Linguistic_Hygenist at PAN 2024 TextDetox: HybridDetox - A Combination of Supervised and Unsupervised Methods for Effective Multilingual Text Detoxification

Notebook for PAN at CLEF 2024

Susmita Gangopadhyay, M.Taimoor Khan and Hajira Jabeen

GESIS Leibniz Institute for the Social Sciences, Köln, Germany

Abstract

Text detoxification is the process of revising toxic comments to neutralize their toxicity by eliminating inappropriate content, while preserving the meaning of the message. Toxicity can manifest in various forms, including the use of curse words, insults, hate speech, cyberbullying, or trolling. The present-day social media landscape is rife with toxic comments, necessitating a text detoxification system. Unlike the conventional practice of blocking offensive content through moderation, detoxification preserves the valuable information contained in the message. This paper details our approach for multilingual text detoxification as part of the *Multilingual Text Detoxification (TextDetox) 2024* [1] Challenge organized by the PAN lab [2]. Our approach consists of two components i.e., the Supervised T5-BART Module for English and Russian languages with parallel corpora and the Unsupervised PLM Detoxifier for the other seven languages. The Supervised T5-BART Module uses T5 and BART as base models, with exponentially weighted moving average and ROUGE scores as loss functions for Russian and English, respectively. The Unsupervised PLM Detoxifier utilizes hashing techniques, log odds ratio, and linguistic patterns to identify and conceal toxic words across all languages. Additionally, it incorporates a mask prediction model to maintain the original sentence's meaning intact. Our proposed approach has achieved an average score of 0.315 across all languages, exhibiting outstanding performance in English, German, and Ukrainian for style transfer, content preservation, and fluency.

Keywords

PAN 2024, text-detoxification, toxicity, mask-prediction, sentence-similarity, sequence-to-sequence models, CEUR-WS

1. Introduction

Internet access has revolutionized information dissemination, providing unprecedented opportunities worldwide. However, this rapid and uncontrolled proliferation of information containing user-generated content could also contain toxic information that is considered harmful, offensive, or inappropriate. Text detoxification is a critical endeavor in the contemporary digital landscape, where the proliferation of toxic comments poses significant challenges to online discourse [3]. Detoxification process involves the meticulous revision of toxic comments to neutralize their toxicity while ensuring that the essence of the original message remains intact [4]. Toxicity can manifest in numerous forms, including the use of curse words, insults, hate speech, cyberbullying, or trolling, contributing to an unhealthy online environment [5]. This pervasive toxicity underscores the urgent need for effective text detoxification systems to maintain a healthier online ecosystem [6]. Identification of toxicity in text is an active area of research. Today, social networks such as *Facebook*, *Instagram* are trying to address the problem of toxicity. However, they usually simply block offensive content through moderation [7]. Text detoxification prioritizes the preservation of valuable information within the message while neutralizing its toxicity.

Significant progress has been made in detecting offensive or toxic speech. The supervised text

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

susmita.gangopadhyay@gesis.org (S. Gangopadhyay); taimoor.khan@gesis.org (M.Taimoor Khan); hajira.jabeen@gesis.org (H. Jabeen)

0009-0009-1520-9070 (S. Gangopadhyay); 0000-0002-6542-9217 (M.Taimoor Khan); 0000-0003-1476-2121 (H. Jabeen)

detoxification techniques are used for languages with abundant resources having parallel corpora [8]. On the other hand, the unsupervised techniques target languages with limited resources by employing alternative methods [9]. In the realm of multilingual text detoxification, existing approaches typically adopt a combination of supervised and unsupervised techniques [10]. The integration of pretrained models has been instrumental in advancing text detoxification efforts [11]. These models, trained on vast amounts of data, possess a remarkable ability to understand and generate human-like text across various languages. However, despite their potential, several open challenges like the inability to generalize to different contexts, inefficiency to handle implicit and subtle toxicity, and below-par performance in multilingual setup remains [12]. Adapting pretrained models to diverse languages and cultural contexts while ensuring their effectiveness in detecting and neutralizing toxic content presents a significant hurdle [13] due to the continuously evolving language or presence of sarcasm, innuendo, or coded language in the text [14].

