PAN 2024 Multilingual TextDetox: Exploring Different Regimes For Synthetic Data Training For Multilingual Text Detoxification

Notebook for PAN at CLEF
Sushko Nikita, 2024

Introduction

Toxic Input	Detoxified Output
What a f**k is this about?	What is this about?
А н**рена ты здесь это писал?	А зачем ты здесь это писал?
Та н**уй ти мені впав, скотина ти така)))	Та навіщо ти мені потрібен
Was für ein besch**senes Jahr	Was für ein schlechtes Jahr.
Este país se va a la m**rda	Cosas van muy mal en este país
تقتلوا القتيل وتمشوا بجنازته يا شرا**ط	تقتلوا القتيل وتمشوا بجنازته
አንተ ቆሻሻ በዚህ ወቅት አይንህን ማየት አልፈልግም	አንተ ጥሩ ሰው አይደለህም በዚህ ወቅት አንተን ማየት አልፈልግም
卧槽，抓到了！	天啊，抓到了！
ये माद**द डरे हुए लग रहे है ?	ये लोग डरे हुए लग रहे है ?

Model selection

- Detoxification is sequence to sequence task, thus, we were selecting a model with encoder-decoder architecture
- mT0 is a perfect model, since it combines multilinguality with instruction tuning
  - mT5 and umT5 showed worse results after testing
  - Aya-101 was too big to fit into our GPU

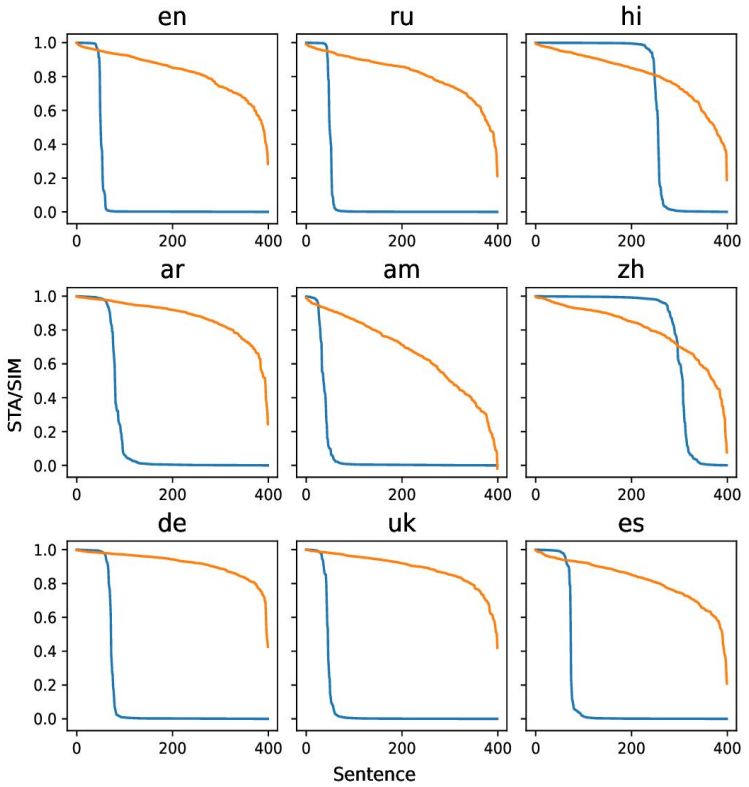
Evaluation Metrics

Four metrics were used for the evaluation:

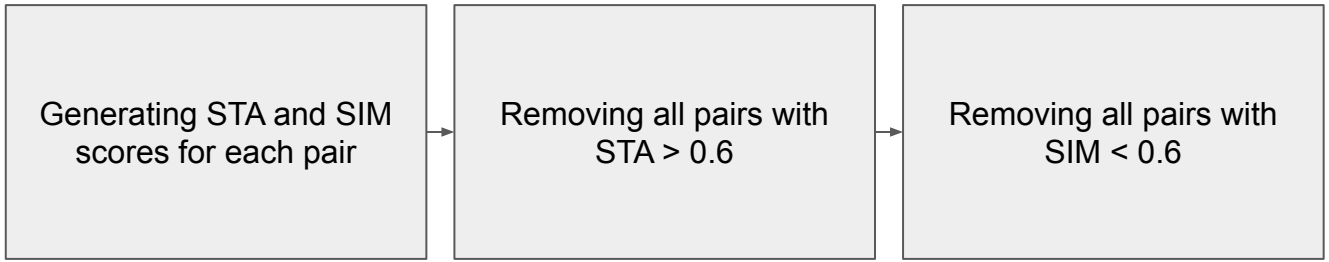
- STA metric measures the overall toxicity of a sentence.
- SIM metric measures the semantic similarity between the original and detoxified sentence.
- ChrF_1 metric measures how natural sounding is the text.
- J metric is a multiplication of STA, SIM and ChrF_1.

