Devulgarization of Polish Texts Using Pre-trained Language Models

Cezary Klamra[1][0000−0003−4321−8862],
Grzegorz Wojdyga[1][0000−0003−4880−5142],
Sebastian Żurowski[2][0000−0003−0457−4979],
Paulina Rosalska[2,3][0000−0002−0231−5933],
Matylda Kozłowska[4][0000−0002−4571−1960], and
Maciej Ogrodniczuk[1][0000−0002−3467−9424]

[1] Institute of Computer Science, Polish Academy of Sciences
[2] Nicolaus Copernicus University in Toruń
[3] Applica.ai
[4] Oracle Poland

Abstract. We propose a text style transfer method for replacing vulgar expressions in Polish utterances with their non-vulgar equivalents while preserving the meaning of the text. We fine-tune three pre-trained language models on a newly created parallel corpus of vulgar/non-vulgar sentence pairs, then we evaluate style transfer accuracy, content preservation and language quality. To the best of our knowledge, the proposed solution is the first of its kind for Polish.

Keywords: text style transfer · removing obscenities · transformer

1 Introduction

Most works on fighting with offensiveness and obscenity in language concentrated on its automatic identification. In this paper we present a mechanism for replacing vulgar expressions in Polish utterances with their non-vulgar equivalents while preserving the overall sense of the text. As a component task, we created a corpus of vulgar expressions and their equivalents.

Due to the low availability of parallel data, most previous works on this topic use non-parallel corpora. Such solutions most often separate the content of the text from its style. A style-independent representation of the content is created and used to reconstruct the text in the target style. Dos Santos et al. [14] describe methods for moderating offensive or hateful language using unsupervised text style transfer. The authors extend the standard encoder and decoder by introducing a classifier and special loss functions that allow the use of a corpus of posts from social media. Tran et al. [15] constructed an unsupervised style transfer pipeline, that uses a vocabulary of restricted words, POS tagging to locate vulgarities, RoBERTa and T5 model to create possible replacements and GPT-2 and BLEU score to select the sentence of the highest quality. Dementieva et al. [4] used a bag-of-words logistic regression model to identify toxic

words and replace them using a BERT model, tuned on a toxic and non-toxic corpus.

With parallel corpora, Cheriyan et al. [2] used synonym generation, the masked language model (BERT), and the sequence-to-sequence model (BART) to rephrase offensive comments on social media. Dementieva et al. [4] used a parallel corpus and the pre-trained ruGPT-3 model to test three approaches: zero-shot, few-shot, and fine-tuning.

2 Theoretical Background for the Corpus Work

Our work concentrates on vulgarisms only, i.e. "lexical units by means of which the speaker reveals his or her emotions towards something or someone, breaking a linguistic taboo" [5]. Grochowski divides vulgar expressions into two categories: systemic and referential-customary. The former are lexical units tabooed solely because of its expressive features and often contain certain sequences of characters (e.g. *-kurw-*, *-jeb-*, *-pierd-*). The dictionary distinguishes three degrees of "strength" of systemic vulgarisms: (i) of the lowest level, generally considered vulgar among the "cultured interlocutors"; (ii) of the medium level, i.e. units commonly considered vulgar; (iii) of the highest level, i.e. regarded as very vulgar. The referential-customary vulgar expressions are tabooed because of their semantic features and the scope of their object reference; they are considered vulgar only in specific contexts. It is common for expressions of this type to cause classification problems since it can be difficult to determine what is their degree of vulgarity.

The construction of the corpus assumes the replacement of vulgarisms with euphemisms which are defined as substitute names used when direct names cannot be used because of negative connotations. The euphemistic term should instead evoke positive (or neutral) connotations. A remark made by Grochowski [5] situates euphemisms in a context: a given unit of language can only be considered euphemistic when juxtaposed with another unit of language (one that it can replace). It implies that vulgar expressions can also be interpreted as euphemisms when replacing even more vulgar ones. Based on this assumption we decide to annotate vulgarisms in context.

