Toxicity Classification in Ukrainian

Daryna Dementieva[1], **Valeriia Khylenko**[1], **Nikolay Babakov**[2] and **Georg Groh**[1]
[1] TU Munich, Department of Informatics, Germany
[2]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela
daryna.dementieva@tum.de, nikolay.babakov@usc.es, grohg@in.tum.de

Abstract

The task of toxicity detection is still a relevant task, especially in the context of safe and fair LMs development. Nevertheless, labeled binary toxicity classification corpora are not available for all languages, which is understandable given the resource-intensive nature of the annotation process. Ukrainian, in particular, is among the languages lacking such resources. To our knowledge, there has been no existing toxicity classification corpus in Ukrainian. In this study, we aim to fill this gap by investigating cross-lingual knowledge transfer techniques and creating labeled corpora by: (i) translating from an English corpus, (ii) filtering toxic samples using keywords, and (iii) annotating with crowdsourcing. We compare LLMs prompting and other cross-lingual transfer approaches with and without fine-tuning offering insights into the most robust and efficient baselines.

This paper contains rude texts that only serve as illustrative examples.

Toxic	І ніх*шеньки їй за те не буде. *And she's not going to get a f*king thing for it.* А зі всіх компліментів які мені казали, це те що я п*ар *And of all the compliments I've been given, the only one I've received is that I'm a f*got.* Увесь твіттер у ваших *бучих котах. *The whole of Twitter is in your f*king cats.*
Non-toxic	І знову дві години на прокидання. *And again, two hours to wake up.* Ну, це тіпа добре, коли хвалять. *Well, it's kind of nice to be praised.* скоро буду своєю серед чужих))) аха *soon I will be my own among strangers))) aha*

Table 1: Toxic and non-toxic examples in Ukrainian.

1 Introduction

Lately, the NLP community has shifted away from exclusively developing monolingual English models and is placing greater emphasis on the development of fair multilingual NLP technologies. There were released plenty of multilingual models, i.e. mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020), mT5 (Xue et al., 2021), mBART (Tang et al., 2020), BLOOMz (Muennighoff et al., 2023), NLLB (Costa-jussà et al., 2022). Additionally, Large Language Models (LLMs) pre-trained on extensive corpora have expanded the realm of potential capabilities (Wei et al., 2022) not only for novel tasks but also for languages.

Nevertheless, the coverage of languages and classical NLP tasks corpora existence is still unequal. In the scope of harmful language detection, we discovered an absence of any toxicity or hateful detection corpora for the Ukrainian language. Thus, the question arises: what is the most effective and promising approach to acquiring a binary toxicity classification corpus for a new language, considering all the recent advancements in the field of NLP. Answering this main research question, the contribution of this work are the following:

- We present the first of its kind toxicity classification corpus for Ukrainian (Table 1) testing three approach for its acquisition: (i) translation from a resource rich language; (ii) toxic samples filtering by toxic keywords; (iii) crowdsourcing data annotation;
- Additionally, we explore three types of cross-lingual knowledge transfer approaches—Backtranslation, LLMs Prompting, and Adapter Training;
- We test both cross-lingual and supervised approaches on all test sets providing insights into the methods effectiveness.

All the obtained data and models are available for the public usage online.[1,2,3]

[1]https://huggingface.co/ukr-detect
[2]https://huggingface.co/textdetox
[3]https://huggingface.co/dardem/xlm-roberta-large-uk-toxicity

Proceedings of the The 8th Workshop on Online Abuse and Harms (WOAH), pages 244–255
June 20, 2024

Method	Models	Datasets	Translation Dependence	Data Creation	Fine tuning	# Inference Steps
			Cross-lingual Transfer Methods			
Backtranslation	- Toxicity detection model for the resource-rich language; - Translation model from resource-rich to the target language;	—	✔	✘	✘	3
LLM prompting	- LLM with the knowledge of the resource-rich language and (emerging) knowledge of the target language;	—	✘	✘	✘	1
Adapter Training	- Auto-regressive multilingual LM where the resource-rich and target languages are present; - Language adapter layers for both languages;	- Toxicity classification dataset in the resource-rich language; - Corpus for translation between the resource-rich and target languages;	✘	✘	✔	1
			Data Acquisition Methods			
Training Data Translation	- Translation model to the target language; - Auto-regressive multilingual or monolingual LM for the target language;	- Toxicity classification dataset in the resource-rich language;	✔	✔	✔	1
Semi-synthetic data by keywords filtering	- Embedding model of texts in the target language;	- Texts in the target language; - List of toxic keywords in the target language;	✘	✔	✔	1
Crowdsourcing data filtering	- Embedding model of texts in the target language;	- Texts in the target language;	✘	✔	✔	1

Table 2: Comparison of the considered approaches for cross-lingual detoxification transfer and corpora acquisition based on required computational and data resources.

