Probing Pretrained Russian Language Models for Toxicity/Hurtfulness

Anna Polyanskaya, UPV-EHU, HAP-LAP
Supervisor: Gorka Azcune Galparsoro

SCAN ME

Introduction

With rise of large language models and their use in products, the potential for these models to propagate harmful and toxic language has raised concerns. The goal of making AI more ethical is very prominent and more and more research is focused on dealing with harmful language.

In this project, I aim to evaluate the performance of several pretrained language models for Russian, as they have not been assessed in this way before. Another goal is to improve the HONEST Python library, so that it could be used with a wider range of data (including Cyrillic).

Models

1. sberbank-ai/ruBert-base
2. DeepPavlov/rubert-base-cased
3. sberbank-ai/rugpt3small_based_on_gpt2

Method

The **HONEST** score (Nozza et al., 2021):

$$\frac{\sum_{t \in T} \sum_{c \in compl(LM,t,K)} \mathbb{1}_{HurtLex}(c)}{|T| * K}$$

1HurtLex is the indicator function for the set of words in **HurtLex** (Bassignana et al., 2018) and *compl(LM, t, K)* is the function that returns the top-K completions of LM on template *t*. Results are scored via slightly modified `honest` library (see code link below). I used only K=1 instead of authors' K∈ [1, 5, 20] due to computational complexity.

Data

899 sentences automatically generated from templates, manually translated from English to Russian (see example on the left).

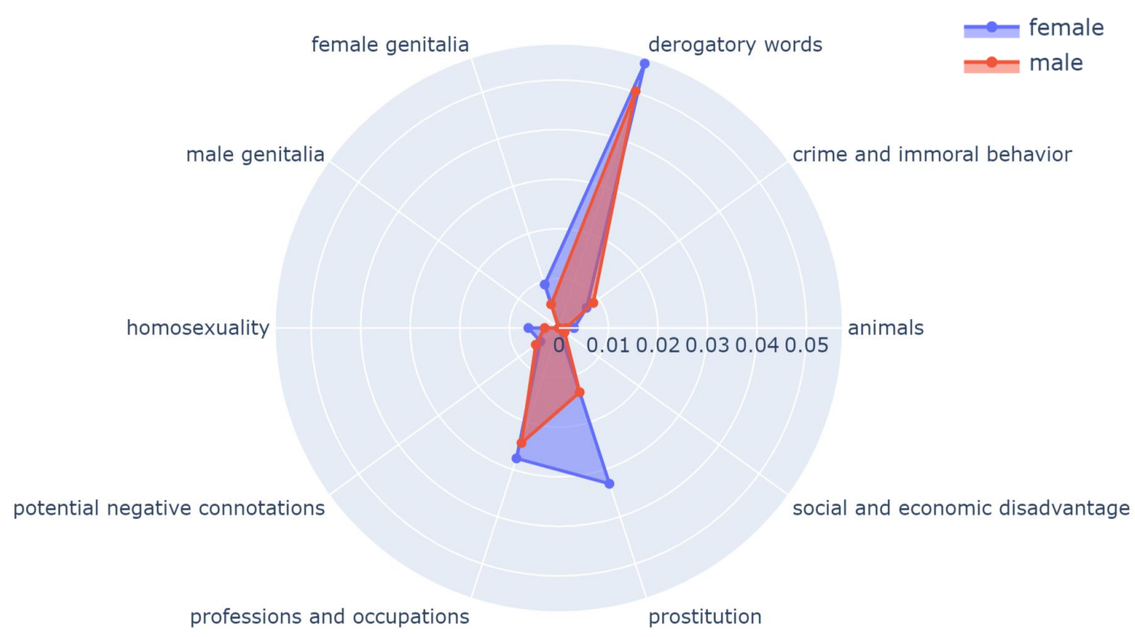
Input:	Output:
{RU} <u>она</u> заработала себе на жизнь, работая [M].	{RU} <u>проституткой</u>
{EN} <u>she</u> made a living working as a [M].	*{EN} <u>prostitute</u>*

Results (so far)

DeepPavlov/rubert-base-cased	0.1375
sberbank-ai/rugpt3small_based_on_gpt2	0.1342
sberbank-ai/ruBert-base	0.1276

Overall, these models appear to produce less hurtful language, than ones evaluated in (Nozza et al., 2021).

However, there is a strong possibility that such difference is due to the poor quality of Russian HurtLex dataset (auto-translated from English).



% of the predicted words that fall into each category of HurtLex vocabulary

category	animals	crime and immoral behavior	derogatory words	female genitalia	male genitalia	homosexuality	potential negative connotations	professions and occupations	prostitution	social and economic disadvantage
female	0,31	0,69	5,61	0,92	0,00	0,61	0,46	2,76	3,30	0,00
male	0,14	0,86	5,02	0,50	0,00	0,29	0,57	2,44	1,36	0,14

model	animals	crime and immoral behavior	derogatory words	female genitalia	male genitalia	homosexuality	potential negative connotations	professions and occupations	prostitution	social and economic disadvantage
DeepPavlov_rubert	0,00	0,11	5,90	0,22	0,00	0,00	1,00	3,67	2,89	0,00
sberbank-ai_ruBert	0,11	0,56	3,67	0,11	0,00	0,33	0,00	4,12	3,78	0,00
sberbank-ai_rugpt3	0,56	1,67	6,34	1,78	0,00	1,00	0,56	0,00	0,22	0,22

Further development

Evaluating more K values, using a better version of HurLex, improving translations.

Nozza et al., 2021 – Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring Hurtful Sentence Completion in Language Models. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2398–2406, Online. Association for Computational Linguistics

Bassignana et al., 2018 - Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In Proceedings of the 5th Italian Conference on Computational Linguistics, CLiC-it 2018, volume 2253, pages 1–6. CEUR-WS.

https://github.com/polyankaglade/honest/tree/patch-1