3 Training Data

In order to obtain representative training data we selected movie dialogues from websites aggregating subtitle translations. Unofficial Polish subtitle translations tend to preserve as much as 80% of vulgarisms present in the original script. The training corpus is based on two sources: (i) the Polish part of the OpenSubtitles corpus,[5] which mostly consists of texts prepared by non-professional translators [6]; (ii) professionally edited dialogue tracks and published scripts of two movies [7,8].

[5] See http://opus.nlpl.eu/OpenSubtitles.php, http://www.opensubtitles.org/

The samples were selected by annotators experienced in linguistic work who also collected and supplemented the material as necessary. The annotations collected include:

- the identifier of the text from which the annotated passage came,
- the annotated vulgar expression and its context,
- the lemma of the annotated expression,
- vulgar equivalents (synonyms) of the vulgar expression,
- common equivalents (synonyms) of the expression,
- euphemistic equivalents (synonyms) of the expression.

In order to establish synonymy relations and to determine the character of particular expressions, the annotators were advised to use the Dictionary of Polish Swearwords and Vulgarisms, the Dictionary of Polish Euphemisms, or general dictionaries of Polish.

Since the corpus contains more than one substitution for some vulgar expressions, it was pre-processed to obtain pairs with different euphemisms for a given context. The grammatical forms of vulgar expressions were manually inflected so that they could be appropriately used in their contexts. The process resulted in 6691 pairs of vulgar and corresponding non-vulgar texts.

4 Devulgarization Tool

For this work, we solve the problem of substitution of vulgar expressions as a text style transfer problem defined as a reformulation of a text without affecting its content and other properties, such as sentiment, bias, degree of formality, etc. Below we compare three approaches using GPT-2, GPT-3, and T5 models. The training details have been shared online (see section 6).

4.1 GPT-2-based Solution

The first presented solution is based on GPT-2 [11], a transformer-type language representation used for language generation but also in sequence-to-sequence tasks by concatenating input and output sequences. The pre-trained model was fine-tuned on concatenated pairs of vulgar and non-vulgar texts from the training corpus (separated with the special token `<|sep|>`). We used papuGaPT-2 model pre-trained on Polish language texts, based on GPT-2 small.[6]

4.2 GPT-3-based Solution

GPT-3 is the latest model in the GPT-n series, largely based on its predecessor but much more extensive. Although GPT-3 is only available for English, Brown et al. [1] argue that the model has some ability to process languages other than English. English words accounted for about 93% of all words in the training

[6] https://huggingface.co/flax-community/papuGaPT2, accessed on April 12, 2022

set, while Polish for only 0.15% but representing over 300 million occurrences nonetheless.[7]

GPT-3 has not been made publicly available but it is possible to use the model, including model tuning, through the OpenAI API.[8] The presented results were obtained using the second most extensive variant of the model available, the Curie model, tuned on the test corpus. The GPT-3 training adopted a similar approach as GPT-2: the training data were presented as pairs of vulgar prompts and non-vulgar completions.

4.3 T5-based Solution

The last presented solution is based on T5 transformer-based model [12]. As a sequence-to-sequence model, T5 is well suited for tasks related to text-based language generation, such as the problem of changing the style of the text. We fine-tuned plT5, a T5-based model pre-trained on large Polish corpora.[9] The model is available in three variants (small, base, and large) differing in the number of parameters. The paper presents the results obtained for the two most extensive models, both of which were fine-tuned on the training corpus.

5 Evaluation

5.1 Evaluation Method

The models were evaluated on the test corpus consisting of 2437 vulgar sentences from the Dictionary of Polish Swearwords and Vulgarisms [5]. We used all of the examples provided in the dictionary, some of which are not strictly vulgar.

The evaluation follows the method used for text style modification [9,15,4], i.e. the quantitative quality assessment of the obtained results in three categories: (i) effectiveness of text style change, (ii) preservation of the content of the original sentence, and (iii) quality of the generated language.

Text style transfer accuracy (STA) was tested using Przetak — a library for checking whether a text contains abusive or vulgar speech in Polish written in Go [3]. Przetak is able to identify offensive language with high accuracy and handles frequent misspellings and out-of-vocabulary words composed of morphemes with vulgar meaning. Evaluated sentences are assigned a score of 0 if they are classified as vulgar, and 1 otherwise.