2 Related Work

The usual case for cross-lingual transfer setup is when data for a specific task is available for English but none for the target language. In such a setup, translation of training data approach has been already explored for sentiment analysis (Kumar et al., 2023) and offensive texts classification (El-Alami et al., 2022; Wadud et al., 2023).

For toxicity, both monolingual and multilingual corpora have been introduced. Thus, English Jigsaw dataset (Jigsaw, 2017) was later extended to the multilingual format (Jigsaw, 2020). Within East European language, there were presented offensive language detection in Polish (Ptaszynski et al., 2024) and Serbian (Jokic et al., 2021) based on Twitter data. In the related domain, Ukrainian bullying detection system was developed based on translated English data in (Oliinyk and Matviichuk, 2023). However, none of the works yet covered specifically Ukrainian toxicity detection.

Definition of Toxicity While there can be different types of toxic language in conversations (Price et al., 2020; Gilda et al., 2021), i.e. sarcasm, hate speech, direct insults, in this work include samples with substrings that are commonly referred to as vulgar or profane language (Costa-jussà et al., 2022; Logacheva et al., 2022) while the whole main message can be both neutral and toxic. Thus, we are considering the task of binary toxicity classification assigning the labels either toxic or non-toxic.

3 Cross-lingual Knowledge Transfer Methods

Firstly, we test three cross-lingual knowledge transfer methods that do not require any training data in the target language acquisition (Table 2): (i) Backtranslation; (ii) LLM Prompting; (iii) Adapter Training. We assume a setup where resource-rich available language is English.

	Translated dataset	Semi-synthetic dataset	Crowdsourced dataset
Train	total: 24616 toxic: 12307 non-toxic: 12309	total: 12606 toxic: 6362 non-toxic: 6244	total: 3000 toxic: 1500 non-toxic: 1500
Val	total: 4000 toxic: 2000 non-toxic: 2000	total: 4202 toxic: 2071 non-toxic: 2131	total: 1000 toxic: 500 non-toxic: 500
Test	total: 52294 toxic: 5800 non-toxic: 46494	total: 4214 toxic: 2114 non-toxic: 2008	total: 1000 toxic: 500 non-toxic: 500

Table 3: Statistics of the obtained datasets: train/val/test splits.

Backtranslation For many tasks, an English classifier may already exist, making it a natural baseline to translate the input text from Ukrainian to English and then employ the English classifier for the task. This Backtranslation approach eliminates the need for fine-tuning but relies on external models—an translation system and an English classifier— for consistent functionality.

LLM Prompting The next approach that as well does not require fine-tuning is prompting of LLMs. Current advances in generative models showed the feasibility of transforming any NLP classification task into text generation task (Chung et al., 2022; Aly et al., 2023). Thus, the prompt can be designed in a zero-shot or a few-shot manner requesting the model to answer with the label. While LLMs were already tested for a hate speech classification task for multiple languages (Das et al., 2023), there were no yet experiments for any text classification task for Ukrainian language which might be underrepresented in such models. We provide the final design of our prompt in Appendix B.

Adapter Training Finally, the most parameter-efficient approach involves employing language-specific Adapter layers (Pfeiffer et al., 2020). Such a layer, firstly, for English, can be added upon multilingual LM. Everything remains frozen while fine-tuning of the final Adapter for the downstream task. Then, English Adapter is replaced with Ukrainian one and inference for the task in the target language can be performed.

4 Data Acquisition Methods

To obtain supervised detection models, we test three ways of training data acquisition for toxicity detection task (Table 2): (i) English toxicity corpus translation into Ukrainian; (ii) filtering toxic samples by pre-defined dictionary of Ukrainian toxic keywords; (iii) crowdsourcing annotation to filter Twitter corpus into toxic and non-toxic samples. The examples of samples from these three dataset can be found in Appendix C.

