Neuro-symboliczna sztuczna inteligencja w walce z cyberprzemocą w Internecie

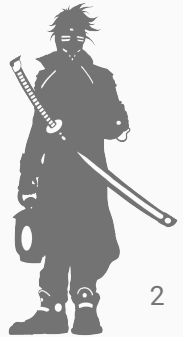
dr inż. Michał Marcińczuk (Samurai Labs)
michal.marcinczuk@samurailabs.ai

MLGdańsk #126
6 czerwca 2022 r.

Agenda

samurai LABS

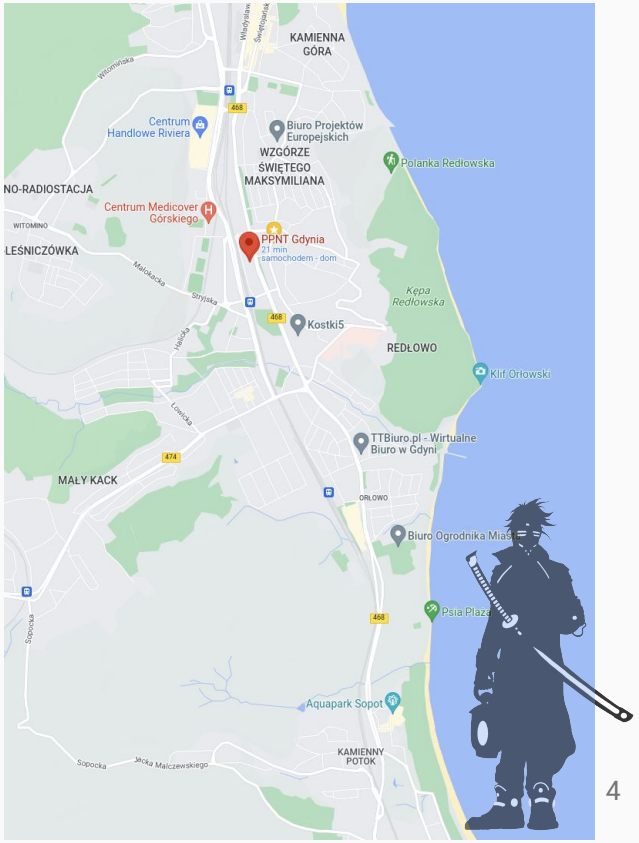
1. O nas
2. Główne usługi
3. Badania
4. Metody symboliczne + maszynowe uczenie



Wprowadzenie

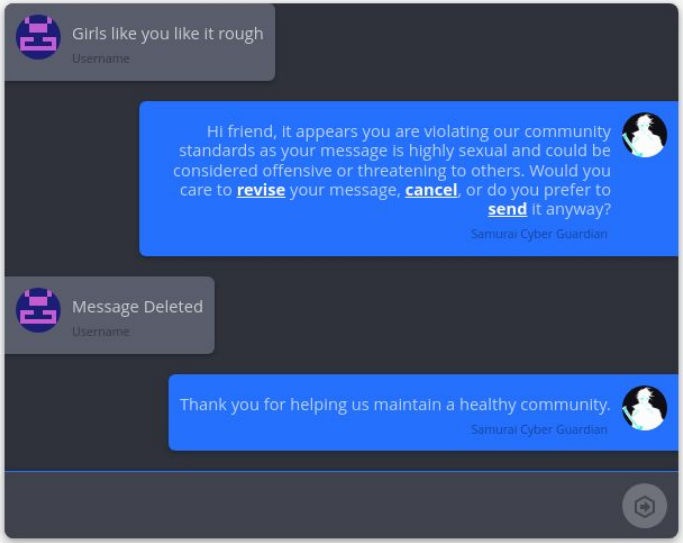
O nas

- **Lokalizacja** — Gdynia (PPNT)
- **Misja** — “Wierzymy, że Internet powinien być bezpieczny dla każdego. Naszą misją jest ochrona dzieci przed przemocą. Jesteśmy pionierami technologii, która sprawia, że internet będzie bezpiecznym miejscem przez walkę z cybernękaniem, prześladowaniem i przemocą internetową.”
- **Interdyscyplinarny zespół** — sztuczna inteligencja, kognitywistyka, lingwistyka, pedagogika, psychologia, matematyka, informatyka, biznes.
- **Obszary zainteresowania** — analiza treści pod kątem wykrywania: cyberprzemocy, mowy nienawiści, treści suicydalnych, pedofilii.