Real “dirty” data

Original data was of bad quality, judging by the amount of non-detoxified “neutral” examples and in pairs and amount of dissimilar examples.

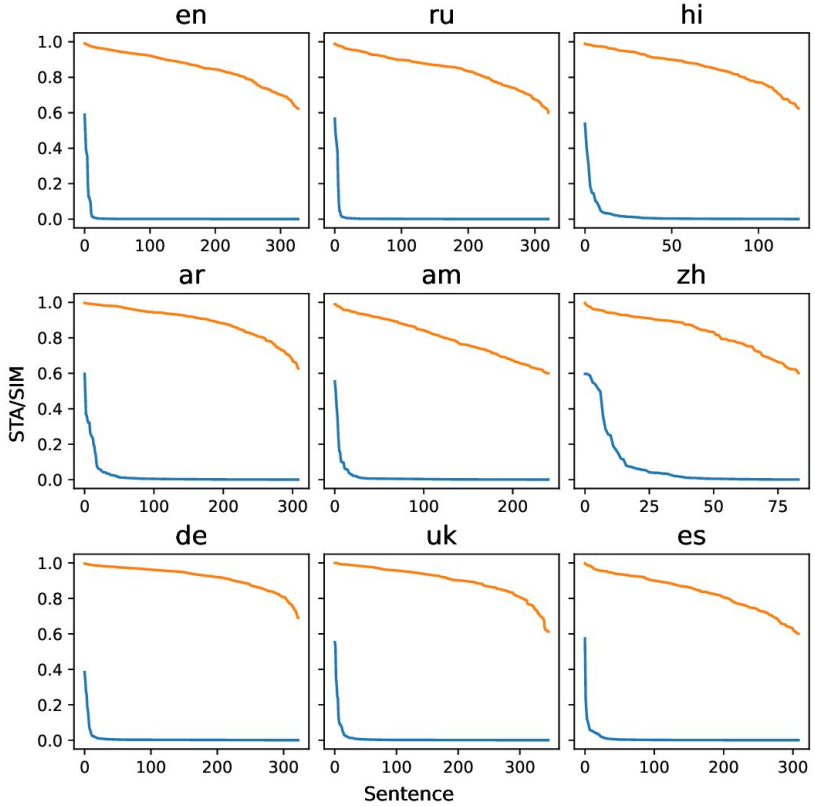


Data cleaning pipeline

