Overview of the Multilingual Text Detoxification Task at PAN 2024

Daryna Dementieva[1,*], Daniil Moskovskiy[2,3], Nikolay Babakov[4], Abinew Ali Ayele[5], Naquee Rizwan[6], Florian Schneider[5], Xintong Wang[5], Seid Muhie Yimam[5], Dmitry Ustalov[7], Elisei Stakovskii[8], Alisa Smirnova[9], Ashraf Elnagar[10], Animesh Mukherjee[6] and Alexander Panchenko[2,3]

[1]*Technical University of Munich, Munich, Germany*
[2]*Skolkovo Institute of Science and Technology, Moscow, Russia*
[3]*Artificial Intelligence Research Institute, Moscow, Russia*
[4]*Universidade of Santiago de Compostela, Santiago de Compostela, Spain*
[5]*Universität Hamburg, Hamburg, Germany*
[6]*Indian Institute of Technology, Kharagpur, India*
[7]*JetBrains, Belgrade, Serbia*
[8]*Independent Researcher*
[9]*Toloka AI, Lucerne, Switzerland*
[10]*University of Sharjah, Sharjah, UAE*

Abstract

Despite different countries and social platform regulations, digital abusive speech persists as a significant challenge. One of the way to tackle abusive, or more specifically, toxic language can be automatic text detoxification—a text style transfer task (TST) of changing register of text from toxic to more non-toxic. Thus, in this shared task, we aim to obtain text detoxification models for 9 languages: English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, and Amharic. This paper presents the Multilingual Text Detoxification (TextDetox) task, the underlying datasets, the evaluation setups, the submissions from participants, and the results obtained.
Warning: This paper contains rude texts that only serve as illustrative examples.

Keywords

PAN 2024, Multilingual Text Detoxification, Text Style Transfer, Multilingualism

1. Introduction

The issue of managing toxic speech remains a crucial aspect of human communication and **digital violence** prevention [1], including the mitigation of toxic responses generated by Large Language

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
daryna.dementieva@tum.de (D. Dementieva); daniil.moskovskiy@skoltech.ru (D. Moskovskiy); nikolay.babakov@usc.ese (N. Babakov); abinew.ali.ayele@uni-hamburg.de (A. A. Ayele); nrizwan@kgpian.iitkgp.ac.in (N. Rizwan); florian.schneider-1@uni-hamburg.de (F. Schneider); xintong.wang@uni-hamburg.de (X. Wang); seid.muhie.yimam@uni-hamburg.de (S. M. Yimam); dmitry.ustalov@jetbrains.com (D. Ustalov); eistakovskii@gmail.com (E. Stakovskii); ashraf@sharjah.ac.ae (A. Elnagar); animeshm@gmail.com (A. Mukherjee); a.panchenko@skol.tech (A. Panchenko)
https://dardem.github.io (D. Dementieva); https://www.researchgate.net/profile/Daniil-Moskovskiy (D. Moskovskiy); https://github.com/bbkjunior/bbkjunior (N. Babakov); https://scholar.google.com/citations?user=g2m1wH4AAAAJ&hl=en (A. A. Ayele); https://www.linkedin.com/in/naquee-rizwan-a97abb159 (N. Rizwan); https://www.linkedin.com/in/flo-schneider-hh (F. Schneider); https://ethanscuter.github.io (X. Wang); https://seyyaw.github.io (S. M. Yimam); https://linkedin.com/in/ustalov (D. Ustalov); https://github.com/eistakovskii (E. Stakovskii); https://www.sharjah.ac.ae/en/academics/Colleges/CI/dept/cs/Pages/ppl_detail.aspx?mcid=4 (A. Elnagar); https://cse.iitkgp.ac.in/~animeshm (A. Mukherjee); https://alexanderpanchenko.github.io (A. Panchenko)
0000-0003-0929-4140 (D. Dementieva); 0009-0006-7943-4259 (D. Moskovskiy); 0000-0002-2568-6702 (N. Babakov); 0000-0003-4686-5053 (A. A. Ayele); 0009-0007-1872-6618 (N. Rizwan); 0000-0003-4141-1415 (F. Schneider); 0009-0002-8005-2259 (X. Wang); 0000-0002-8289-388X (S. M. Yimam); 0000-0002-9979-2188 (D. Ustalov); 0000-0003-2265-7268 (A. Elnagar); 0000-0003-4534-0044 (A. Mukherjee); 0000-0001-6097-6118 (A. Panchenko)

Models (LLMs) [2]. The typical approach to dealing with abusive speech on social platforms involves message blocking [3]. To address this, numerous toxic and hate speech detection models have been developed for different languages, i.e. English [4], Spanish [5], Amharic [6], Code-Mixed Hindi [7], and many others [8]. However, the recent research indicates a necessity for more proactive moderation of abusive speech [9]. One such approach is **text detoxification**.

Within the baselines approaches for automatic text detoxification, multiple unsupervised baselines were created based on ideas of Delete-Retrieve-Generate [10], latent style spaces disentanglement [11], or conditional generation with Masked Language Modeling [12]. However, the latest state-of-the-art outcomes, particularly in English, were attained when parallel data and fine-tuning with text-to-text generation models were employed [13, 14]. At the same time, the availability of such a corpus can be a challenge for new languages and cross-lingual transfer techniques should be applied [15].

Figure 1: In this work, we present a novel benchmark datasets for multilingual and cross-lingual text detoxification for 9 languages: English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, and Amharic.

In this shared task, we explored both setups—cross-lingual and multilingual one—providing new multilingual parallel text detoxification dataset for 9 languages [16]. The remainder of this paper is structured as follows. Section 2 gives an overview of the TextDetox shared task description. Section 3 provides the full overview of the new multilingual parallel text detoxification dataset collection per each language. In the following sections, the evaluation setups essentials are described—baselines in Section 4, automatic evaluation setup in Section 5, and human evaluation setup in Section 6. The submissions from participants are described in Section 7. Section 8 provides the details about final results—both automatic (Section 8.1) and human (Section 8.2) evaluation leaderboards. Finally, Section 9 concludes the paper.

All the resources produced from the task are listed at the shared task page [1] and are also mentioned in the corresponding sections.

2. Shared Task Description

Here, we provide the shared task main definitions—how we understand toxicity, text style transfer task, cross-lingual and multilingual setup, and the competition rules.

[1] https://pan.webis.de/clef24/pan24-web/text-detoxification.html

2.1. Task Definition

Definition of Toxicity While there can be different types of toxic language in conversations [17, 18], i.e. sarcasm, hate speech, threats, in this work, we include samples with substrings that are commonly referred to as **vulgar or profane language** [19, 13] while the whole main message can be both neutral and toxic, but not hateful with direct insult of individuals or groups of people.

Text Detoxification as Text Style Transfer In this work, we adapt the formal task definition of the text style transfer described in [20, 21, 13]:

Having a set of style S and a corpus of texts D, a text style transfer (TST) model is a function $\alpha : S \times S \times D \rightarrow D$ that, given a source style s^{src}, a target style s^{tg}, and an input text d^{src}, produces an output text d^{tg} such that:

- The style of the text changes from the source style s^{src} to the target style s^{tg} and is measured by a style classifier: $\sigma(d^{src}) \neq \sigma(d^{tg}), \sigma(d^{tg}) = s^{tg}$;
- The content of the source text is saved in the target text as much as required for the task and estimated by a content similarity function: $\delta(d^{src}, d^{tg}) \geq t^{\delta}$;
- The fluency of the target text achieves the required level according to the fluency estimator: $\psi(d^{tg}) \geq t^{\psi}$,

where t^{δ} and t^{ψ} are the threshold values for the content preservation (δ) and fluency (ψ) functions and can be adjusted to the specific task. In our task, the source style s^{src} is `toxic` and the target style s^{tg} is `non-toxic`.

Cross-lingual Text Detoxification As parallel text detoxification corpora might not be available for any language, one of the important tasks is to explore cross-lingual text detoxification knowledge transfer. In this case, we assume that training data is available for the resource-rich language (i.e. English) and the task is to obtain a text detoxification system for a new language.

Multilingual Text Detoxification If parallel corpora available for multiple language, then both monolingual text detoxification models per language and multilingual model for all languages can be obtained.

2.2. Competition Rules

The share task timeline was divided in to two phases—development and test.

Development Phase For the first phase, only training parallel data for English and Russian from previous works [13, 22] aiming to provide participants to explore *cross-lingual* transfer techniques.

Test Phase During the test phase, parallel text detoxification corpora were available for all target languages. Participants was invited to submit *monolingual* and *multilingual* solutions.

Leaderboards During both phases, the leaderboards based on automatic evaluation were available. We used Codalab platform [23][2] (and TIRA [24] as a backup platform). However, despite having powerful models capable of classifying texts and embedding their meanings, human judgment remains superior for making final decisions in the text detoxification task [25]. Thus, based on a *test* part subset, we performed human evaluation of the participants submissions. **The final leaderboard** was based on the human judgements results.

[2] https://codalab.lisn.upsaclay.fr/competitions/18243

3. Multilingual Parallel Text Detoxification Dataset

For each of our 9 target languages, we prepared parallel text detoxification corpus. We asked experts and native speakers to contribute for corpora collection. Further, we describe the collection details per each language: English (Section 3.1), Russian (Section 3.2), Ukrainian (Section 3.3), Spanish (Section 3.4), German (Section 3.5), Hindi (Section 3.6), Amharic (Section 3.7), Arabic (Section 3.8), Chinese (Section 3.9). All the instructions per language are available online.[3] We also opensource the obtained resources for the public usage.[4]

For all the data, we adapt the concept of English ParaDetox [26] collection pipeline as it was designed to automate the data collection as well as verification with crowdsourcing. The pipeline consists of three tasks:

Task 1: Rewrite text in a polite way Annotators need to provide the detoxified paraphrase of the text so it becomes non-toxic and the main content is saved or to skip paraphrasing if the text is not possible to rewrite in non-toxic way;

Task 2: Do these sentences mean the same? Check if the content is indeed the same between the original toxic text and its potential non-toxic paraphrase;

Task 3: Is this text offensive? Verification of the provided paraphrase if it is indeed non-toxic.

In the same manner, each language stakeholder asked the annotators to rewrite the toxic samples verifying the main three criteria: (i) the new paraphrase should be non-toxic; (ii) the content should be saved as much as possible; (iii) the resulted text should be fluent but may contain some minor mistakes (as the majority of the original toxic samples are examples from posts from social networks). The rewriting of the texts and verification of their quality could have been done either via crowdsourcing or via manual annotation. The main goal for each language was to obtain 1000 parallel pairs that were later splitted into dev and test sets.

Data Preprocessing For all languages, we maintain the length of samples as sentences of around 5-20 tokens. Also, if a text sample is from a social network, we anonymize any mentioning of usernames and links.

3.1. English

For English, we reused the data from English ParaDetox dataset [13] and additionally manually marked up approximately 400 pairs to form a validation dataset of 1000 examples.

3.1.1. Input Data Preparation

For EnParaDetox, the original toxic texts were taken from Jigsaw toxicity identification challenge train dataset [27]. We have considered only texts labeled as `toxic` and `severe toxic`.

3.1.2. Annotation Process

The training and validation sets of EnParaDetox were acquired through crowdsourcing via Toloka[5] platform with fluent English speakers. Additionally, we employed annotators who are fluent in English and hold a Masters degree in Computer Science to compile additional samples to the test set.

[3] https://github.com/textdetox/textdetox_clef_2024/tree/main/instructions/paradetox_collection
[4] https://huggingface.co/textdetox
[5] https://toloka.ai

3.2. Russian

The same as for English, there were previously available training and validation data from previous work [22, 28]. We reused this data and manually annotated some additional toxic examples taken from various toxicity datasets.

