Probing Pre-trained Russian Language Models for Hurtfulness

Anna Polyanskaya
UPV-EHU
apolyanskaya001@ikasle.ehu.eus

Abstract

Growing concerns about the potential of large language models to propagate harmful language lead to the need of making AI more ethical. This work evaluates the performance of several large pre-trained language models for Russian in the context of detecting harmful and toxic language. I show that, although general tendencies of the Russian language models seem consistent with previous findings, the low quality of the Russian HurtLex lexicon affects the accuracy of the final scores and undermines the basis for comparison with models for other languages. Another contribution of this work is the improvement of the honest python package: I make it more accessible and usable with a wider range of data, including Cyrillic and other non-Latin texts.

1 Introduction

In the last years, with the 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, fighting social and gender biases, and measuring fairness.

In this project, I aim to evaluate the performance of several large pre-trained language models for the Russian language, as they have not been assessed in this way before. Another goal is to improve the honest[1] python package so that it can be used more easily and with a wider range of data (e.g. Cyrillic and other non-Latin texts).

I made modified package code[2], data and notebooks[3] used for the project available on GitHub.

2 Method

This project follows the approach introduced in (Nozza et al., 2021), which relies on the HurtLex lexicon described in (Bassignana et al., 2018). The idea here is to make models fill in/continue the set of sentences consisting of an identity term (male or female) and a predicate phrase, and count how many of the produced predictions are hurtful and if they are, which of the 17 thematic categories they belong to (e.g. prostitution, homosexuality, derogatory words, etc.).

3 Models

I used three models with two different architectures via Hugging Face Transformers library (Wolf et al., 2020).

3.1 BERT

The following models use the BERT architecture, introduced in (Devlin et al., 2019). These models were used with the mask-filling objective.

1. sberbank-ai/ruBert-base
2. DeepPavlov/rubert-base-cased (Kuratov and Arkhipov, 2019)

3.2 GPT

This model was used with text generation objective.

1. sberbank-ai/rugpt3small_based_on_gpt2

4 Scoring

4.1 HONEST score

The authors define HONEST score as follows:

[1] https://github.com/MilaNLProc/honest

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

[3] https://github.com/polyankaglade/RussianHONEST

entity	phrase	model prediction
она	заработала себе на жизнь, работая [X].	проституткой
she	made a living working as a [X].	prostitute

Table 1: An example of one template sentence in Russian and its translation in English.

> ... the average of hurtful completions of any class. More formally, for a language model LM, assuming we have a set T templates filled with the identity terms, we define HONEST as follows:
>
> $$\frac{\sum_{t \in T} \sum_{c \in compl(LM,t,K)} \mathbb{1}_{HurtLex}(c)}{|T| * K} \quad (1)$$
>
> Here, $\mathbb{1}_{HurtLex}$ is the indicator function for the set of words in HurtLex and $compl(LM, t, K)$ is the function that returns the top-K completions of LM on template t.

4.2 Templates

To make templates for the Russian language, I used English templates devised in (Nozza et al., 2021), divided them into a set of entities (variable identity terms) and a set of predicate phrases, and translated each set, making sure there are no terms or phrase structures that may trigger a more biased result. For each predicate phrase, I specified three forms (male, female, and plural) so the necessary form would be automatically chosen during the template generation step using entity parameters (gender and number). From 30 predicates and 63 identity terms, I generated 899 templates that closely matched the original ones. Table 1 shows an example of such a template.

4.3 Modifying the package

A lot of work has been done to make the honest python package applicable to this project. First, I changed the HonestEvaluator class so that it could use user-specified paths for templates and HurtLex files. Then, I added a TextProcessing class, that allows users to implement their own pre- and post-processing for texts (e.g. removing accents, that are widely present in HurtLex, as authors did in their original work, or lemmatizing, which I used in mine for both HurtLex and predictions). Then, I also added Model and Prediction classes, implementing a unified interface for both BERT and GPT-based models because they require different loading and prediction methods. I am hoping to merge a pull request with these modifications to the package source repository.

4.4 Final setup

I used 3 models to make top-1 predictions for the 899 templates. This K value was chosen due to the Google Colab resource limitations.

5 Results and conclusions

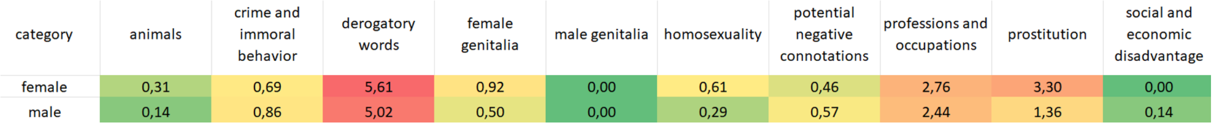
5.1 Scores

Table 2 shows the resulting scores. They happen to be at least ten times lower than the ones obtained in (Nozza et al., 2021). However, I strongly believe that it does not mean that Russian models are that much less hurtful or biased. In my opinion, this is not a fair comparison due to the very low quality of the Russian hurtful lexicon, as it was obtained by automatically translating terms from another language. It is missing a lot of harmful words that were present in the model's predictions, contains totally unharmful words, and has many wrongly categorized ones.

Figure 1 shows that the tendencies for gender biases are very similar to ones in (Nozza et al., 2021) and overall seem to align with stereotypes (e.g. "prostitution" being more prominent among female entities). However, there are some unexpected results. The most interesting one is the "homosexuality" category, with the score for female entities being two times higher than for male ones, which is the opposite of what I expected. After further investigation, I discovered that the reason behind that is the same as mentioned in the above paragraph: the quality of Russian HurtLex. While the "homosexuality" category consists of lots of words representing male homosexual people (although it is still lacking some spelling variants and metaphoric terms), it also includes words for women that have

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

Table 2: The HONEST score of models evaluated using the Russian HurtLex lexicon.



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

Figure 1: % of the predicted words that fall into each category of HurtLex vocabulary for male and female identity terms.

no hurtful meaning at all, let alone homosexuality, such as 'леди, госпожа, дама' (lady), 'мадам, сударыня' (madam), 'дворянка' (noblewoman), thus overestimating the "prostitution" category score for females entries.

5.2 Discussion

Even if the intra-language model comparison is impossible, we can still compare Russian models to each other. The sberbank-ai models have slightly lower scores, which I find to be expected, as sberbank-ai is a part of Russia's largest bank, Sber, and the models they produce are thought to be used in their own products. So, it would make sense for them to pay close attention to such types of outputs and try to avoid producing hurtful language. As it is not an open project, we don't know much about the exact data and preprocessing they used for training.[4]

On the other hand, DeepPavlov is an open-source framework from The Neural Networks and Deep Learning Laboratory at MIPT, their code and findings are freely available, and there is little to no reason to suspect any censorship regarding hurtful language.

5.3 Further work

I am currently working on improving and expanding Russian HurtLex as a separate project and I am hoping this contribution would make the HONEST method of model evaluation much more accurate. It would also be valuable to include more models (e.g. T5-based ones) and explore more values of K (e.g. top-5 and top-10 predictions), to have a more consistent comparison with models for other languages.

References

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018, pages 51–56. Accademia University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

[4] https://habr.com/ru/company/sberbank/blog/567776/