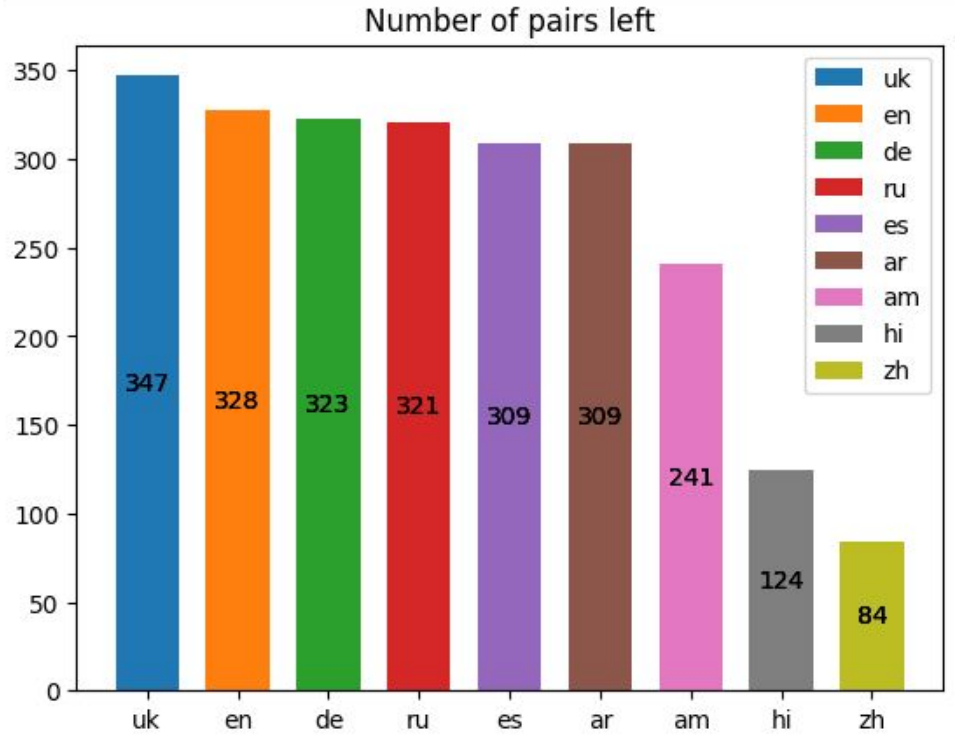


Real “cleaned” data

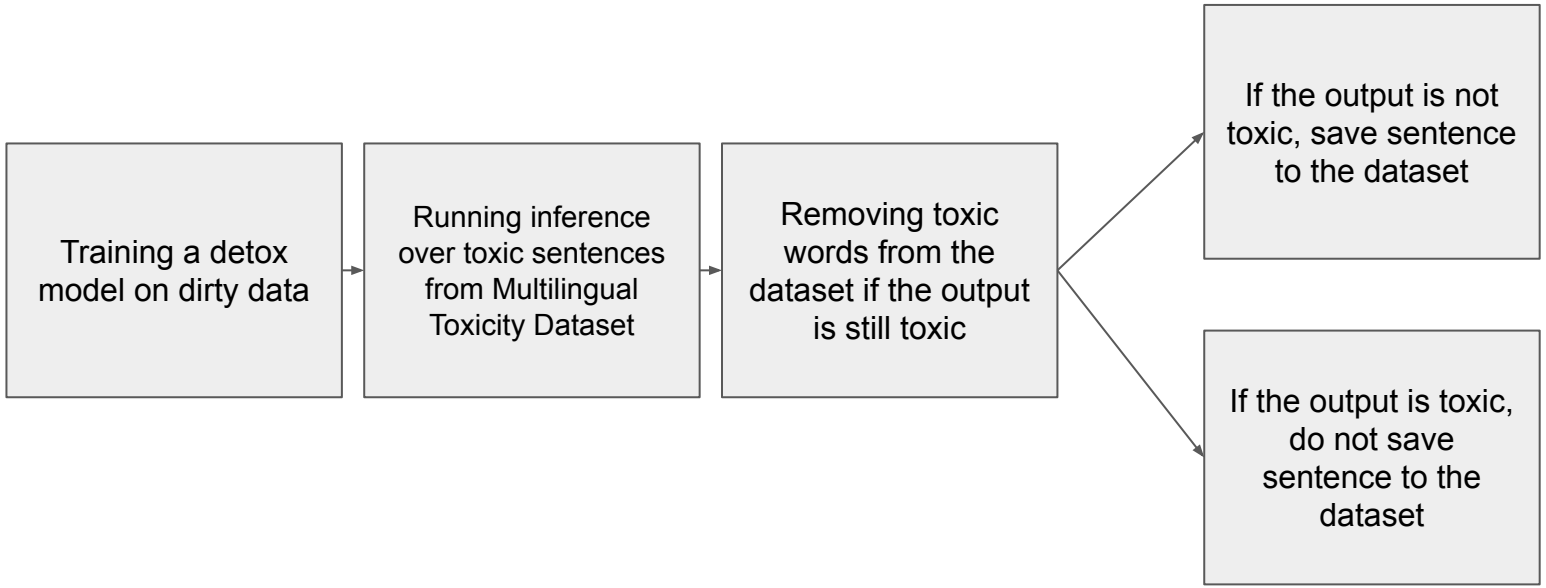
Cleaning the data from dissimilar and toxic pairs drastically improved the quality of the dataset.