Input Toxicity Data The original toxic samples were taken from two binary toxicity classification Kaggle Toxic Comments datasets [29, 30].

3.2.1. Annotation Process

The training and validation sets of RuParaDetox were acquired through crowdsourcing via Toloka platform with fluent Russian speakers. Additionally, we employed annotators who are native in Russian and hold a Masters degree in Computer Science to compile additional samples to the test set.

3.3. Ukrainian

We used the data presented in MultiParaDetox paper [28] providing the main details of data collection:

Input Toxicity Data For the Ukrainian language, there was no existing binary toxicity classification corpus. Therefore, we filtered explicitly toxic samples containing obscene lexicon from the predefined list [31] within the Ukrainian Tweets Corpus [32].

3.3.1. Annotation Process

We adapt ParaDetox [26] collection pipeline and verified the data quality via crowdsourcing. We utilized the Toloka platform for crowdsourcing tasks in Ukrainian. The annotators, who were native Ukrainian speakers, underwent an examination before starting the tasks.

3.4. Spanish

We used the data presented in MultiParaDetox paper [28] providing the main details of data collection:

Input Toxicity Data For Spanish, we selected samples for annotation from three datasets: hate speech detection ones [33, 34] as well as filtered by keywords Spanish Tweets corpus [35].

3.4.1. Annotation Process

We adapt ParaDetox [26] collection pipeline and verified the data quality via crowdsourcing. We utilized the Toloka platform for crowdsourcing tasks in Ukrainian. The annotators, who were native Spanish speakers, underwent an examination before starting the tasks.

3.5. German

German ParaDetox was collected with several annotators with manual quality verification:

3.5.1. Input Data Preparation

The German language source data in this work is based on three datasets containing toxic, offensive, or hate speech comments on social media about primarily political events in Germany or the US. For the two datasets from the GermEval 2018 [36] and GermEval 2021 [37] shared tasks, we used data from both the test and the train split and filtered it as follows. For the GermEval 2018 data, we only used samples labeled with the coarse class "*OFFENSE*" whereas for the GermEval 2021 data – which contains different labels – we only used samples annotated with the "*Sub1_Toxic*" class. The third dataset [38]

was filtered so only samples were kept where both expert annotators classified the samples as hate speech.

The data from the three datasets was merged and deduplicated via exact string matching. Furthermore, we removed all samples that included less than 5 or more than 30 white-space separated tokens.

3.5.2. Annotation Process

To create the final parallel detoxified German dataset, we hired two native German annotators. Annotator A is a female born in 1994 who holds a Master of Arts degree in Social Sciences, and Annotator B is a male born in 1992 who holds a Master of Science degree in Computer Science. The data was distributed so that each sample was transcribed by only one of the annotators.

3.6. Hindi

Hindi dataset was collected manually by a native-speaker annotator gaining data from multiple sources:

3.6.1. Input Data Preparation

Input Toxicity Data We used the HASOC dataset created at FIRE 2019 [39] as source for Hindi language. Contents in this dataset are relevant within Indian subcontinent which are collected from various social media platforms prevalent in India. In this dataset, hostile posts are divided into *HATE SPEECH, OFFENSIVE* and *PROFANE*. For curation, posts containing *OFFENSIVE* and *PROFANE* contents in train and test splits were used. 1455 *PROFANE* posts (1237 train + 218 test) and 873 *OFFENSIVE* posts (676 train + 197 test) were chosen to prepare detoxifiable toxic data for our task.

Input Preprocessing On a total of 2328 samples, we first performed deduplication via exact string matching. Mentions, links and emojis were also removed as part of this step.

3.6.2. Annotation Process

Annotation Task(s) The posts after input preprocessing were manually verified. Those with less than 5 white-space separated tokens were removed and which had more than 25 white-space separated tokens were re-framed to bring them down to this limit. Toxicity and meaning of the posts were unchanged during this re-framing. These posts were then bifurcated into detoxifiable and non-detoxifiable labels. The manual re-framing and bifurcation were carried by a NLP researcher with working experience on hate/toxic speech.

Out of 2328 samples, 1007 samples were marked as detoxifiable. From these detoxifiable samples, we carefully sampled 24 data points and detoxified them. These detoxified samples were evaluated by two experts who are native Hindi language speakers to provide precise samples to the annotators for detoxifying the whole dataset. Annotators were guided based on expert prepared samples and were asked to re-write toxic pairs in a non-toxic manner, keeping the meaning of the original post unchanged. Detoxification was carried out by two annotators and we provide their details in the corresponding subsection.

Annotators One male NLP researcher working in the field of hate/toxic speech and another female student enrolled in Bachelor's Degree and having working knowledge in Machine Learning, were employed to carry out the detoxification of whole dataset. Both annotators are Indian, native Hindi speakers and are well versed with the topicality covered in the dataset.

3.7. Amharic

We compiled new Amharic ParaDetox datasets with the following annotation details, based on prior studies of hate and offensive language:

3.7.1. Input Data Preparation

The Amharic ParaDetox dataset is derived from merging two pre-existing studies conducted on the X/Twitter datasets [6, 40]. The dataset introduced by Ayele et al. [40] was initially annotated into categories of *hate*, *offensive*, *normal*, and *unsure* by three native speaker annotators, with the gold labels determined through a majority voting scheme. In contrast, the dataset presented by Ayele et al. [6] was annotated by two native speakers, with a third adjudicator annotator deciding the gold labels for instances where there was no majority consensus. We extracted a subset of these datasets labeled as *offensive* to create the new Amharic ParaDetox dataset and subsequently reworked this subset using new annotators to determine if the messages could be detoxified and to present non-toxic versions of each message.

Input Toxicity Data The input toxicity data is entirely sourced from the two previous studies, namely Ayele et al. [6] and Ayele et al. [40], and has been adapted to meet the requirements of the ParaDetox task.

3.7.2. Annotation Process

Annotation Task(s) We customized the Potato-POrtable Text Annotation TOol[6] and utilized it for the annotation of Amharic ParaDetox dataset. Annotators were provided annotation guidelines, took hands-on practical training, completed independent training tasks before the main annotation task.

We conducted pilot annotation of 125 sample items with three native Amharic speaker annotators and evaluated the annotation quality with experts and annotators together in a group meeting to improve the understandings of annotators for the main task. The main annotation task comprises of 2,995 tweets, each annotated by one annotator. Annotators were asked to classify each tweet in to two broad categories, detoxifiable and non-detoxifiable. For the detoxifiable category, annotators are asked to detoxify and re-write the text. For non-detoxifiable tweets, annotators choose non-detoxifiable and select reason as; it is hate speech, it is normal speech or indeterminate to decide the label.

Annotators Annotators have previous hate speech annotation experiences and already holds Masters degree in Computer Science. Only two of the annotators were evolved in the main annotation task, where both of them are university lecturers and have basic knowledge of natural language processing tasks. One of the annotators is from Adama Scinece and Technology University with experience of 15 years of teaching Computer Science, who is female. The other annotator is a male, who has been teaching Computer science over 12 years in Kotebe University of Education.

3.8. Arabic

Arabic ParaDetox was collected with several annotators with manual quality verification:

3.8.1. Input Data Preparation

The Arabic ParaDetox dataset was created by combining parts of several existing datasets along with the Arabic-translated version of the Jigsaw dataset [27]. It includes the Levantine Twitter Dataset for Hate Speech and Abusive Language (L-HSAB) [41], which focuses on Levantine dialects, and the Tunisian Hate and Abusive Speech (T-HSAB) dataset [42], which targets Tunisian dialects. It also incorporates the OSACT dataset [43] and the Arabic Levantine Twitter Dataset for Misogynistic Language (LeT-Mi) [44], which specifically addresses gender-based abuse. These resources combine to form the Arabic ParaDetox dataset, aimed at aiding the development of toxicity classifiers capable of handling Arabic content across various dialects and contexts.

[6] https://github.com/davidjurgens/potato

3.8.2. Annotation Process

Annotators The detoxification process was conducted by three annotators, each with a PhD. The team includes two males and one female, all of whom have a strong interest in computational linguistics. These native Arabic speakers possess a deep understanding of the subjects encompassed within the dataset. Each text sample was transcribed by two of the annotators to ensure accuracy and consistency in the data.

3.9. Chinese

We collected new Chinese ParaDetox datasets with the following annotation details:

3.9.1. Input Data Preparation

Input Toxicity Data The Chinese ParaDetox dataset is derived from TOXICN [45], a recently released Chinese toxic language dataset. TOXICN was compiled from social media platforms and comprises 12,011 comments addressing several sensitive topics, including gender, race, region, and LGBTQ issues. From this dataset, we extracted a subset based on multiple criteria: the number of toxic words, the ratio of toxic words in the comments, the length of comments, and the toxic scores of comments.

Input Preprocessing We set thresholds for the criteria mentioned above: the number of toxic words ranged from 1 to 5, the ratio of toxic words in comments was less than 0.5, and the length of comments ranged from 3 to 50 words, ensuring suitability for annotators to rewrite them. Following these criteria, we extracted 1,516 samples from the training set and 231 samples from the test set.

Subsequently, we employed a pre-trained toxic classifier [45] to compute the toxic scores of the selected comments, using a threshold score of 0.978 to filter the candidates. Ultimately, we collected 1,149 samples from the training set and 231 samples from the test set, resulting in a total of 1,380 samples deemed suitable for annotation.

3.9.2. Annotation Process

Annotation Tasks For data annotation and verification, we employed a specifically designed three-task pipeline:

Task 1: Determine if the sentences are toxic or neutral. Annotators were required to choose one of three options: the given sentence is *neutral, toxic but can be rewritten, or toxic and cannot be rewritten.* The last option was included based on the observation that some toxic texts are impossible to rewrite in a non-toxic manner.

Task 2: Rewrite sentences in a non-toxic style. Annotators were instructed to create detoxified versions of the toxic sentences identified in Task 1. They were advised to retain the main content of the original sentences and rewrite the toxic words in a polite manner.

Task 3: Cross-check verification. The rewritten sentences from Task 2 were cross-distributed to different annotators for verification. The goal was to ensure the rewritten sentences were non-toxic and adhered to our guidelines. If annotators selected the *"No" option*, indicating the sentence did not meet the criteria, a further meta-rewrite process was initiated.

From the 1,380 toxic samples, 1,031 samples were successfully detoxified and verified, with 861 from the training set and 170 from the test set. The remaining 349 samples were either considered non-toxic or toxic but could not be rewritten.

Annotators For the detoxification process, we hired three native Chinese annotators. Two female annotators, both aged 22, hold Bachelor's degrees in Engineering, and a male annotator, aged 32, holds a Master's degree in Computer Science. All annotators are native Chinese speakers residing in mainland China, ensuring they deeply understand the Chinese language and the detoxification task.

3.10. Final Dataset

Table 1
The statistics of all ParaDetox datasets used in the shared task. The human detoxified references were collected either via crowdsourcing or locally hired native speaker. For English and Russian, the previously collected train data was available during all shared task's phases. For other languages, 1 000 samples per language were divided correspondingly into development and test parts.