4.1 Training Corpus Translation

To avoid the permanent dependence on a translation system per each request, we can translate the whole English dataset and, as a result, get synthetic training data for the task. Then, a downstream task fine-tuning is possible. This approach's main advantage is that there are no external dependencies during the inference time, but it requires computational resources for fine-tuning. Moreover, some class information might vanish after translation and will not be adapted for the target language.

English Dataset To test this approach, we considered English datasets Jigsaw data (Jigsaw, 2017). We collapsed all labels except from "non-toxic" into one "toxic" class.

Translation Systems Choice To choose the most appropriate translation system, we took into consideration two opensource models—NLLB[4] (Costa-jussà et al., 2022) and Opus[5] (Tiedemann, 2012). We randomly selected 50 samples per each dataset and asked 3 annotators (native speakers in Ukrainian) to verify the quality. As a result, we choose Opus translation system for toxicity classification as it preserves better the toxic lexicon. The system achieved 90% of qualitative translations.

4.2 Semi-synthetic Dataset with Toxic Keywords Filtering

To obtain toxic samples for these approach, we filtered Ukrainian tweets corpus from (Bobrovnyk,

[4]https://huggingface.co/facebook/nllb-200-distilled-600M

[5]https://huggingface.co/Helsinki-NLP/opus-mt-en-uk

	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
	Translated Test Set			Semi-synthetic Test Set			Crowdsourced Test Set		
Prompting of LLMs									
LLaMa-2 Prompting	0.50	0.67	0.42	0.67	0.49	0.67	0.24	0.50	0.32
Mistral Prompting	**0.68**	**0.74**	**0.70**	**0.81**	**0.76**	**0.75**	**0.56**	**0.68**	**0.52**
Cross-lingual transfer approaches									
Backtranslation		—		**0.76**	**0.56**	**0.58**	**0.75**	**0.68**	**0.65**
Adapter Training	0.66	0.63	0.65	0.66	0.58	0.52	0.64	0.58	0.53
Fine-tuning of LMs on different types of data									
XLM-R-finetuned-translated	0.68	0.86	0.70	0.79	0.77	0.77	0.70	**0.68**	**0.67**
XLM-R-finetuned-semisynthetic	0.59	0.53	0.53	0.99	0.99	0.99	**0.75**	0.57	0.48
XLM-R-finetuned-crowdsourced	**0.61**	**0.63**	**0.62**	**0.93**	**0.93**	**0.93**	0.99	0.99	0.99

Table 4: Ukrainian Toxicity Classification results. Within methods comparison, **bold** numbers denote the best results within methods types, gray—in domain results of the fine-tuned models. We do not test Backtranslation approach on the translated data as we cannot guarantee this test set was not present in the English training data of the model.

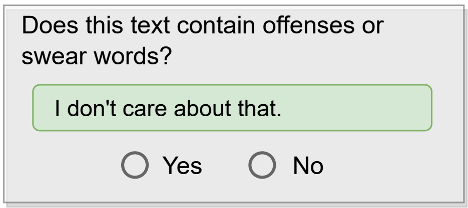

Figure 1: Interface (translated into English for illustration) of the toxicity classification task for data collection with crowdsourcing.

2019a) based on toxic keywords (Bobrovnyk, 2019b). We provide the full description of toxic keywords list construction in Appendix A. Then, tweets that did not contain any toxic words and additional texts from news and fiction UD Ukrainian IU dataset (Kotsyba et al., 2016) were considered as non-toxic.

4.3 Data Filtering with Crowdsourcing

To obtain toxic samples with crowdsourcing, we took Ukrainian tweets corpus (Bobrovnyk, 2019a), erased URL links, and Twitter nicknames, dropped phrases with less than five and more than twenty words, randomly sampled texts for the annotation with Toloka platform[6] (Figure 1). We hired only workers who passed the in-platform test of Ukrainian language knowledge. Each task page contained 9 real tasks, 2 control tasks with known answers, and 1 training task with known answers and explanations. We blocked participants if their answers were inadequately fast (less than 15 seconds per page), if they skipped 5 pages in a row, or if they failed on more than 60% of tasks with known answers. The crowdsourcing instructions and interface are listed in Appendix D.

