Dewulgaryzacja polskich tekstów z wykorzystaniem pretrenowanych modeli językowych

Cezary Klamra, Grzegorz Wojdyga (Instytut Podstaw Informatyki PAN),
Sebastian Żurowski (Uniwersytet Mikołaja Kopernika),
Paulina Rosalska (Uniwersytet Mikołaja Kopernika / Applica.ai),
Matylda Kozłowska (Oracle Polska),
Maciej Ogrodniczuk (Instytut Podstaw Informatyki PAN)

6 czerwca 2022

Plan prezentacji

1. Wprowadzenie
2. Przyjęte podejście
3. Korpus wyrażeń wulgarnych i ich substytutów
4. Przygotowanie narzędzia
5. Ewaluacja
6. Podsumowanie

Wprowadzenie

Motywacja

Problem obraźliwego języka podejmowany był do tej pory głównie w kontekście metod automatycznej identyfikacji wulgarnych lub toksycznych wyrażeń.

Potrzebne są jednak dalej idące rozwiązania, umożliwiające bardziej aktywne przeciwdziałanie szerzeniu się niecenzuralnych treści w przestrzeni internetowej.

Wprowadzenie

Zmiana stylu tekstu

Cel: opracowanie narzędzia umożliwiającego substytucję wyrażeń wulgarnych przy jednoczesnym zachowaniu sensu oryginego tekstu.

Zmiana stylu tekstu (ang. *text style transfer*) – zadanie z dziedziny generowania języka naturalnego, w którym celem jest modyfikacja stylu zadanego tekstu przy jednoczesnym zachowaniu jego treści oraz innych własności (np. wydźwięku, stopnia formalności, obecności biasu).

Wprowadzenie

Zmiana stylu tekstu

Dwa możliwe podejścia do zmiany stylu tekstu:
budowa narzędzia w oparciu o korpus równoległy (Cheriyan i in., 2021; Dementieva i in., 2021)
lub nierównoległy (dos Santos i in., 2018; Tran i in., 2020)

Plan prezentacji

Przyjęte podejście

Wykorzystanie korpusu równoległego.
Problem zmiany stylu tekstu zostaje sprowadzony do „przetłumaczenia” zadanego zdania z języka wulgarnego na niewulgarny.

Przyjęte podejście

- Zestawiliśmy korpus wyrażeń wulgarnych oraz ich zamienników.
- Na postawie korpusu stworzyliśmy narzędzie umożliwiające automatyczną redakcję wyrażeń wulgarnych. Narzędzie przygotowaliśmy w czterech wariantach – w oparciu o modele GPT-2, GPT-3 oraz T-5 base i large.
- Oceniliśmy uzyskane wyniki na trzech płaszczyznach przy pomocy automatycznych metryk.

Plan prezentacji

Korpus wyrażeń wulgarnych

Źródła danych

Teksty wulgarne w korpusie zaczerpnięto z oficjalnych transkrypcji filmów „Psy” i „Psy 2” (1992 i 1994, reż. Władysław Pasikowski) oraz „Dzień Świra” (2002, reż. Marek Koterski), a także z nieoficjalnie opracowanych napisów dostępnych w serwisie OpenSubtitles.

Nieoficjalne napisy zachowują większość (nawet 80%) wulgaryzmów obecnych w filmach (Włuka, 2019).
W niektórych przypadkach przekłady mogą być nawet bardziej wulgarne od oryginału (Gruszczyńska, 2019).

Korpus wyrażeń wulgarnych

Anotacja danych

Korpus składa się z 3000 kontekstów zawierających wulgaryzmy. Teksty zostały zanotowane przez osoby z doświadczeniem w pracach lingwistycznych.

W anotacji uwzględniono:

- identyfikator tekstu źródłowego,
- wyrażenie wulgarne w użytej w tekście formie,
- lemat wyrażenia,
- wulgarne synonimy wyrażenia,
- pospolite synonimy wyrażenia,
- eufemistyczne synonimy wyrażenia.

Korpus wyrażeń wulgarnych

Anotacja danych

Przy wyborze synonimów wyrażeń wulgarnych anotatorzy posługiwali się Słownikiem polskich przekleństw i wulgaryzmów (Grochowski, 2008), Słownikiem eufemizmów polskich (Dąbrowska, 2021) oraz słownikami ogólnymi.

