HybridDetox: A Combination of Supervised andUnsupervised Methods for Effective Multilingual Text Detoxification

PAN 2024 TextDetox (shared task)

Linguistic_Hygenist

Susmita Gangopadhyay, *M. Taimoor Khan* and Hajira Jabeen

GESIS – Leibniz Institute for the social sciences

Outline

- Introduction
- Challenge task
- Proposed approach
- Training
- Conclusion

Introduction

- Multilingual text detoxification
  - Revising toxic messages/comments to neutralize their toxicity while keeping the essence of the message intact (for multiple languages)
- Toxicity
  - The use of curse words, insults, hate speech, cyberbullying, or trolling and contributing to an unhealthy online environment [5]
- Example
  - I hate free speech it is shit ---> I hate free speech it is not good
- Applications
  - Social media platforms can replace toxic content with non-toxic versions
  - This allows the message to be conveyed without blocking it entirely due to toxicity

Languages: English, Russian, Ukrainian, Hindi, Chinese, Arabic, German, Amharic, Spanish

Challenge task

- Multilingual text detoxification (TextDetox) by PAN Lab
  - 2 with parallel corpora: English, Russian
  - 7 with toxic text only: Ukrainian, Hindi, Chinese, Arabic, German, Amharic, Spanish
- Challenge:
  - To detoxify text while keep its content intact
- Evaluation
  - Mode: Automatic and manual
  - Metrics: Style transfer accuracy, content preservation, fluency

Related Work

- Jigsaw/Conversation AI team
  - Toxic comment classification challenge 2 in 2018
  - Unintended bias in toxicity classification challenge 3 in 2019
  - Multilingual Toxic Comment Classification Challenge 4 in 2021
- SemEval
  - SemEval-2019 Task 6 (toxicity detection)
  - SemEval-2020 Task 12 (toxic content identification and categorization)
  - SemEval-2021 (Toxicity span detection)
- Multimedia Automatic Misogyny Identification (MAMI) in 2022
  - Identifying misogynous memes (text and images)
- RUSSE-2022 focused solely on detoxifying Russian texts [22]
- Toxicity detection using deep sequence models i.e., LSTM [15], utilization of embedding models [16], and incorporation of context [17] in the detection of toxic texts.
- Used pretrained seq2seq transformer for text detoxification [18]
- Point-wise corrections with seq2seq models to improve detoxified text fluency and style [9]