[6]https://toloka.ai

5 Experimental Setup

The statistics of train/val/test splits are presented in Table 3. For the Ukrainian texts encoder, XLM-RoBERTa[7] (Conneau et al., 2020) has already been proven as a strong baseline for multiple languages (ImaniGooghari et al., 2023). For LLMs prompting, we experimented with couple setups choosing LLaMa-2[8] (Touvron et al., 2023) and Mistral[9] (Jiang et al., 2023) as the most promising models for the Ukrainian inputs processing. For English toxicity classifier, we used an open fine-tuned version of the DistilBERT model to classify toxic comments.[10]

6 Results

The classification results are presented in Table 4. Within methods that do not require fine-tuning, Backtranslation and Adapter Training look like promising baselines. Mistral outperforms LLaMa with top results on the semi-synthetic test set, but poorly on translated and, most importantly, crowdsourced data. At the same time, Backtranslation achieved top results on these two datasets that illustrates real Ukrainian toxic data the most.

When fine-tuned on the crowdsourced data, XLM-R exhibits almost perfect performance on

[7]https://huggingface.co/FacebookAI/xlm-roberta-large
[8]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[9]https://huggingface.co/mistralai/Mistral-7B-v0.1
[10]https://huggingface.co/martin-ha/toxic-comment-model

both the in- and out-of-domain test sets. Undoubtedly, data collected through human annotations embodies the most accurate understanding of toxicity. However, its performance significantly drops on translated data with the results even lower than unsupervised approaches. That can be due to the reduced toxicity in the translated data: not all labelled originally toxic data remained toxic in Ukrainian. Conversely, the model fine-tuned on the translated data demonstrates the best results on the annotated test set. Thus, the Training Data Translation approach still stands as a viable baseline, showcasing robustness across out-of-domain data.

7 Conclusion

We presented the first of its kind study in toxicity detection in the Ukrainian language. Firstly, we tested several cross-lingual knowledge transfer approaches for the task that have different resources requirements: Backtranslation that requires three inferences steps, LLMs prompting, and Adapter training that requires only adapter layer fine-tuning. Still, the Backtranslation approach showed the best performance within unsupervised baselines.

Next, we explored three methods for acquiring a binary toxicity classification corpus: translating an existing labeled English dataset, filtering toxic samples using a predefined list of Ukrainian toxic keywords, and collecting data through crowdsourcing. The model fine-tuned on translated data exhibited the most resilient performance across out-of-domain datasets, serving as a robust baseline. Ultimately, the model fine-tuned on manually annotated data demonstrated the highest performance.

Limitations & Ethics Statement

In this work, we encounter toxic speech as only speech with obscene lexicon and commonly referred to as vulgar or profane language (Costa-jussà et al., 2022). Thus, this work does not cover any other sides and shades of offensive language like hate, sarcasm, racism, sexism, etc. We believe that this study in toxic language detection will build a new foundation of any harmful language detection in Ukrainian.

Another limitation of this work that we consider only resource-rich language as English. For translated corpus acquisition it might also be beneficial to explore other languages from the linguistic families that are closer to Ukrainian, i.e. Polish or Croatian, if the corpora for the desired task exist in the corresponding languages.

In conclusion, the proposed toxicity detection model is openly shared with the community for further exploration. Deploying this model for specific use cases and domains should be complemented by human-computer interaction solutions that uphold users' freedom of speech while fostering proactive conversations. We firmly believe that our proposed toxicity classification data and models will contribute to the development of more fair and safe multilingual LLMs.

References

Rami Aly, Xingjian Shi, Kaixiang Lin, Aston Zhang, and Andrew Gordon Wilson. 2023. Automated few-shot classification with instruction-finetuned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2414–2432. Association for Computational Linguistics.

Kateryna Bobrovnyk. 2019a. Automated building and analysis of ukrainian twitter corpus for toxic text detection. In *COLINS 2019. Volume II: Workshop*.

Kateryna Bobrovnyk. 2019b. Ukrainian obscene lexicon. https://github.com/saganoren/obscene-ukr. Accessed: 2023-12-14.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti

Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Mithun Das, Saurabh Kumar Pandey, and Animesh Mukherjee. 2023. Evaluating chatgpt's performance for multilingual and emoji-based hate speech detection. *CoRR*, abs/2305.13276.

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

Fatima-Zahra El-Alami, Said Ouatik El Alaoui, and Noureddine En-Nahnahi. 2022. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. *J. King Saud Univ. Comput. Inf. Sci.*, 34(8 Part B):6048–6056.