Content preservation was checked using three metrics: (i) cosine similarity (CS) between embeddings for input and output sentences, calculated using the SBERT multilingual model [13]; (ii) word overlap (WO) between the lemmata of the original (X) and the processed (Y) sentence defined as:

$$\frac{\#(X \cap Y)}{\#(X \cup Y)} \tag{1}$$

[7] https://github.com/openai/gpt-3/tree/master/dataset_statistics, accessed on October 5, 2021

[8] https://beta.openai.com/, accessed on October 5, 2021

[9] https://huggingface.co/allegro/plt5-large, accessed on December 12, 2021

where lemmata have been produced using spaCy; (iii) BLEU, a commonly used and well correlated with human ratings metric for assessing machine translation quality [10].

Quality of the generated language was measured with perplexity (PPL) determined using the pre-trained GPT-2 small model.

As suggested by Pang and Gimpel [9], in order to compare the overall quality of the results we used the geometric mean (GM) of the corresponding metrics in the above categories, according to the formula 2.

$$GM = \Big(\big[100 \cdot max(0, CS)\big] \cdot \big[100 \cdot max(0, STA)\big] \cdot max(0, \frac{1}{PPL}) \Big)^{\frac{1}{3}} \quad (2)$$

5.2 Baselines

Following other studies (see [15,4]), the results achieved by presented methods are compared against two baselines:

- **Delete** — letters of words classified as vulgar by Przetak (described in subsection 5.1) are replaced with asterisks. The first letter of a vulgar word is not changed. This method allows preservation of the content well unless the meaning of the vulgar word is crucial for the understanding of the sentence.
- **Duplicate** — an unchanged copy of the original sentence. This baseline represents the lower bound of the models' performance.

As the models are trained to substitute some expressions which are not recognised by Przetak, metrics measuring content preservation might have disproportionately high values for the Delete baseline.

5.3 Quantitative Evaluation Results

The results of the automatic evaluation are presented in Table 1. In comparison with GPT models, T5 models generate higher-quality language, achieve better results in terms of preserving the content of the original sentence, and the corresponding values of the combined metric are higher.

Table 1: Automatic evaluation results

	style transfer	content preservation			language quality	GM
	STA	**CS**	**WO**	**BLEU**	**PPL**	
Duplicate	0.38	1	1	1	146.86	1.78
Delete	1	0.93	0.84	0.92	246.80	4.14
GPT-2	0.90	0.86	0.71	0.86	258.44	3.71
GPT-3	0.88	0.92	0.79	0.92	359.12	3.58
T5 base	0.90	0.97	0.85	0.95	187.03	4.10
T5 large	**0.93**	**0.97**	**0.86**	**0.95**	**170.02**	**4.31**

5.4 Qualitative Evaluation Results

The presented models can, in most cases, replace vulgar words in the sentence with equivalent non-vulgar words while not changing the rest of the sentence. Replacements usually have an appropriate grammatical form and convey well the sense of the original sentence. Furthermore, the models can often cope with the replacement of the same vulgarism used in different, although homogeneous, grammatical forms or meaning. In situations where the replacement has a different grammatical form from the original (e.g. wrt. gender or number), the models can sometimes generate correct grammatical forms of the subordinate phrases. Some of the processed sentences contain swear words that have not been replaced by euphemisms.

The general effect of model inference is a decrease in the quality of language – output sentences contain numerous typos and modifications of proper names, cases, or punctuation. GPT-2 and GPT-3 models tend to modify non-vulgar parts of sentences much more frequently than T5 models, using synonyms or antonyms. In some cases, there appear words or phrases semantically unrelated to the original sentence. In individual cases, after generating such a word, the model terminates the text generation or adds a sentence ending, which is not coherent or contains numerous repetitions of certain word sequences. Such problems occur much less frequently in sentences processed by T5 models. At the same time, T5 models often perform better with sentences that are more complex, contain numerous proper names, complicated punctuation, or lower-quality language.