Language	Source of Toxic Samples	Annotation Process	Train	Dev	Test
English	[27]	Crowdsourcing + Manual	11 939	400	600
Russian	[29, 30]	CrowdSourcing + Manual	8 500	400	600
Ukrainian	[32]	Crowdsourcing	—	400	600
Spanish	[33, 34, 35]	Crowdsourcing	—	400	600
German	[36, 37, 38]	Manual	—	400	600
Hindi	[39]	Manual	—	400	600
Amharic	[6, 40]	Manual	—	400	600
Arabic	[41, 42, 43, 44]	Manual	—	400	600
Chinese	[45]	Manual	—	400	600

The full picture of the collected ParaDetox data for all target languages is presented in Table 1. While the methods of collecting human annotations vary across languages—some data were gathered via crowdsourcing, others by hiring local native speakers—the quality of the texts was uniformly verified by experts to ensure three key attributes as introduced in [46, 13]: (i) the style of new paraphrases is genuinely non-toxic, (ii) the main content is preserved, and (iii) the new texts are fluent.

For each language for the shared task's phases:

- During the *development* phase: 400 *only* toxic parts were available for participants to perform cross-lingual experiments.
- During the *test* phase: (i) 400 ParaDetox instances were fully released; (ii) participants should provide their final solutions for 600 toxic parts of the test dataset.

For English and Russian during all phases, additional training parallel datasets were available released from previous work [13, 22, 28]. You can find online fully released development part of the data[7] and the test part only toxic instances.[8]

4. Baselines

We provide four baselines for our shared task: (i) trivial Duplicate baseline; (ii) a rule-based Delete approach; (iii) Backtranslation pipeline that reduces the task to the monolingual one; (iv) finally, fine-tuned for the downstream task on the dev dataset mT5 instance. The code for all the baselines is available online.[9]

[7] https://huggingface.co/datasets/textdetox/multilingual_paradetox
[8] https://huggingface.co/datasets/textdetox/multilingual_paradetox_test
[9] https://github.com/pan-webis-de/pan-code/tree/master/clef24/text-detoxification/baselines

Duplicate Trivial baseline: the output sentence is a copy-paste of the input sentence. This baseline has 1.0 (or 100%) SIM score by definition.

Delete For the first unsupervised baseline, we perform an elimination of obscene and toxic substrings from a text according to the predefined lists of keywords. For the shared task, we collected and compiled together the lists of such toxic keywords for all target languages based on openly available sources (see Table 2). The amount of toxic keywords per language differs which displays the diversity of morphological forms and variations of toxicity expressions across languages. For participants and further public usage, we release our compiled list online.[10]

Table 2
The list of the original sources and the corresponding amount of obscene keywords used to compile multilingual toxic lexicon list for our Delete baseline.

Language	Original Source	# of Keywords
English	[13, 47, 19]	3 390
Russian	[22, 19]	141 000
Ukrainian	[48, 19]	7 360
Spanish	[19]	1 200
German	[49, 19]	247
Hindi	[19]	133
Amharic	Ours+[19]	245
Arabic	Ours+[19]	430
Chinese	[50, 45, 19]	3 840

Backtranslation As for a more sophisticated unsupervised baseline, we perform translation of non-English texts in English with NLLB [19] instance[11] and then perform detoxification with the fine-tuned on English ParaDetox train part BART [13] instance.[12]

Fine-tuned mT5 Specifically for the *test* phase, we fine-tuned the multilingual text-to-text generation model mT5 [51]. We tuned the mT5-XL[13] on the released for the test phase parallel *development* part of the presented multilingual data.

5. Automatic Evaluation Setup

We adopt the monolingual evaluation pipelines from [13, 22] to our multilingual setup and provide the detailed description below. We evaluate the outputs based on three parameters—style of text, content preservation, and conformity to human references—combining them into the final **J**oint score. The evaluation script is available online.[14]

Style Transfer Accuracy (STA) ensures that the generated text is indeed more non-toxic. To prepare a model for this metric that covers our target languages, we subsampled 5 000 samples—2 500 toxic and 2 500 neutral—from toxicity classification corpora for each language (see references in Table 1) that were not used for ParaDetox data collection. We released[15] this compiled corpus for participants as an additional dataset for experiments and fine-tuned XLM-R [52] `large` instance for the binary toxicity

[10] huggingface.co/datasets/textdetox/multilingual_toxic_lexicon
[11] https://huggingface.co/facebook/nllb-200-distilled-600M
[12] https://huggingface.co/s-nlp/bart-base-detox
[13] https://huggingface.co/google/mt5-xl
[14] https://github.com/pan-webis-de/pan-code/blob/master/clef24/text-detoxification/evaluate.py
[15] https://huggingface.co/datasets/textdetox/multilingual_toxicity_dataset

classification task. The model is also available for the public usage[16] and is used in the shared task to estimate the level of non-toxicity in the texts.

Content Similarity (SIM) is the cosine similarity between LaBSE[17] embeddings [53] of the source texts and the generated texts.

Fluency (ChrF1) is used to estimate the proximity of the detoxified texts to human references. While in several previous work language acceptability classifiers based on CoLa-like corpora were utilized for fluency estimation [13, 15], the recent work [25] also showed that reference-based metrics achieved high correlations with human evaluation. Thus, we use an implementation of ChrF1 score from `sacrebleu` library [54].

Joint score (J) is the aggregation of the three above metrics. The metrics **STA**, **SIM** and **ChrF1** are subsequently combined into the final **J** score used for the final ranking of approaches. Given an input toxic text x_i and its output detoxified version y_i, for a test set of n samples:

$$\mathbf{J} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{STA}(y_i) \cdot \mathbf{SIM}(x_i, y_i) \cdot \mathbf{ChrF1}(x_i, y_i),$$

where $\mathbf{STA}(y_i)$, $\mathbf{SIM}(x_i, y_i)$, $\mathbf{ChrF1}(x_i, y_i) \in [0, 1]$ for each text detoxification output y_i.

6. Human Evaluation Setup

For the test set, we performed the human evaluation to obtain final judgements on the participants' systems. The details and instructions of the annotation setups are available for the public usage.[18]

6.1. General setup

We used Toloka[19] crowdsourcing platform for manual evaluation of automatic detoxification. For each project, the annotators who indicated the knowledge of the necessary language could access the tasks.

Quality Control First, the annotators underwent a rigorous selection process, starting with a training pool of 20 tasks with predefined answers and instructional feedback. Following this, annotators faced an examination pool comprising 10 tasks with known answers. Only those achieving a minimum accuracy of 65% in both the training and examination pools were permitted to proceed to the main evaluation tasks.

During the execution of the real tasks, annotators' performance was meticulously monitored using several measures. Response speed was a critical factor; annotators who responded rapidly (less than 15 seconds per page) to three consecutive task pages were permanently banned. Additionally, we evaluated the answers of the annotators to mixed tasks with known true answers. Those who demonstrated low performance, were defined as achieving less than 65% accuracy, were required to undergo re-examination. Furthermore, annotators who skipped three task pages were also banned from the study.

Evaluation Dataset We selected 100 random original toxic samples per each language from the *test* part of our dataset. Each pair was annotated within a separate project dedicated to the respective language. These projects were categorized into three distinct types, which will be described in detail in the following section.

[16] https://huggingface.co/textdetox/xlmr-large-toxicity-classifier
[17] huggingface.co/sentence-transformers/LaBSE
[18] https://github.com/textdetox/textdetox_clef_2024/tree/main/instructions/human_evaluation
[19] https://toloka.ai

6.2. Annotation projects and corresponding metrics

In general, the concept of the human evaluation mirrored the approach used in the automatic evaluation. Each project type focused on assessing one of the three key qualities of detoxification; style transfer accuracy, content similarity, or fluency. Below, we provide a detailed description of each project type and its specific focus within the evaluation process.

Style Transfer Accuracy To measure style transfer accuracy, we employed a pairwise comparison between the original toxic text and the generated detoxified text. Annotators were tasked with determining which text was more toxic: the left text, the right text, or neither. An illustration of this task can be found in Figure 2.

To enhance realism, we randomized the sequence of original and detoxified texts. Annotators' votes were then converted into numerical values using the following logic: if the original toxic text was deemed more toxic, the vote was recorded as 1; if the detoxified text was considered more toxic, the vote was recorded as 0; if "none of the texts" was selected, the vote was recorded as 0.5.

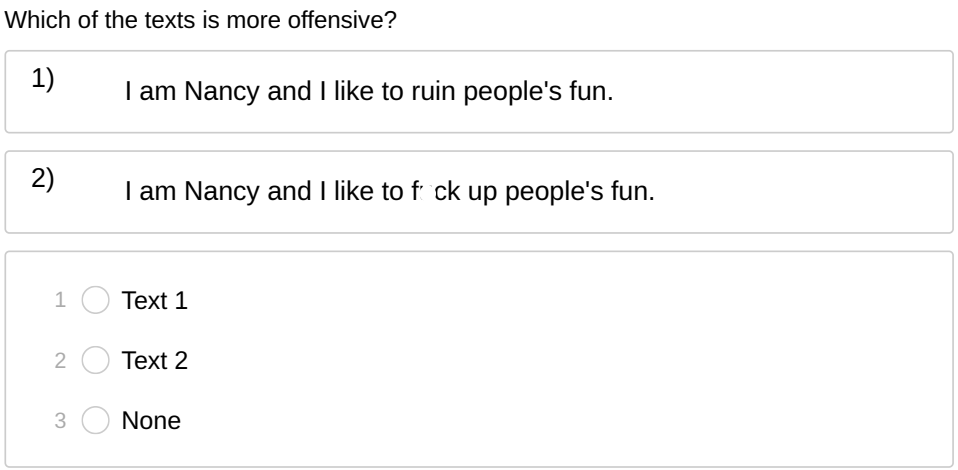

Figure 2: The English interface example of crowdsourcing Task 1: pairwise toxicity comparison.

Content Similarity For the content similarity metric, we presented pairs of texts in a fixed order (the toxic phrase followed by the detoxified phrase) and asked annotators to indicate whether the sense of the two texts was similar. Annotators' responses were binary, with "yes" indicating that the sense was similar and "no" indicating otherwise. The responses were mapped to numerical values, with a "yes" response being recorded as 1 and a "no" response being recorded as 0. An illustration of this task can be found in Figure 3.

Fluency For the fluency metric, sentences were evaluated individually rather than in pairs. Annotators were asked whether each text was intelligible and correct, with three possible responses: yes, partially, and no. Each response was mapped to a numeric value: "yes" was scored as 1, "partially" as 0.5, and "no" as 0. The fluency score for a pair of texts was determined by comparing the scores of the detoxified text to the original text. If the detoxified text received a higher or equal fluency score, the pair was scored as 1; otherwise, it was scored as 0. An illustration of this task can be found in Figure 4.

Joint Score The final scores for each pair within each project were calculated by weighting them according to the skill of the crowdsourcing annotators. This skill was dynamically assigned based on the quality of their responses to both the tasks with known answers and the real tasks. The logic for calculating the Joint score in the manual evaluation mirrors that of the automatic evaluation. For each

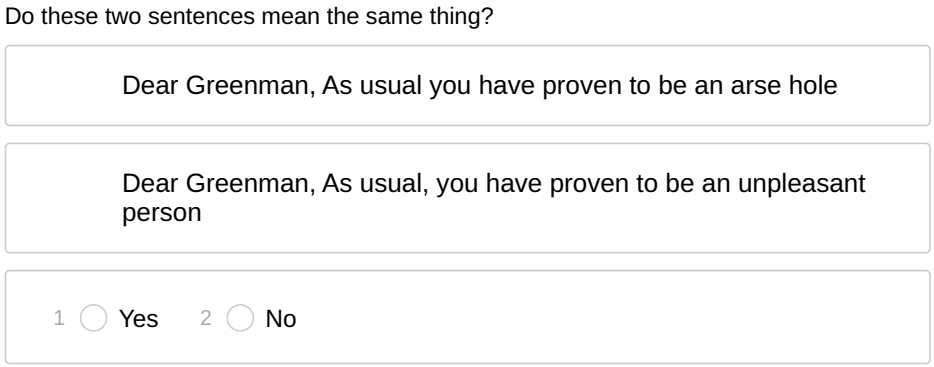

Figure 3: The English interface example of crowdsourcing Task 2: content similarity comparison.