Korpus wyrażeń wulgarnych

Anotacja danych

ID	othr1
Kontekst	— Wiążąc, poluźnij końce. — Postaw mnie, ty łysy chuju! — Błagam cię, nie rób tego.
Wyrażenie	chuju
Lemat	chuj
Wulg.	pała / fiut
Posp.	ciul / menda
Euf.	drań / łajdak

Plan prezentacji

Przygotowanie narzędzia

Korpus treningowy

Na podstawie zaprezentowanego korpusu wyrażeń wulgarnych utworzono równoległy korpus treningowy składający się z 6691 par zdań wulgarnych i niewulgarnych.

Tekst wulgarny	**Tekst niewulgarny**
— On jest tu, żeby się nas pozbyć. — Jestem tu, żeby uratować *twoją zawszoną dupę*. — Nie możesz nikogo uratować.	— On jest tu, żeby się nas pozbyć. — Jestem tu, żeby uratować *twój zawszony tyłek*. — Nie możesz nikogo uratować.
— On jest tu, żeby się nas pozbyć. — Jestem tu, żeby uratować twoją *zawszoną dupę*. — Nie możesz nikogo uratować.	— On jest tu, żeby się nas pozbyć. — Jestem tu, żeby uratować twoją *skórę*. — Nie możesz nikogo uratować.

Przyjęte podejście

Wykorzystane modele

Przeprowadziliśmy eksperymenty z czterema pretrenowanymi reprezentacjami językowymi:

- polskim GPT-2 small,[1]
- angielskim GPT-3 (model Curie),
- polskimi modelami T5 base[2] i large.[3]

Każdy z powyższych modeli został poddany dostrajaniu na równoległym korpusie treningowym.

---

[1] `https://huggingface.co/flax-community/papuGaPT2`

[2] `https://huggingface.co/allegro/plt5-base`

[3] `https://huggingface.co/allegro/plt5-large`

Plan prezentacji

Ocena ilościowa

Działanie modeli zostało ocenione na korpusie 2347 wulgarnych zdań pochodzących ze Słownika polskich przekleństw i wulgaryzmów (Grochowski, 2008) przy użyciu automatycznych metryk.

Teksty należące do korpusu testowego pochodzą z innego źródła niż teksty korpusu treningowego.

Ocena ilościowa

Ocenę wyników przeprowadzono na trzech płaszczyznach:

1. skuteczności zmiany stylu tekstu
2. zachowania treści oryginalnego zdania
3. jakości wygenerowanego języka

Jakość wyników w każdej z powyższych kategorii jest odwrotnie proporcjonalna do jakości w dwóch pozostałych (Pang i Gimpel, 2019).

Ocena ilościowa

Automatyczne metryki

Skuteczność zmiany stylu

Skuteczność zmiany stylu tekstu została oceniona przy pomocy biblioteki Przetak[4] (Ciura, 2019). Wartość metryki **STA** wynosiła 0 dla zdań, w które zostały ocenione jako wulgarne, oraz 1 w przeciwnym przypadku.

Jakośc wygenerowanego języka

Jakość języka oceniono przy pomocy **nieokreśloności (PPL)**, oszacowanej przy pomocy polskiego modelu GPT-2 small[5]

[4] `https://github.com/MarcinCiura/przetak`

[5] `https://huggingface.co/flax-community/papuGaPT2`

Ocena ilościowa

Automatyczne metryki

Zachowanie treści

- **Podobieństwo cosinusowe (CS)** wektorów zdania oryginalnego i przetworzonego, uzyskanych przy pomocy wielojęzycznego modelu SBERT[6],
- **Pokrycie leksykalne (WO)** pomiędzy lematami słów należących do zdania oryginalnego (X) i przetworzonego (Y), wyznaczone zgodnie z wzorem: $\frac{\#(X \cap Y)}{\#(X \cup Y)}$. Lematy zostały wyznaczone przy pomocy biblioteki spaCy[7].
- **BLEU**