Our proposed approach aims to tackle multilingual detoxification by adopting a hybrid method of supervised named Supervised T5-BART Module and unsupervised modules named Unsupevised PLM Detoxifier. The Supervised T5-BART Module fine-tunes T5 Seq2Seq model for Russian, which was originally trained in a teacher-forcing style for multiple NLP tasks like summarization, translation, and text generation. It uses an exponential weighted moving average (EWMA) score for loss evaluation. Additionally, it fine-tunes BART model for English using ROUGE scores for loss evaluation. Meanwhile, the Unsupervised PLM Detoxifier adopts a multi-step process, including the masking of multiple toxic tokens and predicting a suitable mask replacement while still preserving the meaning. By leveraging both supervised and unsupervised methods, our approach offers a robust and versatile solution to the complex problem of multilingual text detoxification.

In the subsequent sections, we describe the problem statement, related previous research, our proposed approach, and present some examples from our results. In addition, we also share our vision of future work that could be adopted in this research direction.

2. Problem Definition

The competition expects a text detoxification system for 9 languages from different linguistic families. Parallel training corpora of several thousand toxic-detoxified pairs are available only for English and Russian languages. For the remaining 7 languages—Spanish, German, Chinese, Arabic, Hindi, Ukrainian, and Amharic—only texts containing toxic content were provided. For all 9 languages, a list of prominent toxic lexicons was provided, varying in number. For languages like English and Russian where parallel training corpora was available, fine-tuning of any text-generation model was allowed. The main challenge of this competition was to use a mix of supervised and unsupervised approaches to develop a multilingual text detoxification system. The evaluation was based on both automatic methods such as duplication, deletion, and backtranslation as mentioned on the challenge website[1] as well as manual verification of the detoxified text.

3. Related Work

There have been numerous studies and shared tasks focusing on toxicity detection, particularly for English language. One of the earliest and most notable efforts came from several Kaggle competitions organized by the Jigsaw/Conversation AI team, which included the "Toxic Comment Classification Challenge"[2] in 2018, the "Unintended Bias in Toxicity Classification Challenge"[3] in 2019, and the "Multilingual Toxic Comment Classification Challenge"[4] in 2021. These competitions provided some of the largest datasets for English toxicity detection, covering multiple types of toxicity such as toxic,

[1] https://pan.webis.de/clef24/pan24-web/text-detoxification.html
[2] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
[3] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification
[4] https://www.kaggle.com/code/bond005/siamese-xlm-r-for-multilingual-sentiment-analysis

obscene, threat, insult, and identity hate, along with multilingual test sets for other languages like Spanish, French, Italian, and Russian.

Starting in 2019, the detection of toxicity and offensive language has been a major focus at SemEval. This began with the SemEval-2019 Task 6 and continued with the SemEval-2020 Task 12, both centered around identifying and categorizing offensive language in Social Media (OffensEval), which garnered considerable interest and participation. The emphasis on toxicity persisted with the SemEval-2021 Task on Toxic Spans Detection. This task was designed to pinpoint the exact spans within a text that make it toxic, providing valuable assistance to human moderators who have to manage lengthy and potentially harmful comments.

In 2022, the arena of toxicity detection continued to buzz with activity, featuring events such as the "Multimedia Automatic Misogyny Identification (MAMI)" competition. This unique challenge focused on identifying misogynous memes, utilizing a comprehensive analysis of both textual content and accompanying images. By shedding light on the pervasive issue of systemic gender inequality and discrimination against women in online spaces. This competition played a pivotal role in raising awareness and fostering discussions on these critical issues.

Apart from these shared tasks and competitions, several other research works focus on the task of toxicity identification and text detoxification. Focus has been on utilizing Deep Learning models like LSTM [15], utilization of embedding models [16], and incorporation of context [17] in the detection of toxic texts. Detoxification is generally framed as a style transfer from toxic to neutral (non-toxic) style, using parallel datasets labeled for toxicity. For example, Logacheva [8] created such a parallel corpus for the English language as a resource for utilization in the detoxification task. Researchers such as Laugier [18] used pretrained text-to-text transformer trained on civil comments dataset to create fluent and neutral sentences from toxic ones. Detoxification efforts often rely on style transfer models tested in other domains. For example, fine-tuning autoencoders with additional style classification and cycle-consistency losses [18] and applying point-wise corrections and seq2seq models to improve text fluency and style [9].