Proposed Approach

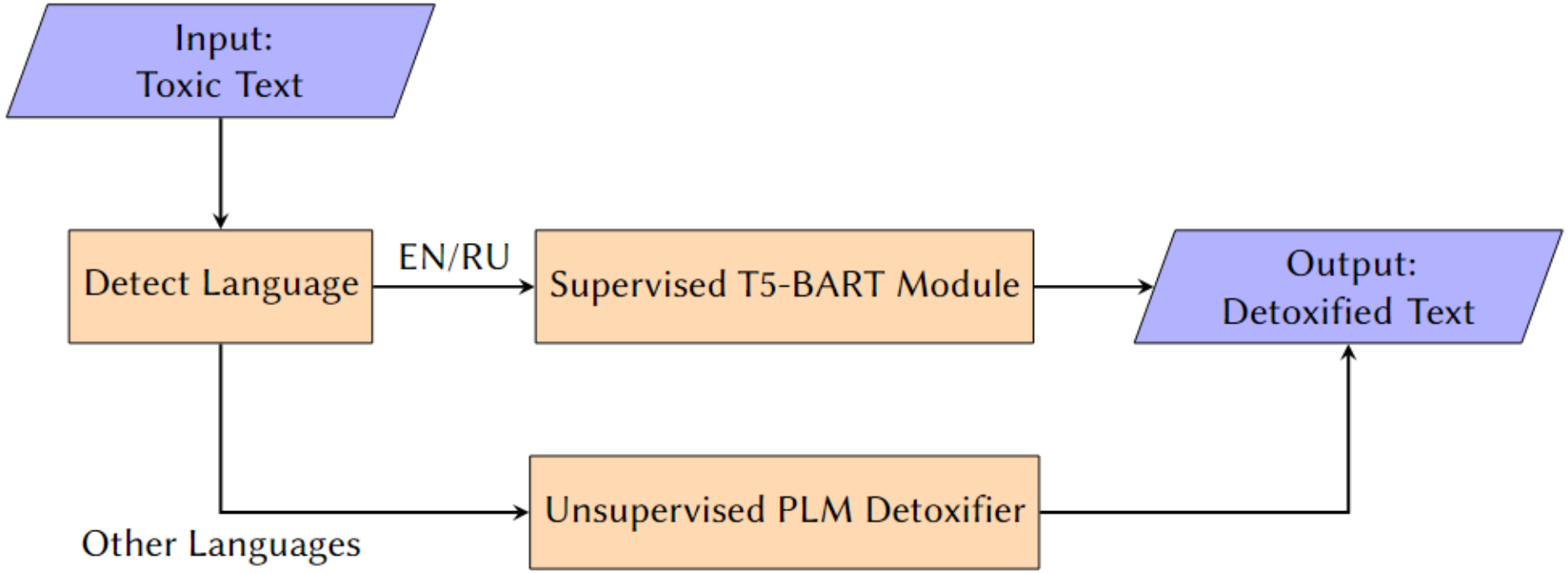


Figure 1: Detoxification Pipeline for all languages

Supervised module

- EN
  - BART model
  - ROUGE measures: ROUGE-1, ROUGE-2 and ROUGE-L

- RU
  - T5 (Text-to-Text Transfer Transformer) - EN
  - Exponentially weighted moving average (EWMA)

Finetuned on parallel corpora

Unsupervised module

- Toxic words identification
  - Detect toxic words using log-odds ratio and hashing
  - Detected toxic words masked based on a threshold
  - Log-odds word frequency in toxic vs. neutral texts.
  - Filter toxic words list on word length

Toxic words masking

  - Mask placement with linguistic patterns
  - Curse words at start/end filtered
  - Others replaced through language model
  - Cumulate masked words to one

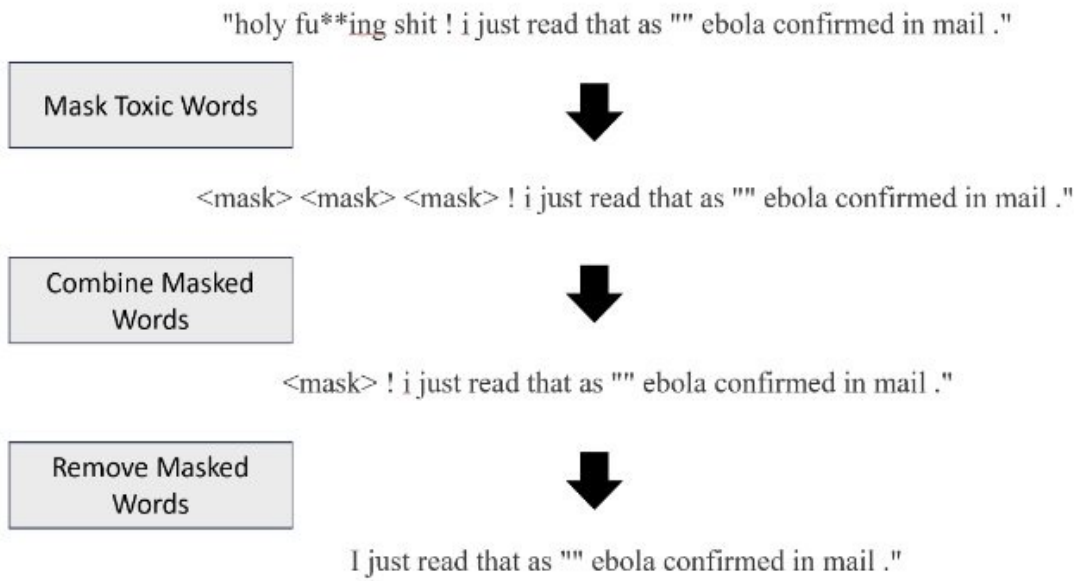


Figure 2: Example of implemented Mask Placement with Linguistic Pattern in our method