[6] `https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1`

[7] `https://spacy.io/models/pl#pl_core_news_lg`

Ocena ilościowa

Automatyczne metryki

Metryka łączona

Ogólna jakość wyników została oceniona przy pomocy metryki łączonej – **średniej geometrycznej (GM)** metryk w każdej z wcześniej omówionych kategorii, zgodnie z wzorem:

$$\mathrm{GM} = \left(\left[100 \cdot \max(0, \mathrm{CS}) \right] \cdot \left[100 \cdot \max(0, \mathrm{STA}) \right] \cdot \max(0, \frac{1}{\mathrm{PPL}}) \right)^{\frac{1}{3}}$$

Ocena ilościowa

Baseline

W ocenie wyników uwzględniono dwie dodatkowe metody pełniące rolę punktów odniesienia (*baselines*):

- **Duplicate** – bezpośrednia kopia oryginalnego zdania
- **Delete** – wszystkie znaki wulgarnych wyrazów, z wyjątkiem pierwszego, są zastąpione symbolem gwiazdki (∗)

Ocena ilościowa

Wyniki

	STA	CS	WO	BLEU	PPL	GM
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**

Ocena jakościowa

Prezentowane modele w większości przypadków są w stanie zastąpić wyrażenia wulgarne występujące w zdaniu równoważnymi wyrażeniami niewulgarnymi, jednocześnie nie zmieniając pozostałej części zdania.

Zamienniki wyrażeń zawierających przekleństwa w przetworzonych zdaniach najczęściej mają odpowiednią formę gramatyczną oraz dobrze oddają sens, w jakim wulgaryzm występował w oryginalnym zdaniu.

Ocena jakościowa

Ogólnym efektem przetwarzania przez wszystkie z prezentowanych modeli jest obniżenie jakości języka: literówki, modyfikacje nazw własnych, wielkości liter, interpunkcji.

Oprócz tego przetworzone teksty często zawierają innego rodzaju błędy, na przykład:

- zastępowanie niewulgarnych części zdania synonimami lub antonimami,
- obecność słów niezwiązanych z pozostałą częścią zdania,
- obecność powtarzających się sekwencji wyrazów.

Ocena jakościowa

Takie problemy znacznie częściej występują w zdaniach przetworzonych przez modele GPT-2 i GPT-3.

Jednocześnie modele T-5 często radzą sobie lepiej ze zdaniami, które są bardziej rozbudowane, zawierają trudne nazwy własne, skomplikowaną interpunkcję lub język niskiej jakości

Przykłady działania narzędzia

zdanie wuglarne	Oni są jak pieczarki – w gównie siedzą i gówno wiedzą.
GPT-2	Oni są jak pieczarki – w bagnie siedzą i guzik wiedzą.
GPT-3	Oni są jak pieczarki – w kłamie siedzą i Pośpieszą.
T-5 base	Oni są jak pieczarki – w bagnie siedzą i guzik wiedzą.
T-5 large	Oni są jak pieczarki – w bagnie siedzą i bagno wiedzą.

Przykłady działania narzędzia

zdanie wuglarne	Rano pakował nad swoim blokiem ruskie zenity i kryształy albo inne gówna.
GPT-2	Rano pakował nad swoimBlokiem zenity i kryształy albo inny badziewie.
GPT-3	Rano pakował nad swoim blokiem rurania i kryształy albo inne rzeczy.
T-5 base	Rano pakował nad swoim blokiem ruskie zenity i kryształy albo inne bzdury.
T-5 large	Rano pakował nad swoim blokiem ruskie zenity i kryształy albo inny syf.

Przykłady działania narzędzia

zdanie wuglarne	Tak więc przyjebać to uderzyć, dojebać – pobić, zajebać – zabić.
GPT-2	Tak więc przypierniczać to uderzyć, wziąć od nas haczyk, wziąć od nas haczyk.
GPT-3	Tak więc przypieprzyć to uderzyć, dojebać – pobić, zapieprzyć – zabić.
T-5 base	Tak więc przypieprzyć to uderzyć, dopieprzyć – pobić, zapieprzyć – zabić.
T-5 large	Tak więc przypieprzać to uderzyć, dopieprzyć – pobić, zapieprzać – zabić.

Plan prezentacji

Możliwości poprawy działania

Jakość działania modeli mogłaby być poprawiona poprzez bardziej przemyślany dobór wartości hiperparametrów treningu oraz wykorzystanie większego korpusu treningowego.