6 Conclusions and Future Work

This work is the first study of text devulgarization in Polish. We conducted experiments with three pre-trained models: GPT-2, GPT-3, and T5. The models were trained on a corpus of vulgar texts in Polish we created especially for this task. We developed an evaluation setup for the problem of substituting offensive language in Polish texts which aims to assess three aspects of the task — style transfer accuracy, content preservation, and language quality. Finally, we evaluated the presented methods and made all resources available.[10]

The results show that the tested approaches can be successfully used for removing offensive language, although there is room for improvement. All of the described approaches could benefit from more careful hyperparameter tuning as well as a larger training corpus. As the availability of large parallel corpora is limited, non-parallel methods for Polish could prove effective. Furthermore, we have only considered substituting profane words with euphemistic equivalents, while in some cases, simply removing the word seems to be the most appropriate strategy.

[10] http://clip.ipipan.waw.pl/DEPOTx (from *Devulgarization of Polish Texts*)

Acknowledgements The work was financed by the European Regional Development Fund as a part of 2014–2020 Smart Growth Operational Programme, CLARIN — Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19 and supported by the Poznan Supercomputing and Networking Center grant number 442. We kindly thank Professor Maciej Grochowski for giving us permission to use the full source material of the Dictionary of Polish Swearwords and Vulgarisms.

References

1. Brown, T.B., et al.: Language Models are Few-Shot Learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Adv. Neural. Inf. Process. Syst. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)
2. Cheriyan, J., et al.: Towards Offensive Language Detection and Reduction in four Software Engineering Communities. In: Eval. and Assess. in Softw. Eng. pp. 254–259. Assoc. Comput. Mach., New York, USA (2021)
3. Ciura, M.: Przetak: Fewer Weeds on the Web. In: Ogrodniczuk, M., Kobyliński, Ł. (eds.) Proc. PolEval 2019 Workshop. pp. 127–133. Institute of Computer Science, Polish Academy of Sciences (2019)
4. Dementieva, D., et al.: Methods for Detoxification of Texts for the Russian Language. Multimodal Technol. Interact. **5**, 54
5. Grochowski, M.: Słownik polskich przekleństw i wulgaryzmów. PWN Scientific Publishers (2008)
6. Lison, P., Tiedemann, J.: OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In: Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC 2016). pp. 923–929. ELRA, Portorož, Slovenia (2016)
7. Miławska, M.: Harmonia czy dysonans? O wulgaryzmach w „Dniu świra" Marka Koterskiego. Słowo. Studia językoznawcze (4), 188–199 (2013)
8. Osadnik, W.: Przeklinanie na ekranie, czyli o tłumaczeniu wulgaryzmów w napisach filmowych. In: P. Fast, N.S. (ed.) Tabu w przekładzie, pp. 61–71. Wydawnictwo Naukowe „Śląsk", Katowice (2007)
9. Pang, R.Y., Gimpel, K.: Unsupervised Evaluation Metrics and Learning Criteria for Non-Parallel Textual Transfer. In: Proc. 3rd Workshop Neural Gener. Transl. pp. 138–147. ACL, Hong Kong (2019)
10. Papineni, K., et al.: Bleu: a method for automatic evaluation of machine translation. In: Proc. 40th Annu. Meet. of ACL. pp. 311–318 (2002)
11. Radford, A., et al.: Language Models are Unsupervised Multitask Learners. OpenAI blog **1**(8) (2019)
12. Raffel, C., et al.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020)
13. Reimers, N., Gurevych, I.: Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In: Proc. 2020 Conf. Empir. Methods in Nat. Lang. Process. (EMNLP). pp. 4512–4525. ACL, Online (2020)
14. dos Santos, C.N., et al.: Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer. In: Proc. 56th Annu. Meet. of ACL. pp. 189–194. ACL, Melbourne, Australia (2018)
15. Tran, M., et al.: Towards a Friendly Online Community: An Unsupervised Style Transfer Framework for Profanity Redaction. In: Proc. 28th Int. Conf. Comput. Linguist. pp. 2107–2114. Int. Comm. Comp. Linguist., Barcelona, Spain (Online) (2020)