Real “cleaned” data

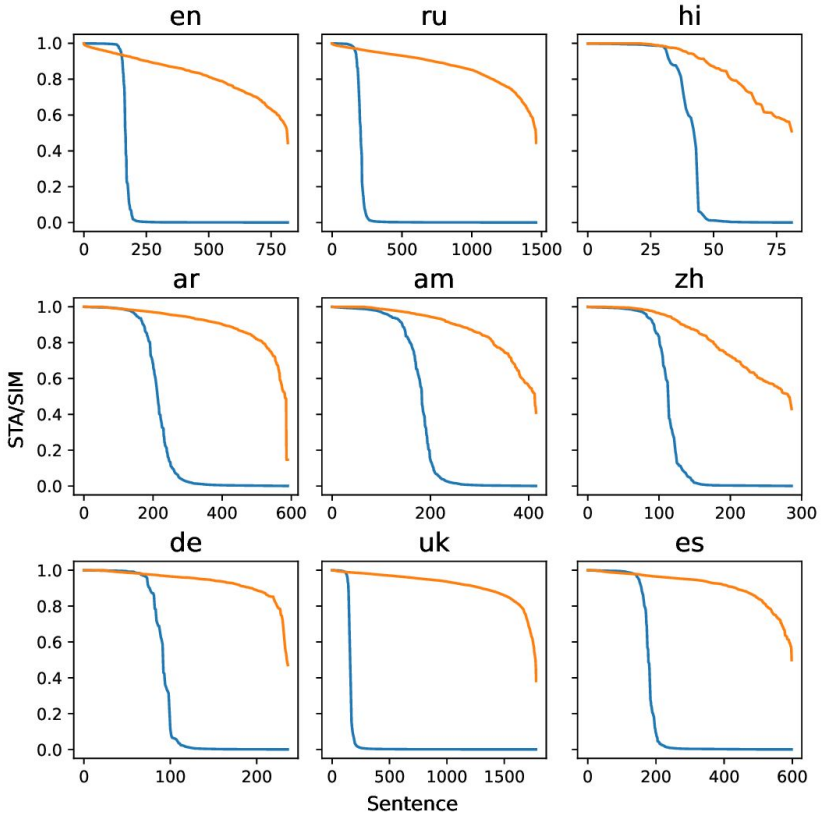


Synthetic data generation pipeline



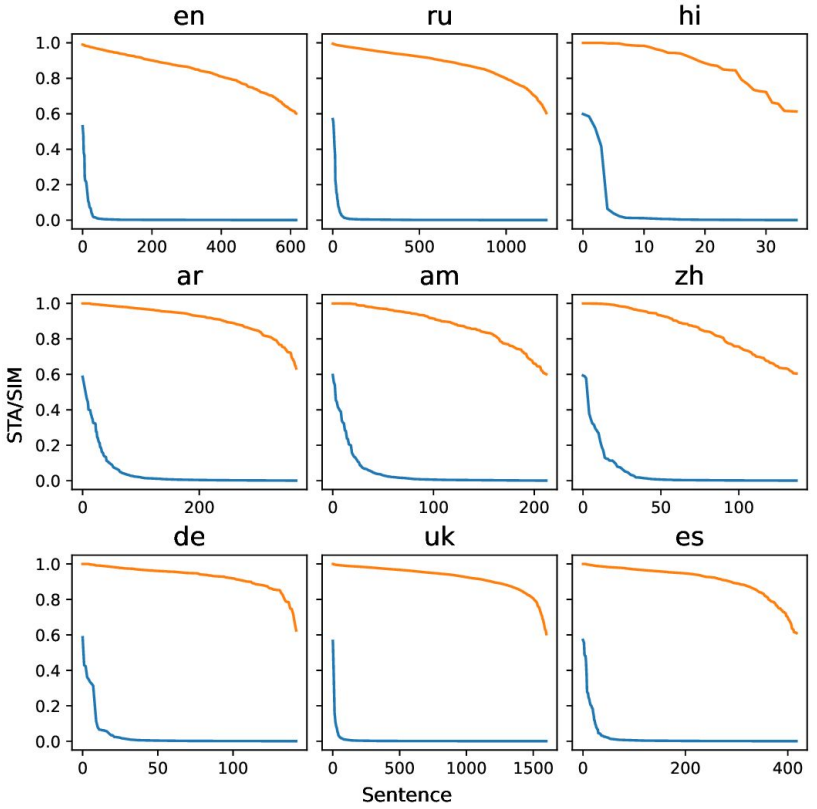
Synthetic “dirty” data

Same cleaning procedure is required after generating synthetic data.