Obecnie prowadzone są prace nad korpusem wulgaryzmów z tekstów pochodzących z Narodowego Korpusu Języka Polskiego, scenariuszy filmowych, Słownika Polskich Przekleństw i Wulgaryzmów oraz tekstów z internetu.

Konkluzje

Prezentowane rozwiązania nie zawsze zwracają oczekiwane wyniki, jednak pokazują, że pretrenowane reprezentacje językowe mogą z sukcesem być wykorzystane do redagowania wulgarnych treści w języku polskim.

Według naszej najlepszej wiedzy jest to pierwsza próba rozwiązania zadania zmiany stylu tekstu dla języka polskiego.

Klamra C., Wojdyga G., Żurowski S., Rosalska P., Kozłowska M., Ogrodniczuk M. (2022). *Devulgarization of Polish Texts Using Pre-trained Language Models*. To appear in Proceedings of International Conference on Computer Science (ICCS 2022).

Materiały

Korpus wyrażeń wulgarnych i ich substytutów, wytrenowane modele oraz pozostałe materiały dostępne są pod adresem

`http://clip.ipipan.waw.pl/DEPOTx`

(od *Devulgarization of Polish Texts*)

Podziękowania

Dziękujemy prof. Maciejowi Grochowskiemu za wyrażenie zgody na wykorzystanie całości materiału ze Słownika polskich przekleństw i wulgaryzmów.

Oświadczenie

Pracę sfinansowano ze środków Europejskiego Funduszu Rozwoju Regionalnego w ramach Programu Operacyjnego Inteligentny Rozwój 2014–2020, „CLARIN — Wspólne Zasoby Językowe i Infrastruktura Technologiczna”, nr projektu POIR.04.02.00-00C002/19. Praca była wspierana przez Poznańskie Centrum Superkomputerowo-Sieciowe, grant nr 442.

Bibliografia

Cheriyan, Jithin i in. (2021). „Towards Offensive Language Detection and Reduction in four Software Engineering Communities”. W: *Eval. and Assess. in Softw. Eng*. Trondheim, Norway: Assoc. Comput. Mach., s. 254–259. ISBN: 9781450390538.

Ciura, Marcin (2019). „Przetak: Fewer Weeds on the Web”. W: *Proc. PolEval 2019 Workshop*. Red. Maciej Ogrodniczuk i Łukasz Kobyliński. Institute of Computer Science, Polish Academy of Sciences, s. 127–133.

Dąbrowska, Anna (2021). *Słownik eufemizmów polskich, czyli w rzeczy mocno, w sposobie łagodnie*. 3 wyd. PWN Scientific Publishers.

Dementieva, Daryna i in. (2021). „Methods for Detoxification of Texts for the Russian Language”. W: *Multimodal Technol. Interact.* 5, s. 54.

dos Santos, Cıcero Nogueira i in. (2018). „Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer”. W: *Proc. 56th Annu. Meet. of ACL*. Melbourne, Australia: ACL, s. 189–194.

Grochowski, Maciej (2008). *Słownik polskich przekleństw i wulgaryzmów*. PWN Scientific Publishers.

Gruszczyńska, Ewa (2019). „Wulgaryzmy w dyskursie medialnym a przekład”. W: *Translatoryczne i dyskursywne oblicza komunikacji*. Red. A. Szczęsny E. Gruszczyńska M. Guławska-Gawkowska. Faculty of Applied Linguistics, University of Warsaw, s. 169–183.

Pang, Richard Yuanzhe i Kevin Gimpel (2019). „Unsupervised Evaluation Metrics and Learning Criteria for Non-Parallel Textual Transfer”. W: *Proc. 3rd Workshop Neural Gener. Transl.* Hong Kong: ACL, s. 138–147.

Tran, Minh i in. (2020). „Towards a Friendly Online Community: An Unsupervised Style Transfer Framework for Profanity Redaction”. W: *Proc. 28th Int. Conf. Comput. Linguist.* Barcelona, Spain (Online): Int. Comm. Comp. Linguist., s. 2107–2114.

Włuka, Mateusz (2019). „Wulgaryzmy w amatorskim tłumaczeniu filmowym”. W: *Rocznik Przekładoznawczy* 14, s. 365–378.