In terms of multilingual and low-resource languages, significant research has been conducted in multilingual text generation [19], language agnostic sentence embeddings [20], and translation of low-resource languages [21]. However, multilingual text detoxification remains a challenge that is relatively under-explored and is still active. A recent challenge, RUSSE-2022 [22], focused solely on detoxifying Russian texts. Our approach contributes to this unique and evolving area of research by proposing a unified pipeline for detoxification across multiple languages, including low-resource ones.

4. HybridDetox Pipeline

We propose a method that allows us to effectively address the challenge of detoxification across all languages in the dataset. Our proposed methodology is a hybrid of supervised and unsupervised approaches. Our pipeline takes toxic text as input, processes it, and rephrases it into detoxified text. *Figure 1* demonstrates the entire detoxification pipeline for the languages in study.

4.1. Language Detection Module

The first step a toxic text passes through is a language detection module. Although sentences and their corresponding languages were provided in the test data, the language detection module was added to simulate a real production scenario. We used the Python langdetect[5] library for this purpose. If the detected language is English or Russian the text is forwarded through the Supervised T5-BART Module. For the remaining seven languages the text is passed through an Unsupervised PLM Detoxifier.

[5] https://pypi.org/project/langdetect/

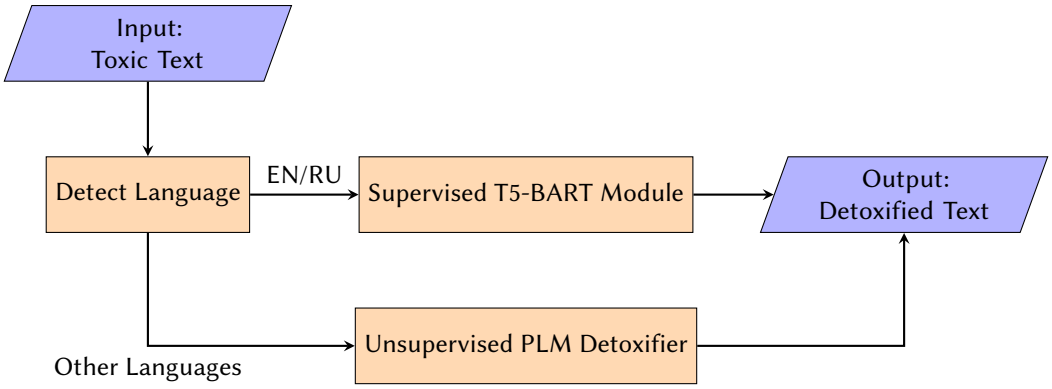

Figure 1: Detoxification Pipeline for all languages

4.2. Supervised T5-BART Module

Supervised T5-BART Module fine-tunes classifiers for English and Russian having parallel corpora. We used T5 (Text-to-Text Transfer Transformer) model as our base model for the Russian language [23]. T5 is designed around the innovative concept of having a single architecture across diverse tasks to benefit from transfer learning. It is trained on large-scale diverse datasets that enhanced its ability to understand and generate close to human-like text. T5 has demonstrated significant advancement for multiple NLP text generation-related tasks e.g., translation, text summarization, etc. We fine-tuned T5 on parallel corpora for Russian using exponentially weighted moving average (EWMA) loss function. It puts more emphasis on the recently generated text and is not effected by extreme values.

We employed BART for the English parallel corpora and fine-tuned for text detoxification [24]. BART model is effective for different text generation and comprehension tasks. Its architecture is flexible that facilitates fine-tuning for specific tasks with parallel corpora. BART has encoder-decoder architecture where the encoder is similar to BERT while the decoder is a GPT model. During fine-tuning, the models are exposed to a labeled dataset containing pairs of toxic and detoxified texts. ROUGE measures is used as loss measure to evaluate the quality of the generated text. It compares n-grams between generated texts and label text with higher overlap desired. We computed ROUGE-1, ROUGE-2 and ROUGE-L to train multiple models where the same measure were used to pick the best model.

4.3. Unsupervised PLM Detoxifier

Unsupervised PLM Detoxifier supports languages without parallel corpora. It is trained for 7 languages including low-resource languages i.e., Chinese, Amharic, German, Hindi, Arabic, Ukrainian, and Spanish. The working of Unsupervised PLM Detoxifier is explained with the following submodules.