Synthetic “clean” data

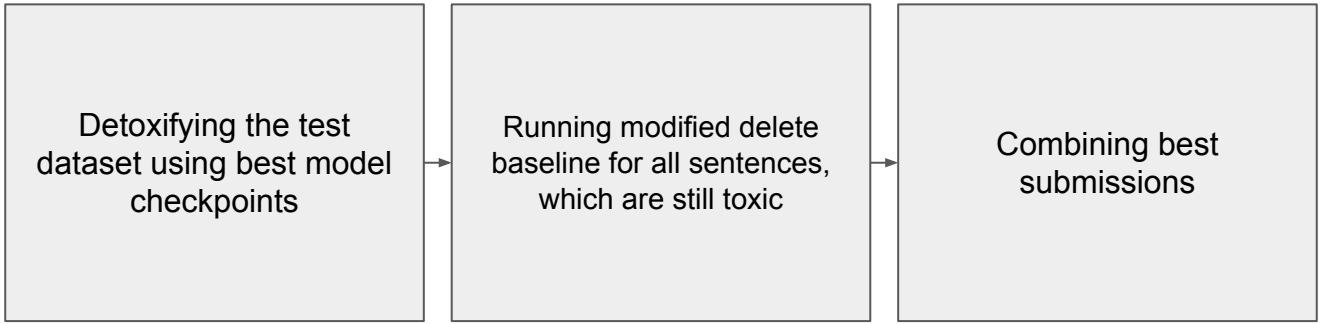
After cleaning the synthetic data we are left with additional high quality examples on “bad” languages of the original dataset.



Training regimes

Regime	STA	SIM	CHRF	J
Dirty Real	0.64	0.89	0.70	0.41
Dirty Synth	0.69	0.88	0.65	0.41
Dirty Real + Synth	0.68	0.85	0.66	0.41
Dirty Synth + Real	0.68	0.90	0.69	0.43
Dirty Mixed	0.7	**0.92**	0.69	0.44
Cleaned Real	0.71	0.90	0.72	0.477
Cleaned Synth	0.71	0.90	0.66	0.437
Cleaned Real + Synth	0.72	0.87	0.71	0.44
Cleaned Synth + Real	0.73	0.88	0.68	0.454
Cleaned Mixed	**0.74**	0.89	**0.73**	**0.481**

Final solution pipeline



Examples of detoxification

Language	Toxic input	Neutral output
English	almost as f[redacted]d up as the cia funding and arming bin laden.	Almost as bad as the cia funding and arming bin laden.
German	Die wollen von nix ablenken. Die sind naturgemäß so d[redacted]m.	Die wollen von nix ablenken. Die sind naturgemäß so naiv.
Spanish	Que m[redacted]a es esa ubicación millenial ahora estoy re enojado loko	Que mala es esa ubicación ahora estoy re enojado
Ukranian	Б[redacted]ть, у[redacted]ю н[redacted]й в Острог і не вертаюсь. в[redacted]у	Уже йду в Острог і не вертаюсь.
Russian	дело даже не в iq - просто х[redacted]м там не место	дело даже не в iq - просто плохим людям там не место

Automatic evaluation results

User	average	en	es	de	zh	ar	hi	uk	ru	am
adugeen	0.523	**0.602**	**0.562**	**0.678**	**0.178**	**0.626**	**0.355**	**0.692**	**0.634**	**0.378**
lmeribal	0.515	**0.593**	**0.555**	**0.669**	0.165	**0.617**	**0.352**	**0.686**	**0.628**	**0.374**
nikita.sushko	0.465	**0.553**	0.480	**0.592**	**0.176**	**0.575**	0.241	**0.668**	**0.570**	**0.328**
VitalyProtasov	0.445	0.531	0.472	0.502	0.175	0.523	0.320	0.629	0.542	0.311
erehulka	0.435	0.543	0.497	0.575	0.160	0.536	0.185	0.602	0.529	0.287

Manual evaluation results

User	average	en	es	de	zh	ar	hi	uk	ru	am
Human References	0.85	0.88	0.79	0.71	0.93	0.82	0.97	0.90	0.80	0.85
SomethingAwful	0.77	0.86	**0.83**	**0.89**	0.53	0.74	**0.86**	**0.69**	**0.84**	**0.71**
adugeen	0.74	0.83	0.73	0.70	0.60	**0.82**	0.68	**0.84**	**0.76**	**0.71**
VitalyProtasov	0.72	0.69	**0.81**	0.77	0.49	**0.79**	**0.87**	0.67	0.73	0.68
nikita.sushko	0.71	0.70	0.62	**0.79**	0.47	**0.89**	**0.84**	0.67	0.74	0.68
erehulka	0.71	0.88	0.71	**0.85**	**0.68**	0.78	0.52	0.63	0.65	0.69

Conclusions

- The optimal approach for training a multilingual seq2seq model for text detoxification tasks was identified
- When combined with the detoxification via toxic word deletion baseline, our resulting model achieved third place in the automatic evaluation stage of the PAN 2024 TextDetox competition
- The model and dataset are available for download on HuggingFace

Models and synthetic dataset

Thx

Contacts:

Email: nikita.sushko@skoltech.ru
Telegram: @chameleon_lizard