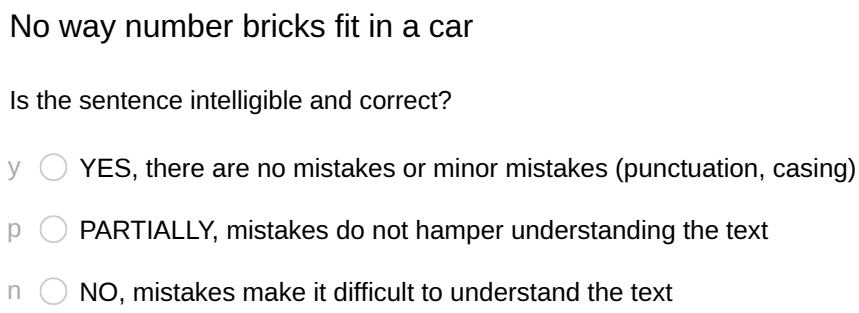

Figure 4: The English interface example of crowdsourcing Task 3: fluency evaluation.

language, the Joint score was determined by multiplying the three individual scores (style transfer accuracy, content similarity, and fluency) using the same formula as in the automatic evaluation.

7. Participants

We received 20 submissions for the development phase leaderboard and 31 submissions for the test phase leaderboard; the final manually evaluated leaderboard was based on 17 submissions who confirmed their participation in the competition. Here, we briefly describe the solutions of our final participants. Each team is presented with its respective leaderboard name (in some cases, additionally, by a username of the corresponding team member that did a submission) and approach used in brackets:

Team cake, Submission d1n910 (few-shot Kimi.AI) [55] The participants achieved the resulting score with a few-short LLM inference by using a two stage process: first, 400 samples from EN and RU provided datasets were used to be detoxified by a proprietary LLM—Kimi.AI [56] which is a large language model chatbot developed by Moonshot AI, a Beijing-based startup. In the second step, the participants employed newly detoxified samples to construct a prompt where they were included as examples of the desired behavior and the model Kimi.AI, thus, was prompted to perform detoxification in target languages.

Team SINAI, Submission estrella (Tree-of-Thought with GPT-3.5) [57] To get the results, Team SINAI employed the Tree-of-Thought prompting strategy based on the OpenAI's model GPT-3.5 [58]. Given a toxic sentence, the model was prompted to output three options of potential detoxified sentences. Then the model was asked to decide in terms of offensiveness, content, and fluency which one out these sentences was detoxified the most appropriate way.

Team MarSanAI, Submission maryam.najafi (Mistral-7b with PPO) [59] This team offered a solution only for two languages: English and Russian. A reinforcement learning method was utilized to fine-tune an LLM—Mistral-7b [60]—coupled with a Proximal Policy Optimization (PPO) [61] using the implementation from HuggingFace TRL [62]; the reward was obtained using the provided toxicity classifier.

Team Linguistic_Hygienist, Submission gangopsa (T5 & BART) [63] The solution consisted of two components: i) the supervised solution for the English and Russian languages; ii) the unsupervised solution for the other 7 languages. The supervised solution used T5 [64] and BART [65] as base models; the exponentially weighted moving average and ROUGE scores were used as loss functions for Russian and English, respectively. The unsupervised solution utilized hashing techniques, log odds ratio, and grammatical rules to identify and conceal toxic words across other 7 languages; additionally, it incorporated a mask prediction model to maintain the original sentences meaning intact.

Team VitalyProtasov (mT0-large) [66] In the proposed solution, the team used a text-to-text model—mT0-large [67]—which was trained on different combinations of languages. In addition, before training, certain filtering techniques were applied to the data.

Team nikita.sushko (mT0-XL) [68] The participant used the text-to-text mT0-XL [67] model that was fine-tuned in two stages. In the first stage, a model was fine-tuned on the parallel data of all languages; this model was used to generate synthetic parallel data from non-parallel samples. The resulting data was cleaned and filtered using a cosine distance between LaBSE embeddings and the toxicity scores by the provided classification models followed by a modification with improved delete approach. At the end, the synthetic and filtered "golden" data were merged into new training set to fine-tune a new instance of the text-to-text multilingual model.

Team SmurfCat, Submission adugeen (mT0-XL) [69] Multilingual model mT0-XL [67] was as well used by this team. First, the model was fine-tuned for text generation using a combination of parallel and translated datasets. The model was further aligned with the Odds Ratio Preference Optimization (ORPO) [70]. During the inference stage, the best candidate generated by the model was chosen by calculating scores from STA and SIM models.

Team gleb.shnshn (zero-shot LLaMa-3) This solution was based on a modern open-source LLM—LlaMa3-70B [71]. The model was prompted using the zero-shot prompting method for the detoxification task.

Team memu_pro_kotow, Submission SomethingAwful (few-shot LLaMa-3 & mT0-XL) [72] In this solution, "uncensored" LLaMa3 [71] was introduced and initialized for every target language except Amharic. Using the recent alignment jailbreaking method by identifying "refusal" directions and subtracting them from model weights [73], they used LLaMa3-70B to get predictions using a few-shot prompting strategy. So, the model received 10 examples of detoxification via starting prompt. For the Amharic language, the text-to-text mT0-XL [67] model was used: the model was fine-tuned on the Amharic parallel dataset.

Team Magnifying_Glass, Submission ZhongyuLuo (Translation & BART-detox, ruT5-detox & Postprocessing) [74] The team used a combination of different methods and models depending on the language. For the majority of languages, the participant used a text-to-text encoder-decoder NLLB translation model [19] to translate data from various languages into English. Then, the translated data was detoxified using the English BART-detox model [13]. After that, the resulting parallel synthetic data was translated back into the original languages. For Russian, the specifically Russian text-to-text model—ruT5-base-detox [22]—was employed for the detoxification. In the case of Chinese, the

participants, firstly, applied filtering of the training dataset, fine-tuned the pre-trained ruGPT3 [75] model, and applied the Delete method.

Team nlp_enjoyers, Submission shredder67 (mT5) [76] The participant employed a text-to-text model mT5 [51]. The provided multilingual parallel data from the development phase was used for fine-tuning.

Team NaiveNeuron, Submission erehulka (few-shot LLaMa-3) [77] The team used a text-to-text Llama3 [71] which was prompted using a few-shot method.

Team team0, Submission dkenco (few-shot Cotype-7b) In this case, the team put a stress solely on the English and Russian languages. Two language-specific approaches were used based on Cotype-7b model [78]. For English, there was employed a zero-shot prompting technique where the prompt included brief instructions for the text detoxification task. For the Russian language, the team used a few-shot approach: the system prompt included brief instructions for the task to be performed as well as five randomly picked samples from the parallel development set. During inference, for both languages, there were applied regular expressions intended as filters.

Team NLPunks, Submission bmmikheev (few-shot LlaMa-3) This team used a text-to-text Llama3-70B [71] by with a few-shot prompting method. For English and Russian, the generated output was evaluated manually. For other languages, GPT-3.5 [58] was used to evaluate outputs. For all languages, the system prompt was formulated in English.

Team Iron Autobots, Submission razvor (few-shot LlaMa-3) The participant as well used a text-to-text Llama3-70b [71] with a few-shot prompting method.

Team MBZUAI-UnbabelDetox, Submission mkrisnai (few-shot GPT-3.5) In this team, a two-step prompting approach was utilized. At the first step, GPT-3.5 [58] was prompted with a few-shot method to produce synthetic detoxification data. Then, the resulting data was employed in the prompt to GPT-3.5 to perform detoxification.

Team Yekaterina29 (mT5-XL) The participant fine-tuned mT5-XL instances [51] on the provided development multilingual parallel dataset.

Almost all of the participant used the current state-of-the-art Large Language Models (LLMs), among which are GPT-3.5 [58] and LLaMa-3 [71] models. To enhance the model's performance on the task of detoxification, most participants used the few-shot prompting method. Among smaller models, mT5 [51] and mT0 [67] were utilized: usually, these models were fine-tuned using ad hoc filtering and data augmentation techniques, for instance, as RAG and backtranslation. Additionally, region-specific LLMs were also employed—Cotype [78] and Kimi.AI [56].

8. Results

Here, we provide the final results of the final test phase, of our tasks. The full detailed tables of results per each language and per each metric can be found in Appendix A.

8.1. Automatic Evaluation Leaderboard

We received 20 submissions for the development phase automatic leaderboard and 31 submissions for the test phase automatic leaderboard. Automatic evaluation leaderboards are publicly available online.[20] The final leaderboard from the test automatic phase evaluation is presented in Table 3.

[20] https://codalab.lisn.upsaclay.fr/competitions/18243#results

The leading solutions were consistent across most languages, except for Spanish, Chinese, and Hindi. However, with the automatic evaluation leaderboard publicly available to all participants, some teams focused on optimizing their models specifically for the evaluation metrics, leading to potential overfitting.

Most solutions surpassed the baseline for at least one language, and in some cases, participant systems approached the performance of human references. However, except for Hindi, no participant solution outperformed human references in the automatic evaluation across any language. Although the automatic evaluation scores for human references across most languages hovered around a J score of 0.7, the results for Chinese were notably poor, with the highest participant score being 0.178 and the best human reference score at 0.201. The results leads to a further investigations of the robustness of the automatic evaluation metrics.

The top three teams across the majority of the languages generally employed a similar strategy, fine-tuning the mT0-XL text-to-text model. Team **SmurfCat** is holding the best automatic evaluation scores for all the languages, which was achieved by additionally fine-tuning mT0-XL with a recent ORPO alignment method. The majority of the submissions were multilingual, designed to cover all languages within a single model. These models demonstrated consistent score distributions across languages, with notable declines in performance for Chinese and Hindi. An exception was user *ansafronov*, who achieved the top score specifically for Chinese.

Table 3
Results of the *automatic* evaluation of the test phase. Scores are sorted by the average **J**oint score. The scores for each language are respective **J**oint scores. Baselines are highlighted with gray, Human References are highlighted with green. Three best scores for each language are highlighted with **bold**, the best score is **underlined bold**.