Shlok Gilda, Luiz Giovanini, Mirela Silva, and Daniela Oliveira. 2021. Predicting different types of subtle toxicity in unhealthy online conversations. In *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*, pages 360–366. Elsevier.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André FT Martins, François Yvon, et al. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. *arXiv preprint arXiv:2305.12182*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

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

Jigsaw. 2020. Multilingual toxic comment classification. https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification. Accessed: 2024-03-18.

Danka Jokic, Ranka Stankovic, Cvetana Krstev, and Branislava Sandrih. 2021. A twitter corpus and lexicon for abusive speech detection in serbian. In *3rd Conference on Language, Data and Knowledge, LDK 2021, September 1-3, 2021, Zaragoza, Spain*, volume 93 of *OASIcs*, pages 13:1–13:17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Natalia Kotsyba, Bohdan Moskalevskyi, Mykhailo Romanenko, Halyna Samoridna, Ivanka Kosovska, Olha Lytvyn, and Oksana Orlenko. 2016. Ud ukrainian iu. `https://universaldependencies.org/treebanks/uk_iu/index.html`.

Puneet Kumar, Kshitij Pathania, and Balasubramanian Raman. 2023. Zero-shot learning based cross-lingual sentiment analysis for sanskrit text with insufficient labeled data. *Appl. Intell.*, 53(9):10096–10113.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. Association for Computational Linguistics.

V Oliinyk and Matviichuk. 2023. Low-resource text classification using cross-lingual models for bullying detection in the ukrainian language. Адаптивні системи автоматичного управління: міжвідомчий науково-технічний збірник, 2023,№ 1 (42).

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Ilan Price, Jordan Gifford-Moore, Jory Flemming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversations. In *Proceedings of the Fourth Workshop on Online Abuse and Harms, WOAH 2020, Online, November 20, 2020*, pages 114–124. Association for Computational Linguistics.

Michal Ptaszynski, Agata Pieciukiewicz, Pawel Dybala, Pawel Skrzek, Kamil Soliwoda, Marcin Fortuna, Gniewosz Leliwa, and Michal Wroczynski. 2024. Expert-annotated dataset to study cyberbullying in polish language. *Data*, 9(1):1.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Md. Anwar Hussen Wadud, Muhammad F. Mridha, Jungpil Shin, Kamruddin Nur, and Aloke Kumar Saha. 2023. Deep-bert: Transfer learning for classifying multilingual offensive texts on social media. *Comput. Syst. Sci. Eng.*, 44(2):1775–1791.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *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*, pages 483–498. Association for Computational Linguistics.

A The Full List of Toxic Keywords Used for Filtering

This list only serves to increase reproducibility of our work and has no intention to offend the reader.

Additionally to the openly available list of Ukrainian toxic keywords[11], we also came up with some additional words that can be divided into the following groups:

Slurs towards a group of people under discrimination (nationality, race, sexual orientation etc.):

"хохол", "хохли", "хохлом", "хохлами", "жид", "жиди", "жидом", "жидами", "жидовка", "жидовський", "жидовські", "жидовська", "вузькоглазий", "вузькоглазі", "ніга", "нігга", "ніггерам", "ніггери", "ніггерів", "нігер", "нігера", "нігерами", "нігери", "нігерка", "нігерська", "нігерський", "нігерських", "нігерські", "нігерів", "нігріла", "нігєр", "нігєра", "нігєри", "нігєрський", "педарастів", "педераст", "педерастія", "педик", "педики", "педиків", "педік", "педіковського", "підар", "підари", "підором", "підараси", "підорський", "підорас", "підарас", "підарам", "підару", "підарасу", "підарасам", "тьолка", "тьолкою", "тьолки", "тьолками", "тьолкам", "тьолці", "блядь", "бляді", "шалава", "шалави", "прошмандовка"

Most often used toxic or hate appeals to the opponent:

"уйобок", "хуйло", "ахуєл", "уєбан", "уїбан", "довбойоб", "долбойоб", "залупа", "гандон", "пизда", "їблан", "єблан", "їбанутий", "єбанутий"

Different obscene words (forms without endings):

"їбат", "їбан", "пизд", "бля"

B LLM Toxicity Classification Prompt

Denote: even if we perform classification for texts in Ukrainian, the core structure of the note is still in Enlglish. Such a design was proven to be the most successful in our experiments.