4.3.1. Toxic Words Identification and Masking

To identify toxic words in the sentences, we adopted a combination of hashing-based techniques and log-odds ratio. As a starting point, we utilized the list of toxic lexicons provided in the challenge[6]. Each language has a list of toxic lexicons containing prominent curse words specific to that language. We employed a hashing-based sequence-matching mechanism[7] to identify words similar to these lexicons beyond a certain threshold. These identified toxic words were then removed from the sentences and replaced with masks. Suitable threshold values t_1 are identified based on manual evaluation.

In the next step, our approach relied on the principle that curse words are relatively rare and would appear less frequently in a neutral corpus compared to a toxic one. Therefore, the log-odds ratio between any normal neutral corpus and a toxic corpus would highlight a list of toxic words. The log-odds ratio

[6]https://huggingface.co/datasets/textdetox/multilingual_toxic_lexicon
[7]https://docs.python.org/3/library/difflib.html

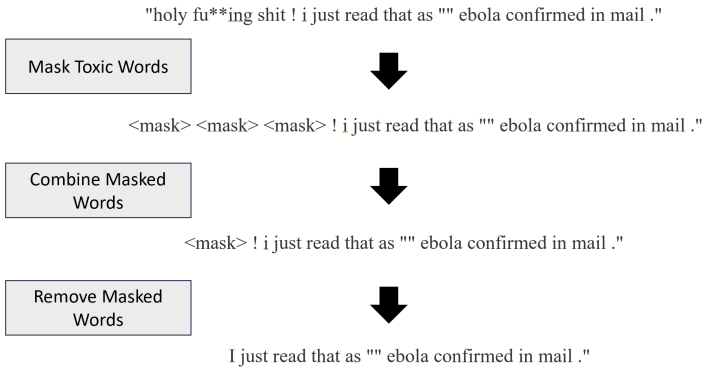

Figure 2: Example of implemented Mask Placement with Linguistic Pattern in our method

defines the relative frequency comparison and measures how often or less frequent a word is in one corpus compared to another. A higher log-odds ratio indicates that the word is much more common in the target corpus (e.g., toxic text) than in the reference corpus (e.g., neutral text). Mathematically, the log-odds ratio for a word w can be defined as:

$$\text{Log-Odds Ratio} = \log\left(\frac{\frac{p(w|C_{\text{toxic}})}{1-p(w|C_{\text{toxic}})}}{\frac{p(w|C_{\text{neutral}})}{1-p(w|C_{\text{neutral}})}}\right) \tag{1}$$

where $p(w \mid C_{toxic})$ and $p(w \mid C_{neutral})$ are the occurrence probabilities of the word w in toxic and neutral corpora, respectively.

In summary, the log-odds ratio helps identify words that are significantly more likely to appear in toxic texts compared to neutral ones, thus aiding in the detection of toxic language. We utilized the development set's toxic and neutral pairs for this experiment, but it could also have been conducted with any toxic and neutral corpus in the target languages. From the extracted list of words and their log-odds ratios, we selected those with a score above t_2 as toxic words. This threshold was chosen because the log-odds ratio values are in range [0, 1]. We aimed to maintain a balanced value to ensure that we accurately identify toxic words while minimizing false positives and negatives. Additionally, we cleansed our generated toxic lexicon list by filtering out stopwords and words that are less than 3 characters long. This was done to eliminate special characters, symbols, or incomplete random words. The filtering criteria for removing unwanted stopwords was based on general observation. This approach ensured that we effectively masked words likely to be curse words, thereby excluding stopwords and special characters that might have got on to the list of toxic words.

4.3.2. Mask Placement with Linguistic Patterns

Languages follow certain grammatical paradigms or linguistic rules that aid in constructing sentences. By observing these rules, we were able to better process the masks in sentences. We found that for any language if curse words appear at the beginning or end of a sentence, they can be simply removed. Additionally, when multiple consecutive masked words were present, they could be combined into a single mask without losing the overall meaning of the sentence. *Figure 2* shows an example of our implemented linguistic paradigms. For ease of understanding, the provided example is in English.