Unsupervised module

- **Mask Prediction**
  - Used XLM-RoBERTa model, pretrained on 100 languages
  - The model predicts 15% randomely masked words
  - For ultiple masks, it generates multiple possibilities (3^n combinations).
  - Chose XLM-RoBERTa for its bidirectional sentence understanding, unlike RNNs or GPT.
- **Sentence Similarity**
  - Generate sentence embeddings
  - Semantic similarity computed for pairs of toxic, neutral sentences
  - Sentence with highest dissimilarity to toxic sentence is chosen
  - The model utilizes a Siamese network with cosine similarity loss for similarity measurement

Training (Fine-tuning supervised module)

Training and validation loss of the supervised models (BART for EN) and (T5 for RU)

English		Russian	
Training Loss	Validation Loss (ROUGE Score)	Training Loss	Validation Loss (EWMA)
1.422600	1.197713	0.6551	7.6154
1.359800	1.141416	0.6522	7.7473
1.297000	1.114076	0.6881	7.7505
1.389900	1.111269	0.5655	7.7502
1.461900	1.109282	0.6839	7.4104
1.208900	1.140405	0.6881	7.7505
1.377100	1.129589	0.6098	7.9673
1.215600	1.122608	0.6349	7.6759

Example

Language	Toxic Sentence	Neutral Sentence
Ukranian	Просто забирайся нахуй з моєї голови	Просто забирайся вже з моєї голови
Hindi	तो फिर क्यो रंडी रोना कर रहा है ?	तो फिर क्यो ऐसा कर रहे है ?
English	I hate free speech it is shit	I hate free speech it is not good
Chinese	就算是编的，我也相信，臭母狗	就算是编的，我也相信，臭母狗
Arabic	هازارد كل مسيرته بدوري عبط واكبر مقل	هازارد في مسيرته ، واكبر مقل
German	Was für ein beschissenes Jahr. KANNSTE LAUT SAGEN!	Was für ein Jahr. KANNSTE LAUT SAGEN!
Russian	кулибин хренов, на хонор это не прокатывает	кулиин ренов, на онор то не рокатвает
Amharic	Hahaha አፍሪካን ???? መጀመሪያ አንተ ራስህን መረጋጋት በቻልክ	Hahaha አፍሪካን ???? መጀመሪያ አንተ ራስህን መረጋጋት በቻልክ
Spanish	se ve como una mierda la foto perooo	e ve como una de la foto perooo

Figure 3: Sample results of toxic and detoxified text in each of the languages

Results

Evaluation	average	en	es	de	zh	ar	hi	uk	ru	am
Manual	0.50	0.74	0.20	0.72	0.37	0.61	0.75	0.48	0.00	0.61
Automatic	0.315	0.472	0.356	0.414	0.069	0.425	0.198	0.528	0.090	0.280

Manual and automatic scores of our proposed approach for individual languages and their average. The evaluation is based on *removing toxicity*, *style transfer*, *accuracy*, *content preservation* and *fluency*

Limitations

- Exponentially weighted moving average (EWMA) was not an appropriate choice for this task
- The toxic text samples for other languages could be used with few-shot learning
- Our approach didn't explicit attempt to determine the content/message in the toxic text
- The unsupervised approach for all 7 languages could be separated for languages with shared roots

Conclusion and limitations

- *Text detoxification is a challenging task depending diverse presence of toxicity and its detoxified versions*
- *A hybrid approach for text detoxification is a plausible direction however it needs more ground truth for higher accuracy*
- *Our proposed model received 0.315 score and can be improved by addressing the limitations highlighted*
- *Using multilingual embeddings and transfer learning is not explored for this task*

Methods Hub

This method along with many other interesting methods applicable on digital behavioral data for the social science use cases are available on the portal for reuse.