Classify the text into two categories: contains obscene words or contains none obscene words. Reply with only one word: obscene or normal.

Examples:
Text: Сьогодні знайти у відкритих джерелах точну суму, витрачену на будівництво об'єкта, що про нього мова, майже неможливо. *Sentiment: normal.*
Text: знаєте, якщо свої дебільні коментарі ще й ілюструвати посиланнями на російську вікі, хтось може здогадатися, що ви тупий єблан. *Sentiment: obscene.*

Text: {text}
Sentiment:

[11] https://github.com/saganoren/obscene-ukr

C Corpora Data Examples

C.1 Translated Data

Here, we provide examples in Table 5 of data gained from the Data Translation approach. The translation are the ones obtained from the utilized translation model (Tiedemann, 2012).

Toxic	ви всі тупі осли. *youre all dumb asses*
Non-toxic	Є два адміністратори, які досить добре працюють з такими статтями, можливо, ви могли б зв'язатися з ними. *there are two admins that do handle such articles pretty well you could maybe contact and*
Toxic	І Роберт - це чорне лайно *and robert is a black shit*
Non-toxic	Гаразд, я почав трансляцію нової статті, я використав вашу запропоновану назву може використовувати більше деталей зараз *ok i started the transtion the new article i used your suggested title could use a lot more detail now*
Toxic	Що за купа ср*них ботанів? *what a bunch of f**king nerds*
Non-toxic	Зупиніться, будь ласка, якщо ви продовжите вандализувати сторінки, ви будете заблоковані від редагування wikipedia *please stop if you continue to vandalize wikipedia you will be blocked from editing*
Toxic	Альтернативна поп-культура, що означає п*зда чи ци, розкидає таємничу сучку, яка руйнує все, що примара називає когось, це спосіб дати людині знати, що вони є п*зда в той час як цензують інших навколо вас в громадських місцях або в соціальних кутах, сучасний сленг попереджаючи інших про небезпеку. *alternative pop culture meaning c*nt or cee unt a percieved mysterious bitch that destroys everything whem calling someone this is a way of letting anyone know they are a c*nt while censoring others around you in public or in social corners a modern slang alerting other of the danger*
Non-toxic	Адміністратори виконують дії, що ґрунтуються на громадському консенсусі, вони не приймають односторонніх рішень далі, тому у зв'язку з цим редактори, які зосереджують свою увагу на виборах або канадалях, не мають можливості перенаправити кандидатів на партійні статті. *admins execute actions based on community consensus they do not make unilateral decisions further that afd did not have the involvement of editors who focus on ontario or canadawide elections so they were likely unfamiliar with the option of redirecting to party candidate articles*

Table 5: Examples of translated samples for **Toxicity Classification** task. English translation are taken from the Jigsaw dataset (Jigsaw, 2017).

C.2 Semi-synthetic Data

Here, we provide examples in Table 6 of data gained by filtering with toxic keywords.

Toxic	@USER нє, китай рулить, то однозначно. ден сяопін був генієм економіки. але це було підписано бо більше ні на шо пі**рович не заслужив:) *@USER no, the Chinese drive, of course. The shoopin was an economic genius. But it was signed because no more on the sublarcier was worthy of:)*
Non-toxic	@USER 1) доброго часу, коліжанці дав почитати збірку, багато що оцінила, але запитала про "Самонедостатність" .. *@USER 1) good time, the colts gave the collection a lot of reading and appreciated, but asked about "Memo-nysity..."*
Toxic	вже не пі**рас? *Isn't that a f**got?*
Non-toxic	Не раз заявляв про наміри зайти на наш ринок ірландський Ryanair . *More than once, he claimed to visit our market in Irish Ryanair.*
Toxic	сьогоднішня мрія - адекватний транспорт в крим, щоб не доводилося щоразу мозок собі ї**ти стиковкою цих жахливих людиноненависницьких рейсів *Today's dream is a safe transport into the ice so that every brain doesn't have to f**k its way through these terrible man - hated flights.*
Non-toxic	Співрозмовники досягли домовленості про проведення чергового засідання Спільної міжурядової українсько - туркменської комісії з економічного та культурно - гуманітарного співробітництва вже ближчим часом . *Coordinators have reached an agreement to hold a joint Intergovernmental Union Commission on Economic and Cultural Cooperation for a longer time.*
Toxic	нема відчуття гіршого, ніж коли розумієш, шо ти конкретно так тупанув, і через це все йде по п**ді. *There's no worse feeling than when you realize that you were exactly f**king that way, and that's why everything goes on p*ss.*
Non-toxic	Державне підприємство « Конструкторське бюро „ Південне " ім . М . К . Янгеля » було створено 1951 як конструкторський відділ Південного машинобудівного заводу з виробництва військових ракет . *The state enterprise (C) was created by 1951 as the South Carworker's design department for the production of military rockets.*