Team	Average*	EN	ES	DE	ZH	AR	HI	UK	RU	AM
Human References	0.608	0.711	0.709	0.733	0.201	0.695	0.298	0.790	0.732	0.601
Team SmurfCat	**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
SomethingAwful	0.431	0.522	0.475	0.551	0.147	0.514	0.269	0.584	0.516	0.299
mareksuppa	0.424	0.537	0.492	0.577	0.156	0.547	0.181	0.615	0.540	0.173
kofeinix	0.395	0.497	0.420	0.502	0.095	0.501	0.189	0.569	0.490	0.298
Yekaterina29	0.372	0.510	0.439	0.479	0.131	0.453	0.173	0.553	0.507	0.102
AlekseevArtem	0.366	0.427	0.401	0.465	0.071	0.465	0.217	0.562	0.406	0.278
Team NLPunks	0.364	0.489	0.458	0.487	0.150	0.415	0.212	0.466	0.402	0.194
pavelshtykov	0.364	0.489	0.458	0.487	0.150	0.415	0.212	0.466	0.402	0.194
gleb.shnshn	0.359	0.462	0.437	0.464	0.155	0.415	0.244	0.460	0.445	0.147
Volodimirich	0.342	0.472	0.410	0.388	0.095	0.431	0.181	0.483	0.452	0.169
ansafronov	0.340	0.506	0.319	0.362	**0.178**	0.456	0.133	0.328	0.507	0.270
MOOsipenko	0.326	0.411	0.352	0.326	0.067	0.442	0.104	0.474	0.507	0.252
mkrisnai	0.324	0.475	0.422	0.396	0.109	0.270	0.194	0.460	0.383	0.205
Team MarSanAI	0.316	0.504	0.305	0.315	0.069	0.456	0.105	0.315	0.508	0.269
Team nlp_enjoyers	0.316	0.418	0.359	0.384	0.104	0.389	0.172	0.432	0.431	0.157
Team cake	0.316	0.408	0.361	0.503	0.086	0.283	0.158	0.471	0.394	0.178
mT5	0.315	0.418	0.359	0.384	0.096	0.389	0.170	0.433	0.432	0.157
gangopsa	0.315	0.472	0.356	0.414	0.069	0.425	0.198	0.528	0.090	0.280
Team SINAI	0.309	0.413	0.404	0.403	0.126	0.283	0.225	0.436	0.397	0.097
Delete	0.302	0.447	0.319	0.362	0.175	0.456	0.105	0.328	0.255	0.270
Team Iron Autobots	0.288	0.345	0.351	0.364	0.124	0.373	0.204	0.404	0.367	0.058
LanaKlitotekhnis	0.253	0.460	0.161	0.298	0.062	0.274	0.110	0.341	0.384	0.184
Anastasia1706	0.242	0.349	0.271	0.191	0.064	0.404	0.088	0.334	0.248	0.227
ZhongyuLuo	0.240	0.506	0.330	0.024	0.052	0.225	0.138	0.284	0.507	0.096
cocount	0.210	0.271	0.265	0.320	0.100	0.315	0.079	0.245	0.214	0.080
Backtranslation	0.205	0.506	0.275	0.233	0.027	0.206	0.104	0.201	0.223	0.075
etomoscow	0.204	0.293	0.244	0.197	0.025	0.149	0.092	0.266	0.507	0.067
cointegrated	0.175	0.160	0.265	0.245	0.050	0.183	0.070	0.253	0.223	0.123
dkenco	0.163	0.183	0.090	0.287	0.069	0.294	0.035	0.032	0.265	0.217
FD	0.144	0.061	0.189	0.166	0.069	0.294	0.035	0.215	0.048	0.217
Duplicate	0.126	0.061	0.090	0.287	0.069	0.294	0.035	0.032	0.048	0.217

8.2. Human Evaluation Leaderboard

After participants confirmed their submissions via a form, we received 17 entries for the human evaluation phase. This evaluation was conducted on a subsample of 100 test set items through crowdsourcing. The results of the human evaluation, organized by team and language, are publicly available.[21] The **final leaderboard** based on human evaluation is presented in Table 4.

Table 4
Results of the *human* final evaluation of the test phase. Scores are sorted by the average **J**oint score. The scores for each language are respective **J**oint scores. Baselines are highlighted with gray, Human References are highlighted with green. Three best scores for each language are highlighted with **bold**, the best score is **underlined bold**.

Team	Average*	EN	ES	DE	ZH	AR	HI	UK	RU	AM
Human References	0.851	0.885	0.794	0.715	0.925	0.823	0.965	0.902	0.797	0.852
SomethingAwful	**0.774**	0.864	**0.834**	**0.889**	0.534	0.741	**0.863**	**0.686**	**0.839**	**0.715**
Team SmurfCat	**0.741**	0.832	0.726	0.697	0.598	**0.819**	0.683	**0.840**	**0.760**	**0.715**
VitalyProtasov	**0.723**	0.691	**0.810**	0.775	0.493	**0.788**	**0.873**	0.666	0.733	0.680
nikita.sushko	0.712	0.702	0.618	**0.792**	0.474	**0.885**	**0.840**	0.674	0.743	0.680
erehulka	0.708	0.879	0.709	**0.850**	**0.678**	0.778	0.520	0.627	0.646	0.686
Team NLPunks	0.685	0.842	0.764	0.785	**0.604**	0.692	0.780	0.632	0.508	0.563
mkrisnai	0.681	**0.890**	**0.833**	0.697	0.341	0.629	0.732	**0.734**	**0.784**	0.489
Team cake	0.654	**0.907**	0.768	0.774	**0.838**	0.442	0.340	0.499	0.709	0.611
Yekaterina29	0.639	0.749	0.635	0.737	0.300	0.704	0.664	0.654	0.703	0.603
Team SINAI	0.576	0.858	0.681	0.527	0.334	0.765	0.542	0.658	0.678	0.146
gleb.shnshn	0.564	0.737	0.676	0.545	0.408	0.544	0.647	0.436	0.614	0.471
Delete	0.560	0.470	0.551	0.574	0.426	0.649	0.653	0.598	0.491	0.629
mT5	0.541	0.677	0.472	0.635	0.435	0.627	0.601	0.416	0.399	0.608
Team nlp_enjoyers	0.524	0.670	0.423	0.546	0.231	0.558	0.666	0.421	0.502	0.698
Team Iron Autobots	0.516	0.741	0.536	0.647	0.527	0.617	0.583	0.478	0.449	0.065
ZhongyuLuo	0.513	0.735	0.519	0.009	0.564	0.486	0.485	0.417	0.679	**0.724**
gangopsa	0.500	0.741	0.200	0.718	0.374	0.613	0.750	0.484	0.003	0.615
Backtranslation	0.411	0.726	0.557	0.343	0.344	0.417	0.326	0.226	0.221	0.544
Team MarSanAI	0.177	**0.889**	—	—	—	—	—	—	0.704	—
dkenco	0.119	0.679	—	—	—	—	—	—	0.392	—

The human evaluation leaderboard saw significant changes compared to the automatic evaluation phase. Human references received high scores from the annotators, with J scores around 0.8 or higher. However, not all teams surpassed the mT5 and Delete baselines. Interestingly, the Delete baseline outperformed the mT5 text-to-text generation baseline in languages such as Arabic, Hindi, Ukrainian, Russian, and Amharic. This indicates that not all multilingual models are equally proficient in understanding and handling toxicity across different languages.

In the human evaluation phase, participants' solutions closely matched the human references, even surpassing the provided references from parallel datasets in some languages. The top solution, after manual evaluation, was presented by user **SomethingAwful** and was based on the "uncensored" LLaMa3-70B language model. Interestingly, **SomethingAwful**'s solution did not achieve the highest scores across all nine languages but excelled in Spanish, German, and Russian. The leader of the automatic evaluation leaderboard, Team **SmurfCat**, secured second place. Participants **nikita.sushko** and **VitalyProtasov** switched places in the manual leaderboard.

Similar to the automatic leaderboard, human assessments revealed that certain models excelled in specific languages. For instance, *nikita.sushko* and *VitalyProtasov* achieved top results in Arabic and Hindi. Despite Team *mkrisnai* ranking 7th overall, their solution performed exceptionally well in

[21] https://github.com/textdetox/textdetox_clef_2024/tree/main/human_evaluation_results

English, Spanish, Russian, and Ukrainian. Additionally, Team *Team cake* secured the highest scores specifically for English and Chinese.

From the detailed results in Appendix A, it is evident that the solutions surpassed human references in English, Spanish, and German, often achieving near-perfect fluency. However, this success does not extend to other languages. These results highlight the impressive human-like text generation capabilities of modern LLMs, though they still struggle with handling toxicity and maintaining consistent controllable generation across languages. Future work should focus on developing more challenging tasks, particularly in cross-lingual contexts.

9. Conclusion

In Multilingual Text Detoxification task at PAN 2024, participants were tasked with transforming text style from toxic to non-toxic across nine languages: English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, and Amharic. The task was divided into two phases: (i) *development* phase focused on cross-lingual transfer approaches; (ii) *test* phase utilized parallel training data for all languages and encouraged multilingual solutions. Participants' submissions in both phases underwent evaluation using a set of automatic metrics, followed by human evaluation of the test subset to determine the final leaderboard rankings.

Participants employed modern state-of-the-art Large Language Models either by prompting them in few-shot formats or fine-tuning medium-sized models. For certain languages with sufficient training data, these models approached or even exceeded human reference provided in the shared task. However, this was primarily observed for resource-rich European languages. Opportunities for enhancement remain significant for less resource-rich languages and those with limited data, highlighting the need for further exploration in cross-lingual text detoxification and knowledge transfer.

Acknowledgment

We express our deepest gratitude to Toloka platform to support our shared task. Crowdsourced data collection and human evaluation were made possible through the provided research grant.

References

[1] Z. R. Shi, C. Wang, F. Fang, Artificial intelligence for social good: A survey, CoRR abs/2001.01818 (2020). URL: http://arxiv.org/abs/2001.01818. arXiv:2001.01818.

[2] Y. Yao, J. Duan, K. Xu, Y. Cai, E. Sun, Y. Zhang, A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly, CoRR abs/2312.02003 (2023). URL: https://doi.org/10.48550/arXiv.2312.02003. doi:10.48550/ARXIV.2312.02003. arXiv:2312.02003.

[3] J. Cobbe, Algorithmic censorship by social platforms: Power and resistance, Philosophy & Technology 34 (2021) 739–766.

[4] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A benchmark dataset for explainable hate speech detection, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, AAAI Press, 2021, pp. 14867–14875. URL: https://doi.org/10.1609/aaai.v35i17.17745. doi:10.1609/AAAI.V35I17.17745.

[5] J. M. Molero, J. Pérez-Martín, Á. Rodrigo, A. Peñas, Offensive language detection in spanish social media: Testing from bag-of-words to transformers models, IEEE Access 11 (2023) 95639–95652. URL: https://doi.org/10.1109/ACCESS.2023.3310244. doi:10.1109/ACCESS.2023.3310244.

[6] A. A. Ayele, S. M. Yimam, T. D. Belay, T. Asfaw, C. Biemann, Exploring Amharic hate speech data collection and classification approaches, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the 14th

International Conference on Recent Advances in Natural Language Processing, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 49–59. URL: https://aclanthology.org/2023.ranlp-1.6.

[7] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, M. Shrivastava, A dataset of Hindi-English code-mixed social media text for hate speech detection, in: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018, pp. 36–41. URL: https://aclanthology.org/W18-1105. doi:10.18653/v1/W18-1105.

[8] M. R. Costa-jussà, M. C. Meglioli, P. Andrews, D. Dale, P. Hansanti, E. Kalbassi, A. Mourachko, C. Ropers, C. Wood, Mutox: Universal multilingual audio-based toxicity dataset and zero-shot detector, CoRR abs/2401.05060 (2024). URL: https://doi.org/10.48550/arXiv.2401.05060. doi:10.48550/ARXIV.2401.05060. arXiv:2401.05060.

[9] E. Kulenović, Should democracies ban hate speech? hate speech laws and counterspeech, Ethical Theory and Moral Practice 26 (2023) 511–532.

[10] J. Li, R. Jia, H. He, P. Liang, Delete, retrieve, generate: a simple approach to sentiment and style transfer, in: M. A. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 1865–1874. URL: https://doi.org/10.18653/v1/n18-1169. doi:10.18653/V1/N18-1169.

[11] C. Nogueira dos Santos, I. Melnyk, I. Padhi, Fighting offensive language on social media with unsupervised text style transfer, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 189–194. URL: https://aclanthology.org/P18-2031. doi:10.18653/v1/P18-2031.

[12] D. Dale, A. Voronov, D. Dementieva, V. Logacheva, O. Kozlova, N. Semenov, A. Panchenko, Text detoxification using large pre-trained neural models, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7979–7996. URL: https://aclanthology.org/2021.emnlp-main.629. doi:10.18653/v1/2021.emnlp-main.629.