GESIS Leibniz-Institut für Sozialwissenschaften

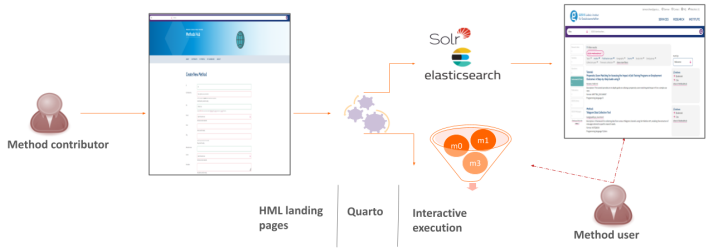
Methods Hub: A Platform for Sharing Computational Social Science Methods

M. Taimoor Khan, Arnim Bleier, Chung-Hong Chan, Po-Chun Chang, Raniere, Costa da Silva, Danilo Dessi, Stefan Dietze, Gabriella Lapesa, Brigitte Mathiak, David Schoch, Claudia Wagner and Hajira Jabeen*

KTS & CSS Departments at GESIS – Leibniz Institute for the Social Sciences

Methods Hub is a niche platform for advanced computational methods carefully curated and documented for the needs of social scientists. It offers:

- Advanced AI-based computational methods to collect, preprocess, analyze and visualize digital behavioral data (DBD).
- Tutorials demonstrate the application of methods for specific use cases as a step-wise guide.
- All methods and tutorials are public access, open licensed and follow documentation standards to promote reusability.
- The portal facilitates to search, access and work with the methods and tutorials through integration with other services e.g., GESIS search.



Sample methods and tutorials

- *TelegramToolkit* for data collection and enrichment
- *RTOOT:* Interact with the mastodon API from R
- *Keyword's Finder:* A tool for comparative keyword analysis
- Quantifying implicit associations among words using word embeddings
- Tutorial on similar tweets using locality sensitive hashing
- *OOLONG:* Create Validation Tests for Automated Content Analysis
- *SWEATER:* Test for associations among words in word embedding spaces
- *ScienceLinker:* A date linking, enrichment and analysis toolkit
- Tutorial on using SSciBERT politics model to detect political science domain texts from a dataset of scientific abstracts
- Tutorial demonstrate the use of Flair NLP framework to extract NERs from scientific acknowledgement

Why Methods Hub

Social Science Relevance
- Address use case(s) and research question(s) from social science domain
- Evidence of applicability to Digital Behavioral Data (DBD)

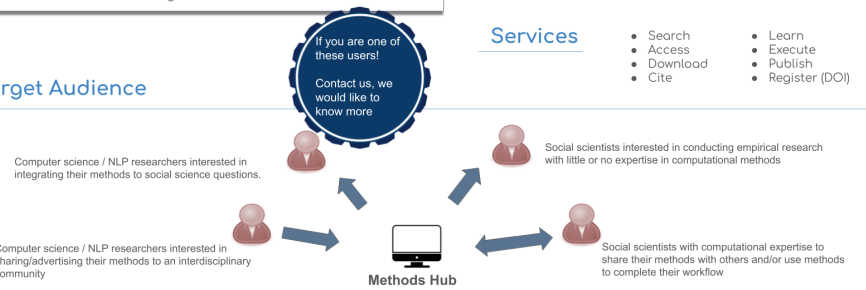
Well documented & Reusable code
- The methods and tutorials follow documentation standard to understand and reuse
- Update and generalize the code for newer research questions and share with the community

Open Science
- Publicly available methods and tutorial repos
- Open licensed e.g., MIT, Apache 2.0, CC-BY 4.0

Services

- Search
- Access
- Download
- Cite
- Learn
- Execute
- Publish
- Register (DOI)

Target Audience



Contact

For future information please contact:
David Schoch david.schoch@gesis.org
M. Taimoor Khan taimoor.khan@gesis.org

GESIS - Leibniz Institute for the Social Sciences

Thanks for listening