pool's complexity and number of questions. These rates exceed the hourly minimum wage in Russia.

Potential risks. We recognize that the dataset may be used to cause harm if employed in bad faith. It contains displays of bias against several groups and can, in theory, either be used for online harassment directly or be used to fine-tune a model capable of online harassment. However, we believe that putting the dataset online will not have any significant negative social impact, as the dataset's contents are sparse and limited (intended for evaluation and not training) and, by design, lack any meaningful metada. As such, we doubt that this dataset will be suffient for creating a model that can purposefully, meaningfully and maliciously reproduce bias.

14 Other

15 Glossary

- *LLM* - large language model