[13] V. Logacheva, D. Dementieva, S. Ustyantsev, D. Moskovskiy, D. Dale, I. Krotova, N. Semenov, A. Panchenko, ParaDetox: Detoxification with parallel data, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 6804–6818. URL: https://aclanthology.org/2022.acl-long.469. doi:10.18653/v1/2022.acl-long.469.

[14] K. Atwell, S. Hassan, M. Alikhani, APPDIA: A discourse-aware transformer-based style transfer model for offensive social media conversations, in: N. Calzolari, C. Huang, H. Kim, J. Pustejovsky, L. Wanner, K. Choi, P. Ryu, H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022, International Committee on Computational Linguistics, 2022, pp. 6063–6074. URL: https://aclanthology.org/2022.coling-1.530.

[15] D. Dementieva, D. Moskovskiy, D. Dale, A. Panchenko, Exploring methods for cross-lingual text style transfer: The case of text detoxification, in: J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, A. A. Krisnadhi (Eds.), Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023, Association for Computational Linguistics, 2023, pp. 1083–1101. URL: https://doi.org/10.18653/v1/2023.ijcnlp-main.70. doi:10.18653/V1/2023.IJCNLP-MAIN.70.

[16] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking

Analysis, and Generative AI Authorship Verification, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[17] I. Price, J. Gifford-Moore, J. Flemming, S. Musker, M. Roichman, G. Sylvain, N. Thain, L. Dixon, J. Sorensen, Six attributes of unhealthy conversations, in: S. Akiwowo, B. Vidgen, V. Prabhakaran, Z. Waseem (Eds.), Proceedings of the Fourth Workshop on Online Abuse and Harms, WOAH 2020, Online, November 20, 2020, Association for Computational Linguistics, 2020, pp. 114–124. URL: https://doi.org/10.18653/v1/2020.alw-1.15. doi:10.18653/V1/2020.ALW-1.15.

[18] S. Gilda, L. Giovanini, M. Silva, D. Oliveira, Predicting different types of subtle toxicity in unhealthy online conversations, in: N. Varandas, A. Yasar, H. Malik, S. Galland (Eds.), The 12th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2021) / The 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2021), Leuven, Belgium, November 1-4, 2021, volume 198 of *Procedia Computer Science*, Elsevier, 2021, pp. 360–366. URL: https://doi.org/10.1016/j.procs.2021.12.254. doi:10.1016/J.PROCS.2021.12.254.

[19] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Y. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, J. Wang, No language left behind: Scaling human-centered machine translation, CoRR abs/2207.04672 (2022). URL: https://doi.org/10.48550/arXiv.2207.04672. doi:10.48550/ARXIV.2207.04672. arXiv:2207.04672.

[20] D. Dementieva, D. Moskovskiy, V. Logacheva, D. Dale, O. Kozlova, N. Semenov, A. Panchenko, Methods for detoxification of texts for the russian language, Multimodal Technol. Interact. 5 (2021) 54. URL: https://doi.org/10.3390/mti5090054. doi:10.3390/MTI5090054.

[21] D. Moskovskiy, D. Dementieva, A. Panchenko, Exploring cross-lingual text detoxification with large multilingual language models, in: S. Louvan, A. Madotto, B. Madureira (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, 2022, pp. 346–354. URL: https://doi.org/10.18653/v1/2022.acl-srw.26. doi:10.18653/V1/2022.ACL-SRW.26.

[22] D. Dementieva, V. Logacheva, I. Nikishina, A. Fenogenova, D. Dale, I. Krotova, N. Semenov, T. Shavrina, A. Panchenko, RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora, COMPUTATIONAL LINGUISTICS AND INTELLECTUAL TECHNOLOGIES (2022). URL: https://api.semanticscholar.org/CorpusID:253169495.

[23] A. Pavao, I. Guyon, A.-C. Letournel, D.-T. Tran, X. Baro, H. J. Escalante, S. Escalera, T. Thomas, Z. Xu, Codalab competitions: An open source platform to organize scientific challenges, Journal of Machine Learning Research 24 (2023) 1–6. URL: http://jmlr.org/papers/v24/21-1436.html.

[24] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. URL: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.

[25] V. Logacheva, D. Dementieva, I. Krotova, A. Fenogenova, I. Nikishina, T. Shavrina, A. Panchenko, A study on manual and automatic evaluation for text style transfer: The case of detoxification, in: Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 90–101. URL: https://aclanthology.org/2022.humeval-1.8. doi:10.18653/v1/2022.humeval-1.8.

[26] V. Logacheva, D. Dementieva, S. Ustyantsev, D. Moskovskiy, D. Dale, I. Krotova, N. Semenov, A. Panchenko, Paradetox: Detoxification with parallel data, in: S. Muresan, P. Nakov, A. Villav-

icencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, 2022, pp. 6804–6818. URL: https://doi.org/10.18653/v1/2022.acl-long.469. doi:10.18653/V1/2022.ACL-LONG.469.

[27] Jigsaw, Toxic comment classification challenge, https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge, 2017. Accessed: 2024-03-18.

[28] D. Dementieva, N. Babakov, A. Panchenko, Multiparadetox: Extending text detoxification with parallel data to new languages, arXiv preprint arXiv:2404.02037 (2024).

[29] A. Belchikov, Russian language toxic comments, https://www.kaggle.com/blackmoon/russian-language-toxic-comments, 2019. Accessed: 2023-12-14.

[30] A. Semiletov, Toxic Russian Comments: Labelled comments from the popular Russian social network, https://www.kaggle.com/alexandersemiletov/toxic-russian-comments, 2020. Accessed: 2023-12-14.

[31] K. Bobrovnyk, Ukrainian obscene lexicon, https://github.com/saganoren/obscene-ukr, 2019. Accessed: 2023-12-14.

[32] K. Bobrovnyk, Automated building and analysis of ukrainian twitter corpus for toxic text detection, in: COLINS 2019. Volume II: Workshop, 2019. URL: https://ena.lpnu.ua:8443/server/api/core/bitstreams/c4c645c1-f465-4895-98dd-765f862cf186/content.

[33] J. C. Pereira-Kohatsu, L. Q. Sánchez, F. Liberatore, M. Camacho-Collados, Detecting and monitoring hate speech in twitter, Sensors 19 (2019) 4654. URL: https://doi.org/10.3390/s19214654. doi:10.3390/S19214654.

[34] M. Taulé, M. Nofre, V. Bargiela, X. Bonet, Newscom-tox: a corpus of comments on news articles annotated for toxicity in spanish, Language Resources and Evaluation (2024) 1–41.

[35] J. M. Pérez, D. A. Furman, L. Alonso Alemany, F. M. Luque, RoBERTuito: a pre-trained language model for social media text in Spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7235–7243. URL: https://aclanthology.org/2022.lrec-1.785.

[36] M. Wiegand, M. Siegel, J. Ruppenhofer, Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language (2018).

[37] J. Risch, A. Stoll, L. Wilms, M. Wiegand, Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments, in: Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments, Duesseldorf, Germany, 2021, pp. 1–12.

[38] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, M. Wojatzki, Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis, in: M. Beißwenger, M. Wojatzki, T. Zesch (Eds.), Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, volume 17 of *Bochumer Linguistische Arbeitsberichte*, Bochum, Germany, 2016, pp. 6–.9.

[39] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 14–17. URL: https://doi.org/10.1145/3368567.3368584. doi:10.1145/3368567.3368584.

[40] A. A. Ayele, S. Dinter, T. D. Belay, T. T. Asfaw, S. M. Yimam, C. Biemann, The 5Js in Ethiopia: Amharic hate speech data annotation using Toloka Crowdsourcing Platform, in: Proceedings of the 4th International Conference on Information and Communication Technology for Development for Africa (ICT4DA), Bahir Dar, Ethiopia, 2022, pp. 114–120. URL: https://ieeexplore.ieee.org/document/9971189.

[41] H. Mulki, H. Haddad, C. B. Ali, H. Alshabani, L-hsab: A levantine twitter dataset for hate speech and abusive language, in: Proceedings of the third workshop on abusive language online, 2019, pp. 111–118.

[42] H. Haddad, H. Mulki, A. Oueslati, T-hsab: A tunisian hate speech and abusive dataset, in:

International conference on Arabic language processing, Springer, 2019, pp. 251–263.
[43] H. Mubarak, K. Darwish, W. Magdy, T. Elsayed, H. Al-Khalifa, Overview of osact4 arabic offensive language detection shared task, in: Proceedings of the 4th Workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection, 2020, pp. 48–52.
[44] H. Mulki, B. Ghanem, Let-mi: An Arabic Levantine Twitter dataset for misogynistic language, in: Proceedings of the Sixth Arabic Natural Language Processing Workshop, Association for Computational Linguistics, Kyiv, Ukraine (Virtual), 2021, pp. 154–163. URL: https://aclanthology.org/2021.wanlp-1.16.
[45] J. Lu, B. Xu, X. Zhang, C. Min, L. Yang, H. Lin, Facilitating fine-grained detection of Chinese toxic language: Hierarchical taxonomy, resources, and benchmarks, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023, pp. 16235–16250. URL: https://aclanthology.org/2023.acl-long.898.
[46] D. Dementieva, S. Ustyantsev, D. Dale, O. Kozlova, N. Semenov, A. Panchenko, V. Logacheva, Crowdsourcing of parallel corpora: the case of style transfer for detoxification, in: D. Ustalov, F. Casati, A. Drutsa, I. Stelmakh, N. Pavlichenko, D. Baidakova (Eds.), Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale co-located with 47th International Conference on Very Large Data Bases (VLDB 2021), Copenhagen, Denmark, August 20, 2021, volume 2932 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 35–49. URL: https://ceur-ws.org/Vol-2932/paper2.pdf.
[47] R. J. Gabriel, English full list of bad words and top swear words banned by google, https://github.com/coffee-and-fun/google-profanity-words/blob/main/data/en.txt, 2023. Accessed: 2023-12-12.
[48] K. Bobrovnyk, The dictionary of ukrainian obscene words, https://github.com/saganoren/obscene-ukr, 2019. Accessed: 2023-12-12.
[49] I. Shutterstock, List of dirty, naughty, obscene, and otherwise bad words, https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words, 2020. Accessed: 2023-12-12.
[50] A. Jiang, X. Yang, Y. Liu, A. Zubiaga, SWSR: A chinese dataset and lexicon for online sexism detection, Online Soc. Networks Media 27 (2022) 100182. URL: https://doi.org/10.1016/j.osnem.2021.100182. doi:10.1016/J.OSNEM.2021.100182.
[51] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics, 2021, pp. 483–498. URL: https://doi.org/10.18653/v1/2021.naacl-main.41. doi:10.18653/V1/2021.NAACL-MAIN.41.
[52] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://doi.org/10.18653/v1/2020.acl-main.747. doi:10.18653/V1/2020.ACL-MAIN.747.
[53] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic BERT sentence embedding, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, 2022, pp. 878–891. URL: https://doi.org/10.18653/v1/2022.acl-long.62. doi:10.18653/V1/2022.ACL-LONG.62.
[54] M. Post, A call for clarity in reporting BLEU scores, in: O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. Jimeno-Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. L. Neves, M. Post, L. Specia, M. Turchi, K. Verspoor (Eds.), Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels,

October 31 - November 1, 2018, Association for Computational Linguistics, 2018, pp. 186–191. URL: https://doi.org/10.18653/v1/w18-6319. doi:10.18653/V1/W18-6319.