4.3.3. Mask Prediction

Following the process of identifying and masking toxic words, and implementing linguistic rules, we were left with sentences containing masked toxic words. To handle these, we used the XLM-RoBERTa large model [25], which is pretrained in a self-supervised manner on 2.5TB of filtered CommonCrawl data spanning 100 languages, including all languages featured in the competition. The model employs a

Masked Language Modeling (MLM) objective. It randomly masks 15% of the words in the input sentence, processes the entire masked sentence through the model, and predicts the masked words. We chose this model because, unlike traditional recurrent neural networks (RNNs) that process words sequentially or autoregressive models like GPT that internally mask future tokens, XLM-RoBERTa learns a bidirectional representation of the sentence. Using this model, we predicted the top three probable replacements for each mask and generated sentences accordingly. For sentences with multiple masks, this resulted in 3^n possible sentences.

4.3.4. Sentence Similarity

From our resultant 3^n sentences generated from the masked predictions, we used a sentence transformer model [26] to generate embeddings for each of the sentences along with their parallel toxic input sentence. The model works in a way that sentences with similar meanings are associated with embeddings that are close in the vector space. Then, the semantic textual similarity between two sentences is computed, and we have sentence pairs annotated together with a score indicating the similarity between them. The model uses a Siamese network structure that was trained using CosineSimilarityLoss [27]. Among all the sentence pairs generated, we chose the one that had the lowest score indicating the resultant sentence closest to the input toxic sentence as our selected output sentence. The code for both Supervised[8] and Unsupervised[9] pipelines are made available at GitHub.

5. Results

The experimental setup consists of setting up the threshold values. For masking toxic words in the unsupervised module, the threshold t_1 is set to 0.8. While the threshold t_2 for identifying toxic words based on their log-odds ratio is set to 0.5. The threshold values are determined based on manual evaluation. In supervised learning, English has 19744 and Russian has 11090 training samples. Our method achieves an average score of 0.315 on the leaderboard's automatic evaluation securing 22nd position and comparable average results with the mT5 baseline. Notably, we observe exceptional performance in languages such as English, German, Ukrainian, and Arabic, with scores of 0.47,0.41,0.42, and 0.52 respectively. In the manual evaluation conducted via crowdsourcing on a random subsample of 100 texts per language, our method secured the 18th place with an average score of 0.50.

Upon observing the results of both automatic and manual evaluation, we found that our proposed approach demonstrated suboptimal performance in Chinese and Spanish and was particularly ineffective for Russian. Despite the loss function indicating convergence and the text being detoxified, manual evaluation for Russian revealed that the generated text lacked meaningfulness. Further exploration of our method revealed that the fine-tuned T5 multi-task text generator model that was used for our method generated smaller tokens resulting in the generation of out-of-vocabulary words. This raises a significant concern that although language models trained on multilingual text generation may exhibit promising scores and reduced loss metrics, verifying their effectiveness for languages outside the researcher's linguistic proficiency remains challenging. This issue underscores the necessity of incorporating native speakers in the evaluation process to ensure the semantic integrity of the detoxified text.

We also found that using BART with ROUGE for English performed much better on the test set than using T5 with EWMA for the Russian language. In general, these findings indicate that unsupervised approaches to multilingual text detoxification using pretrained language models hold promising results despite the lack of parallel training corpora. *Table 1* shows training and validation loss of Supervised T5-BART Module. Samples of toxic sentences and their detoxified sentence for all 9 languages involved in the study are given in *Figure 3*.

[8] https://github.com/taimoorkhan-nlp/RuEn-supervised-detoxifier
[9] https://github.com/susmita3107/mDetoxifier-Multilingual-unsupervised-text-detoxifier

Table 1
Training and validation loss of the supervised models (BART and T5) for English and Russian, respectively.

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

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

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

Table 2
Manual and automatic evaluation scores of our proposed approach for individual languages and average of all languages. The evaluation is based on removing toxicity, style transfer accuracy, content preservation and fluency.

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

6. Conclusion and Future Work

In this work, we propose a novel approach that combines both supervised and unsupervised methods for text detoxification across nine languages, including some low-resource ones. Our work forms a part of the CLEF Multilingual TextDetox challenge 2024, achieving an average score of 0.315 on the leaderboard's automatic evaluation and 0.50 in manual evaluation. While our results are promising, we acknowledge that there is room for improvement. In the future, we aim to explore diverse methodologies, such as leveraging multilingual embedding features to identify linguistic similarities among different languages. Additionally, we intend to experiment with various clustering techniques to investigate potential hierarchical relationships among toxic words. Furthermore, we plan to explore domain adaptation and transfer learning methods, particularly for languages that share similar roots e.g., Italian,

Spanish, Portuguese and Latin. We anticipate that models trained on languages with similar linguistic roots might effectively perform on others with comparable linguistic characteristics.