Table 6: Examples of natural samples for **Toxicity Classification** task obtained from Ukrainian tweets corpus from (Bobrovnyk, 2019a) and news and fiction UD Ukrainian IU dataset (Kotsyba et al., 2016) by the toxic keywords filtering.

C.3 Crowdsourced Data

Here, we provide examples in Table 7 of data gained by crowdsource collection.

Toxic	Фу ти що той як його п*дар? *What are you, like his f*got?*
Non-toxic	Якщо не думати про певні проблеми, то жити дуже навіть можна. *If you don't think about certain problems, you can live very well.*
Toxic	П*здєц, я такий безкоштовно могла будь-де знайти *F*ck, I could have found one for free anywhere)*
Non-toxic	Бо я щось зовсім від новин відвик... *Because I'm not used to the news at all...*
Toxic	Завод по проізводству сосатєльних х*ов. *A factory for the production of sucking d*ks.*
Non-toxic	Нарізав вам фрагменти вчорашнього ефіру з Мураєвим. *I've cut you fragments of yesterday's broadcast with Muraev.*
Toxic	Тому от вони, а не х*рь якась *So here they are, not some bul*hit*
Non-toxic	Особливо, коли в тебе другий день шалена слабкість. *Especially when you've been feeling crazy weak for two days.*
Toxic	Давайте, розкажіть нам що це просте співпадіння, оце х*та з Мо*нкою. *Go ahead, tell us that it's a simple coincidence, this f*k with the sc*tum.*
Non-toxic	Не люблю свята, бо це лише витрати та клопіт, а так жодної різниці зі звичайним рутинним днем. *I don't like holidays because they're just expenses and hassle, and there's no difference between them and a normal day.*
Toxic	Ну для мене люба френдзона це ху*ве місце *Well, for me, any friendzone is a fu*ing place*
Non-toxic	Є цікаві персонажі й діалоги, сюжет середній. *There are interesting characters and dialogues, but the plot is average.*

Table 7: Examples of crowdsourced samples for **Toxicity Classification** task obtained from Ukrainian tweets corpus from (Bobrovnyk, 2019a).

D Crowdsourcing platform instructions and interface

Here, we list the full instruction and task interface in the original Ukrainian language. Per each page, the annotators were paid 0.10$.

D.1 General instructions for the task.

Вам потрібно прочитати речення і визначити, чи містять вони образи або нецензурні та грубі слова.

You need to read the sentences and determine if they contain insults or obscene and rude words.

Увага! Необразне речення може містити критику і бути негативно забарвленим.

WARNING! A non-figurative sentence can contain criticism and be negatively colored.

Приклади

Examples

Образливі речення:

Offensive sentences:

- Інтернет-шпана, не тобі мене повчати.
  Internet-nasty crew, it's not for you to teach me.
- Яка підписка, що ти несеш, поїхавший?
  What is the subscription, what are you talking about, are you mad?
- Щонайменше два малолітніх дегенерати в треді, мда.
  At least two juvenile degenerates in a thread, huh?
- Взагалі не бачу сенсу сперечатися з приводу дюймів, хуєвий там ips чи ні, машина не цим цікава.
  In general, I don't see any point in arguing about inches, whether the ips is fucked up or not, this is not what makes the car interesting.

Нейтральні (не образливі) речення:

Neutral (not offensive) sentences:

- У нас є убунти і текнікал прев'ю.
  We have Ubuntu and Teknical previews.
- він теж був хоробрим!
  He was brave too!
- Це безглуздо, ти ж знаєш
  It makes no sense, you know that.
- Якщо він мріє напакостити своїм сусідам, то це погано.
  If he dreams of hurting his neighbors, that's bad.

D.2 Task interface

Чи містить цей текст образи або нецензурні слова?

Does the text contain insults or obscenities?

- Так
  Yes
- Ні
  No