[55] J. Peng, Z. Han, H. Zhang, J. Ye, C. Liu, B. Liu, M. Guo, H. Chen, Z. Lin, Y. Tang, A Multilingual Text Detoxification Method Based on Few-shot Learning and CO-STAR Framework, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[56] M. AI, Kimi chatbot, 2024. URL: https://kimi.moonshot.cn, accessed: 2024-05-31.

[57] M. Vallecillo-Rodríguez, A. M. Martín-Valdivia, SINAI at PAN 2024 TextDetox: Application of Tree of Thought Strategy in Large Language Models for Multilingual Text Detoxification, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[58] OpenAI, Chatgpt: Optimizing language models for dialogue, 2022. URL: https://openai.com/blog/chatgpt, accessed: 2024-05-31.

[59] M. Najafi, E. Tavan, S. Colreavy, Marsan at PAN 2024 TextDetox: ToxiCleanse RL and Paving the Way for Toxicity-Free Online Discourse, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[60] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, CoRR abs/2310.06825 (2023). URL: https://doi.org/10.48550/arXiv.2310.06825. doi:10.48550/ARXIV.2310.06825. arXiv:2310.06825.

[61] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, CoRR abs/1707.06347 (2017). URL: http://arxiv.org/abs/1707.06347. arXiv:1707.06347.

[62] L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang, Trl: Transformer reinforcement learning, https://github.com/huggingface/trl, 2020.

[63] S. Gangopadhyay, M. Khan, H. Jabeen, HybridDetox: Combining Supervised and Unsupervised Methods for Effective Multilingual Text Detoxification, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[64] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67. URL: http://jmlr.org/papers/v21/20-074.html.

[65] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 7871–7880. URL: https://doi.org/10.18653/v1/2020.acl-main.703. doi:10.18653/V1/2020.ACL-MAIN.703.

[66] V. Protasov, PAN 2024 Multilingual TextDetox: Exploring Cross-lingual Transfer in Case of Large Language Models, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[67] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. L. Scao, M. S. Bari, S. Shen, Z. X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, C. Raffel, Crosslingual generalization through multitask finetuning, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, Association for Computational Linguistics, 2023, pp. 15991–16111. URL: https://doi.org/10.18653/v1/2023.acl-long.891. doi:10.18653/V1/2023.ACL-LONG.891.

[68] N. Sushko, PAN 2024 Multilingual TextDetox: Exploring Different Regimes For Synthetic Data Training For Multilingual Text Detoxification, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[69] E. Rykov, K. Zaytsev, I. Anisimov, A. Voronin, SmurfCat at PAN TexDetox 2024: Alignment of Multilingual Transformers for Text Detoxification, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[70] J. Hong, N. Lee, J. Thorne, ORPO: monolithic preference optimization without reference model, CoRR abs/2403.07691 (2024). URL: https://doi.org/10.48550/arXiv.2403.07691. doi:10.48550/ARXIV.2403.07691. arXiv:2403.07691.

[71] AI@Meta, Llama 3 model card (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[72] S. Pletenev, Memu_pro_kotow at PAN 2024 TextDetox: Uncensored Llama3 Helps to Censor Better, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[73] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A. Dombrowski, S. Goel, N. Li, M. J. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, J. Z. Kolter, D. Hendrycks, Representation engineering: A top-down approach to AI transparency, CoRR abs/2310.01405 (2023). URL: https://doi.org/10.48550/arXiv.2310.01405. doi:10.48550/ARXIV.2310.01405. arXiv:2310.01405.

[74] Z. Luo, M. Luo, A. Wang, Multilingual Text Detoxification Using Google Cloud Translation and Post-Processing, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[75] A. Forever, rugpt3, 2022. URL: https://huggingface.co/ai-forever, accessed: 2024-05-31.

[76] V. Zinkovich, S. Karpukhin, N. Kurdiukov, P. Tikhomirov, nlp_enjoyers at Multilingual Textual Detoxification (CLEF-2024, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[77] E. Řehulka, M. Šuppa, RAG Meets Detox: Enhancing Text Detoxification Using Open-Source Large Language Models with Retrieval Augmented Generation, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[78] MTS.AI, Cotype: Generative ai solutions, 2022. URL: https://mts.ai, accessed: 2024-05-31.

A. Automatic and Manual Evaluation Results per Language

Here, we provide the extended results—from both automatic and human evaluation setups—based on three evaluation parameters for all languages: English (Table 5), Spanish (Table 6), German (Table 7), Chinese (Table 8), Arabic (Table 9), Hindi (Table 10), Ukrainian (Table 11), Russian (Table 12), and Amharic (Table 13). In every table, the baselines are highlighted with gray; Human References are highlighted with green; the ordering is made by **J** score from **Human Evaluation** results. The automatic evaluation is based on the full test set of 600 samples per language; human evaluation was performed on 100 set of the test set per language.

Table 5
Automatic and human evaluation results for English.

	Automatic Evaluation				Human Evaluation			
	STA	**SIM**	**ChrF**	**J**	**STA**	**SIM**	**FL**	**J***
Team cake	0.911	0.790	0.542	0.407	0.965	0.940	1.000	0.907
mkrisnai	0.807	0.865	0.661	0.475	0.946	0.950	0.990	0.890
Team MarSanAI	0.788	0.859	0.723	0.504	0.955	0.980	0.950	0.889
Human References	0.864	0.820	1.000	0.711	0.970	0.960	0.950	0.884
erehulka	0.871	0.869	0.697	0.543	0.976	0.900	1.000	0.879
SomethingAwful	0.876	0.860	0.670	0.522	0.968	0.910	0.980	0.863
Team SINAI	0.910	0.787	0.553	0.412	0.953	0.900	1.000	0.858
Team NLPunks	0.875	0.849	0.635	0.489	0.945	0.900	0.990	0.841
Team SmurfCat	0.934	0.886	0.706	0.601	0.973	0.900	0.950	0.832
Yekaterina29	0.793	0.879	0.704	0.509	0.963	0.810	0.960	0.749
Team Iron Autobots	0.969	0.706	0.477	0.344	0.938	0.790	1.000	0.741
gangopsa	0.737	0.888	0.698	0.473	0.897	0.878	0.939	0.740
gleb.shnshn	0.870	0.773	0.661	0.462	0.966	0.770	0.990	0.736
ZhongyuLuo	0.807	0.868	0.693	0.506	0.953	0.820	0.940	0.734
Backtranslation	0.807	0.868	0.693	0.506	0.941	0.820	0.940	0.725
nikita.sushko	0.851	0.892	0.710	0.552	0.971	0.760	0.950	0.701
VitalyProtasov	0.841	0.864	0.699	0.531	0.970	0.750	0.950	0.691
dkenco	0.951	0.567	0.311	0.182	0.956	0.710	1.000	0.679
mT5	0.676	0.868	0.670	0.417	0.906	0.770	0.970	0.677
Team nlp_enjoyers	0.676	0.868	0.670	0.417	0.908	0.760	0.970	0.669
Delete	0.662	0.956	0.691	0.447	0.848	0.630	0.880	0.470

Table 6
Automatic and human evaluation results for Spanish.

	Automatic Evaluation				Human Evaluation			
	STA	**SIM**	**ChrF**	**J**	**STA**	**SIM**	**FL**	**J***
SomethingAwful	0.885	0.830	0.625	0.475	0.916	0.910	1.000	0.834
mkrisnai	0.867	0.806	0.584	0.421	0.886	0.940	1.000	0.833
VitalyProtasov	0.892	0.835	0.619	0.472	0.910	0.890	1.000	0.809
Human References	0.875	0.811	1.000	0.708	0.901	0.890	0.990	0.794
Team cake	0.928	0.765	0.488	0.360	0.891	0.870	0.990	0.767
Team NLPunks	0.861	0.848	0.615	0.458	0.906	0.860	0.980	0.764
Team SmurfCat	0.959	0.885	0.644	0.562	0.871	0.850	0.980	0.726
erehulka	0.884	0.865	0.634	0.496	0.930	0.770	0.990	0.708
Team SINAI	0.899	0.781	0.546	0.404	0.851	0.800	1.000	0.681
gleb.shnshn	0.900	0.799	0.584	0.436	0.890	0.760	1.000	0.676
Yekaterina29	0.745	0.888	0.646	0.439	0.835	0.760	1.000	0.634
nikita.sushko	0.788	0.896	0.657	0.480	0.866	0.720	0.990	0.617
Backtranslation	0.812	0.770	0.423	0.275	0.865	0.650	0.990	0.556
Delete	0.479	0.972	0.669	0.318	0.685	0.830	0.970	0.551
Team Iron Autobots	0.947	0.742	0.479	0.351	0.933	0.580	0.990	0.535
ZhongyuLuo	0.808	0.810	0.483	0.329	0.831	0.630	0.990	0.518
mT5	0.649	0.873	0.616	0.358	0.796	0.630	0.940	0.471
Team nlp_enjoyers	0.653	0.870	0.616	0.359	0.775	0.600	0.910	0.423
gangopsa	0.788	0.822	0.542	0.356	0.810	0.280	0.880	0.199

Table 7
Automatic and human evaluation results for German.

	Automatic Evaluation				Human Evaluation			
	STA	**SIM**	**ChrF**	**J**	**STA**	**SIM**	**FL**	**J***
SomethingAwful	0.799	0.904	0.759	0.550	0.898	0.990	1.000	0.889
erehulka	0.829	0.899	0.760	0.574	0.923	0.930	0.990	0.850
nikita.sushko	0.774	0.940	0.808	0.591	0.833	0.950	1.000	0.791
Team NLPunks	0.820	0.867	0.670	0.487	0.891	0.880	1.000	0.784
VitalyProtasov	0.646	0.951	0.813	0.502	0.798	0.980	0.990	0.774
Team cake	0.795	0.887	0.710	0.502	0.890	0.870	1.000	0.774
Yekaterina29	0.807	0.869	0.671	0.478	0.896	0.830	0.990	0.736
gangopsa	0.651	0.892	0.714	0.413	0.788	0.980	0.930	0.718
Human References	0.809	0.909	1.000	0.732	0.863	0.920	0.900	0.714
Team SmurfCat	0.921	0.923	0.781	0.677	0.856	0.830	0.980	0.696
mkrisnai	0.683	0.888	0.659	0.395	0.810	0.860	1.000	0.696
Team Iron Autobots	0.934	0.734	0.514	0.364	0.943	0.700	0.980	0.647
mT5	0.746	0.837	0.603	0.383	0.873	0.750	0.970	0.635
Delete	0.454	0.989	0.802	0.361	0.591	0.990	0.980	0.574
Team nlp_enjoyers	0.750	0.835	0.602	0.384	0.870	0.640	0.980	0.545
gleb.shnshn	0.910	0.803	0.617	0.464	0.940	0.580	1.000	0.545
Team SINAI	0.876	0.803	0.563	0.403	0.810	0.650	1.000	0.526
Backtranslation	0.796	0.747	0.372	0.232	0.858	0.400	1.000	0.343
ZhongyuLuo	0.815	0.222	0.130	0.024	0.876	0.010	0.990	0.008

Table 8
Automatic and human evaluation results for Chinese.