References

[1] D. Dementieva, D. Moskovskiy, N. Babakov, A. A. Ayele, N. Rizwan, F. Schneider, X. Wang, S. M. Yimam, D. Ustalov, E. Stakovskii, A. Smirnova, A. Elnagar, A. Mukherjee, A. Panchenko, Overview of the multilingual text detoxification task at pan 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.

[2] 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.

[3] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the international AAAI conference on web and social media, volume 11, 2017, pp. 512–515.

[4] S. Poria, E. Cambria, D. Hazarika, P. Vij, A deeper look into sarcastic tweets using deep convolutional neural networks, arXiv preprint arXiv:1610.08815 (2016).

[5] Z. Zhang, D. Robinson, J. Tepper, Detecting hate speech on twitter using a convolution-gru based deep neural network, in: The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15, Springer, 2018, pp. 745–760.

[6] L. Zhou, A. Caines, I. Pete, A. Hutchings, Automated hate speech detection and span extraction in underground hacking and extremist forums, Natural Language Engineering 29 (2023) 1247–1274.

[7] P. Liu, J. Guberman, L. Hemphill, A. Culotta, Forecasting the presence and intensity of hostility on instagram using linguistic and social features, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 12, 2018.

[8] 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), 2022, pp. 6804–6818.

[9] D. Dale, A. Voronov, D. Dementieva, V. Logacheva, O. Kozlova, N. Semenov, A. Panchenko, Text detoxification using large pre-trained neural models, arXiv preprint arXiv:2109.08914 (2021).

[10] G. Hassan, J. Rabah, P. Madriaza, S. Brouillette-Alarie, E. Borokhovski, D. Pickup, W. Varela, M. Girard, L. Durocher-Corfa, E. Danis, Protocol: Hate online and in traditional media: A systematic review of the evidence for associations or impacts on individuals, audiences, and communities, Campbell systematic reviews 18 (2022) e1245.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[12] J. H. Park, P. Fung, One-step and two-step classification for abusive language detection on twitter, arXiv preprint arXiv:1706.01206 (2017).

[13] Y. Khan, W. Ma, S. Vosoughi, Lone pine at semeval-2021 task 5: fine-grained detection of hate speech using bertoxic, arXiv preprint arXiv:2104.03506 (2021).

[14] J. Risch, R. Krestel, Toxic comment detection in online discussions, Deep learning-based approaches for sentiment analysis (2020) 85–109.

[15] M. Taleb, A. Hamza, M. Zouitni, N. Burmani, S. Lafkiar, N. En-Nahnahi, Detection of toxicity in social media based on natural language processing methods, in: 2022 International Conference on

Intelligent Systems and Computer Vision (ISCV), 2022, pp. 1–7. doi:10.1109/ISCV54655.2022.9806096.

[16] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[17] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, I. Androutsopoulos, Toxicity detection: Does context really matter?, arXiv preprint arXiv:2006.00998 (2020).

[18] L. Laugier, J. Pavlopoulos, J. Sorensen, L. Dixon, Civil rephrases of toxic texts with self-supervised transformers, arXiv preprint arXiv:2102.05456 (2021).

[19] Z. Wang, S. Mayhew, D. Roth, et al., Extending multilingual bert to low-resource languages, arXiv preprint arXiv:2004.13640 (2020).

[20] M. Artetxe, H. Schwenk, Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond, Transactions of the association for computational linguistics 7 (2019) 597–610.

[21] S. Ranathunga, E.-S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, R. Kaur, Neural machine translation for low-resource languages: A survey, ACM Computing Surveys 55 (2023) 1–37.

[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, Cited by 1 (2022) 114–131.

[23] 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, Journal of machine learning research 21 (2020) 1–67.

[24] 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, arXiv preprint arXiv:1910.13461 (2019).

[25] 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, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[26] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[27] F. Rahutomo, T. Kitasuka, M. Aritsugi, et al., Semantic cosine similarity, in: The 7th international student conference on advanced science and technology ICAST, volume 4, University of Seoul South Korea, 2012, p. 1.