	Automatic Evaluation				Human Evaluation			
	STA	**SIM**	**ChrF**	**J**	**STA**	**SIM**	**FL**	**J***
Human References	0.266	0.789	1.000	0.201	0.963	0.990	0.970	0.925
Team cake	0.549	0.665	0.238	0.086	0.930	0.910	0.990	0.837
erehulka	0.389	0.789	0.551	0.160	0.950	0.870	0.820	0.677
Team NLPunks	0.462	0.815	0.395	0.150	0.648	0.980	0.950	0.603
Team SmurfCat	0.529	0.822	0.415	0.177	0.773	0.920	0.840	0.597
ZhongyuLuo	0.633	0.650	0.122	0.051	0.838	0.830	0.810	0.563
SomethingAwful	0.459	0.733	0.449	0.147	0.888	0.770	0.780	0.533
Team Iron Autobots	0.602	0.714	0.284	0.123	0.806	0.860	0.760	0.527
VitalyProtasov	0.411	0.868	0.504	0.175	0.891	0.970	0.570	0.493
nikita.sushko	0.415	0.869	0.504	0.176	0.920	0.990	0.520	0.473
mT5	0.289	0.809	0.411	0.095	0.726	0.920	0.650	0.434
Delete	0.384	0.887	0.524	0.174	0.693	0.990	0.620	0.425
gleb.shnshn	0.531	0.799	0.364	0.154	0.728	0.700	0.800	0.407
gangopsa	0.129	0.999	0.535	0.069	0.511	1.000	0.730	0.373
Backtranslation	0.661	0.591	0.070	0.026	0.831	0.600	0.690	0.344
mkrisnai	0.452	0.805	0.328	0.108	0.653	0.550	0.950	0.341
Team SINAI	0.608	0.741	0.286	0.126	0.558	0.720	0.830	0.333
Yekaterina29	0.344	0.778	0.472	0.130	0.840	0.830	0.430	0.299
Team nlp_enjoyers	0.375	0.770	0.403	0.104	0.778	0.430	0.690	0.230

Table 9
Automatic and human evaluation results for Arabic.

	Automatic Evaluation				Human Evaluation			
	STA	**SIM**	**ChrF**	**J**	**STA**	**SIM**	**FL**	**J***
nikita.sushko	0.780	0.930	0.783	0.575	0.921	0.990	0.970	0.885
Human References	0.795	0.875	1.000	0.694	0.941	0.920	0.950	0.823
Team SmurfCat	0.921	0.890	0.747	0.625	0.918	0.910	0.980	0.818
VitalyProtasov	0.730	0.921	0.775	0.522	0.891	0.930	0.950	0.787
erehulka	0.788	0.896	0.752	0.535	0.920	0.890	0.950	0.777
Team SINAI	0.883	0.699	0.425	0.282	0.921	0.830	1.000	0.764
SomethingAwful	0.825	0.860	0.719	0.513	0.931	0.820	0.970	0.741
Yekaterina29	0.695	0.904	0.710	0.452	0.828	0.850	1.000	0.704
Team NLPunks	0.728	0.857	0.652	0.414	0.866	0.840	0.950	0.691
Delete	0.597	0.974	0.777	0.455	0.750	0.920	0.940	0.648
mkrisnai	0.759	0.755	0.466	0.270	0.796	0.790	1.000	0.629
mT5	0.713	0.841	0.642	0.389	0.868	0.760	0.950	0.626
Team Iron Autobots	0.757	0.809	0.596	0.373	0.828	0.810	0.920	0.617
gangopsa	0.776	0.826	0.643	0.424	0.920	0.900	0.740	0.612
Team nlp_enjoyers	0.718	0.834	0.640	0.388	0.863	0.710	0.910	0.557
gleb.shnshn	0.794	0.825	0.616	0.415	0.920	0.650	0.910	0.544
ZhongyuLuo	0.771	0.719	0.366	0.225	0.832	0.590	0.990	0.486
Team cake	0.917	0.672	0.420	0.282	0.970	0.480	0.950	0.442
Backtranslation	0.836	0.682	0.319	0.205	0.915	0.460	0.990	0.416

Table 10
Automatic and human evaluation results for Hindi.

	Automatic Evaluation				Human Evaluation			
	STA	SIM	ChrF	J	STA	SIM	FL	J*
Human References	0.367	0.814	1.000	0.297	0.975	0.990	1.000	0.965
VitalyProtasov	0.615	0.713	0.680	0.320	0.938	0.940	0.990	0.873
SomethingAwful	0.460	0.826	0.666	0.269	0.948	0.910	1.000	0.862
nikita.sushko	0.351	0.882	0.709	0.240	0.923	0.910	1.000	0.840
Team NLPunks	0.393	0.837	0.613	0.212	0.896	0.870	1.000	0.780
gangopsa	0.351	0.844	0.646	0.197	0.928	0.860	0.940	0.750
mkrisnai	0.476	0.786	0.509	0.193	0.871	0.840	1.000	0.732
Team SmurfCat	0.634	0.799	0.631	0.355	0.961	0.710	1.000	0.682
Team nlp_enjoyers	0.302	0.804	0.619	0.171	0.905	0.800	0.920	0.666
Yekaterina29	0.261	0.905	0.662	0.173	0.790	0.840	1.000	0.663
Delete	0.146	0.974	0.706	0.104	0.673	0.970	1.000	0.653
gleb.shnshn	0.497	0.790	0.595	0.244	0.975	0.670	0.990	0.646
mT5	0.295	0.808	0.620	0.170	0.871	0.690	1.000	0.601
Team Iron Autobots	0.461	0.781	0.550	0.204	0.896	0.650	1.000	0.582
Team SINAI	0.586	0.750	0.490	0.224	0.960	0.570	0.990	0.541
erehulka	0.324	0.806	0.635	0.184	0.940	0.700	0.790	0.519
ZhongyuLuo	0.439	0.773	0.376	0.137	0.816	0.600	0.990	0.485
Team cake	0.771	0.583	0.310	0.157	0.953	0.360	0.990	0.339
Backtranslation	0.443	0.731	0.289	0.103	0.853	0.390	0.980	0.326

Table 11
Automatic and human evaluation results for Ukrainian.

	Automatic Evaluation				Human Evaluation			
	STA	SIM	ChrF	J	STA	SIM	FL	J*
Human References	0.877	0.899	1.000	0.790	0.990	0.980	0.930	0.902
Team SmurfCat	0.951	0.913	0.780	0.691	0.971	0.900	0.960	0.839
mkrisnai	0.895	0.842	0.592	0.460	0.963	0.770	0.990	0.734
SomethingAwful	0.875	0.887	0.733	0.584	0.966	0.710	1.000	0.686
nikita.sushko	0.886	0.919	0.804	0.668	0.965	0.720	0.970	0.673
VitalyProtasov	0.846	0.922	0.792	0.628	0.956	0.710	0.980	0.665
Team SINAI	0.944	0.797	0.551	0.436	0.983	0.690	0.970	0.658
Yekaterina29	0.804	0.891	0.742	0.553	0.940	0.710	0.980	0.654
Team NLPunks	0.771	0.869	0.665	0.466	0.936	0.710	0.950	0.631
erehulka	0.882	0.899	0.743	0.602	0.975	0.670	0.960	0.627
Delete	0.423	0.974	0.791	0.327	0.708	0.870	0.970	0.597
Team cake	0.804	0.863	0.658	0.470	0.966	0.580	0.890	0.498
gangopsa	0.816	0.884	0.721	0.527	0.943	0.540	0.950	0.483
Team Iron Autobots	0.861	0.807	0.561	0.403	0.930	0.530	0.970	0.478
gleb.shnshn	0.857	0.826	0.634	0.460	0.936	0.500	0.930	0.435
Team nlp_enjoyers	0.704	0.856	0.678	0.431	0.905	0.490	0.950	0.421
ZhongyuLuo	0.884	0.773	0.385	0.283	0.966	0.440	0.980	0.416
mT5	0.704	0.858	0.679	0.433	0.911	0.480	0.950	0.415
Backtranslation	0.914	0.704	0.293	0.201	0.981	0.230	1.000	0.225

Table 12
Automatic and human evaluation results for Russian.

	Automatic Evaluation				Human Evaluation			
	STA	SIM	ChrF	J	STA	SIM	FL	J*
SomethingAwful	0.819	0.873	0.695	0.515	0.986	0.850	1.000	0.838
Human References	0.887	0.824	1.000	0.732	0.990	0.830	0.970	0.797
mkrisnai	0.758	0.825	0.600	0.382	0.901	0.870	1.000	0.784
Team SmurfCat	0.957	0.885	0.736	0.634	0.953	0.830	0.960	0.759
nikita.sushko	0.843	0.901	0.728	0.570	0.948	0.800	0.980	0.743
VitalyProtasov	0.807	0.893	0.731	0.542	0.933	0.810	0.970	0.733
Team cake	0.881	0.791	0.540	0.394	0.958	0.740	1.000	0.709
Team MarSanAI	0.779	0.878	0.723	0.507	0.916	0.800	0.960	0.704
Yekaterina29	0.811	0.875	0.689	0.507	0.953	0.760	0.970	0.702
ZhongyuLuo	0.812	0.863	0.705	0.507	0.958	0.770	0.920	0.678
Team SINAI	0.890	0.792	0.533	0.396	0.935	0.740	0.980	0.678
erehulka	0.858	0.868	0.686	0.528	0.975	0.690	0.960	0.645
gleb.shnshn	0.857	0.817	0.627	0.445	0.955	0.670	0.960	0.614
Team NLPunks	0.709	0.858	0.630	0.402	0.938	0.570	0.950	0.508
Team nlp_enjoyers	0.762	0.842	0.638	0.431	0.920	0.580	0.940	0.501
Delete	0.372	0.971	0.708	0.254	0.743	0.750	0.880	0.490
Team Iron Autobots	0.907	0.776	0.506	0.367	0.965	0.490	0.950	0.449
mT5	0.762	0.844	0.638	0.431	0.955	0.440	0.950	0.399
dkenco	0.595	0.817	0.554	0.264	0.825	0.480	0.990	0.392
Backtranslation	0.903	0.697	0.328	0.222	0.970	0.230	0.990	0.220
gangopsa	0.414	0.575	0.345	0.090	0.905	0.180	0.020	0.003

Table 13
Automatic and human evaluation results for Amharic.

	Automatic Evaluation				Human Evaluation			
	STA	SIM	ChrF	J	STA	SIM	FL	J*
Human References	0.893	0.683	1.000	0.601	0.935	0.990	0.920	0.851
ZhongyuLuo	0.819	0.665	0.165	0.095	0.875	0.890	0.930	0.724
SomethingAwful	0.776	0.855	0.438	0.299	0.801	0.980	0.910	0.714
Team SmurfCat	0.900	0.888	0.456	0.378	0.768	1.000	0.930	0.714
Team nlp_enjoyers	0.837	0.640	0.269	0.157	0.863	0.940	0.860	0.697
erehulka	0.586	0.971	0.482	0.286	0.700	1.000	0.980	0.686
nikita.sushko	0.742	0.908	0.478	0.328	0.755	0.990	0.910	0.680
VitalyProtasov	0.754	0.872	0.458	0.310	0.786	0.950	0.910	0.680
Delete	0.539	0.979	0.486	0.269	0.661	1.000	0.950	0.628
gangopsa	0.584	0.956	0.478	0.280	0.690	0.990	0.900	0.614
Team cake	0.559	0.836	0.360	0.178	0.691	0.960	0.920	0.610
mT5	0.836	0.641	0.270	0.157	0.893	0.840	0.810	0.607
Yekaterina29	0.794	0.589	0.204	0.102	0.891	0.980	0.690	0.602
Team NLPunks	0.555	0.865	0.372	0.194	0.743	0.880	0.860	0.562
Backtranslation	0.819	0.618	0.135	0.075	0.856	0.690	0.920	0.543
mkrisnai	0.467	0.946	0.453	0.205	0.515	0.990	0.960	0.489
gleb.shnshn	0.649	0.725	0.298	0.146	0.805	0.960	0.610	0.471
Team SINAI	0.623	0.588	0.234	0.096	0.778	0.520	0.360	0.145
Team Iron Autobots	0.672	0.355	0.118	0.057	0.845	0.250	0.310	0.065