Główne usługi

Moderacja treści — Cyber Guardian



76% of cyberbullies voluntarily change their tone after being educated and redirected by Samurai's Cyber Guardian

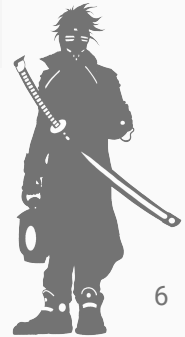
Blackmail Detection Example:

Send me more nudes or I'll publish the ones I already have

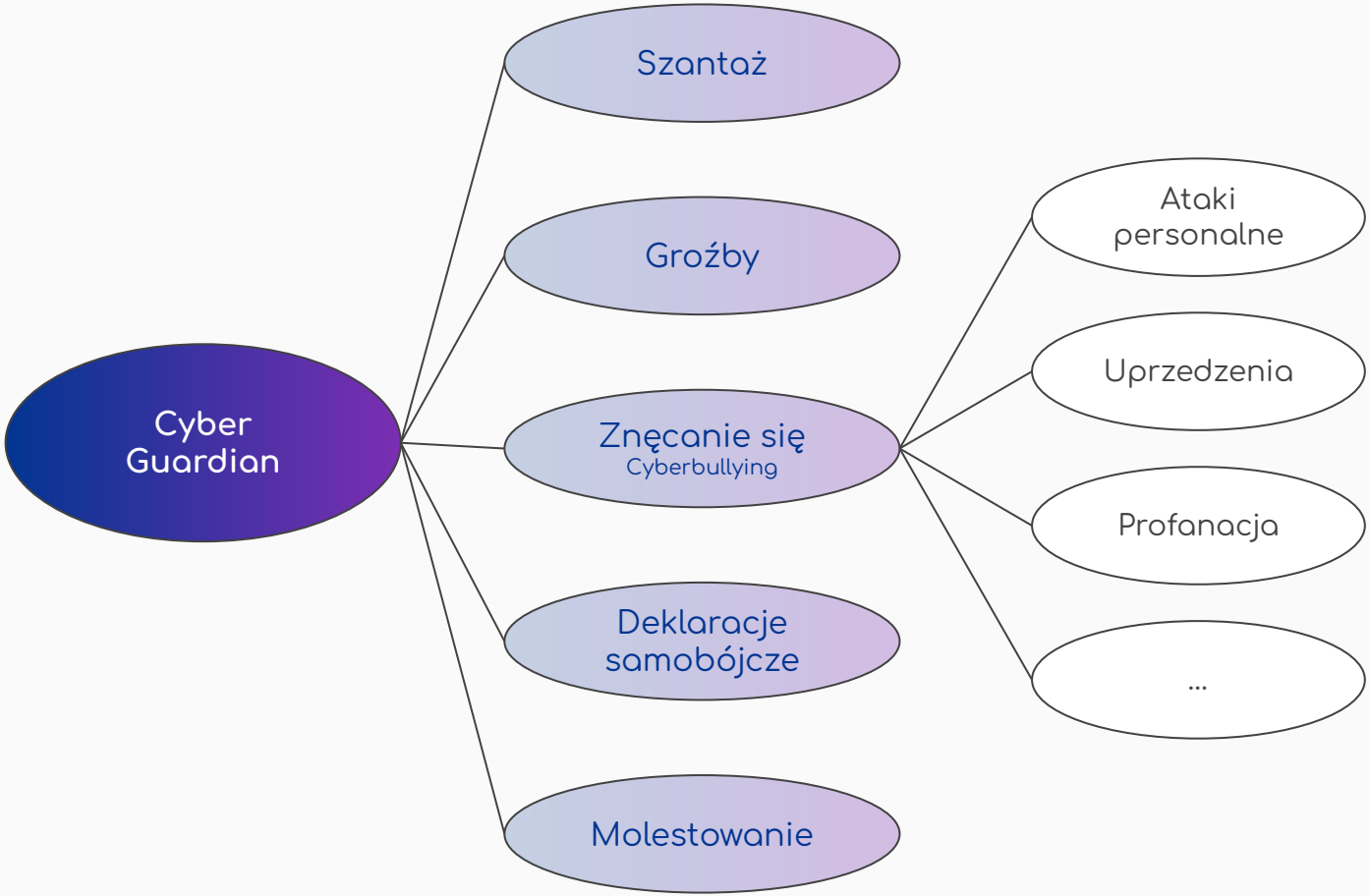
Request Alternative Negative Consequences

Samurai Cyber Guardian Detection
Blackmail threat to reveal information
Sexual harassment attempt to solicit intimate photographs

https://username.samurailabs.ai/#cyberguardian

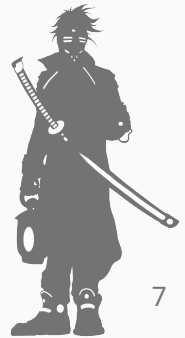


Moderacja treści — Cyber Guardian cd.

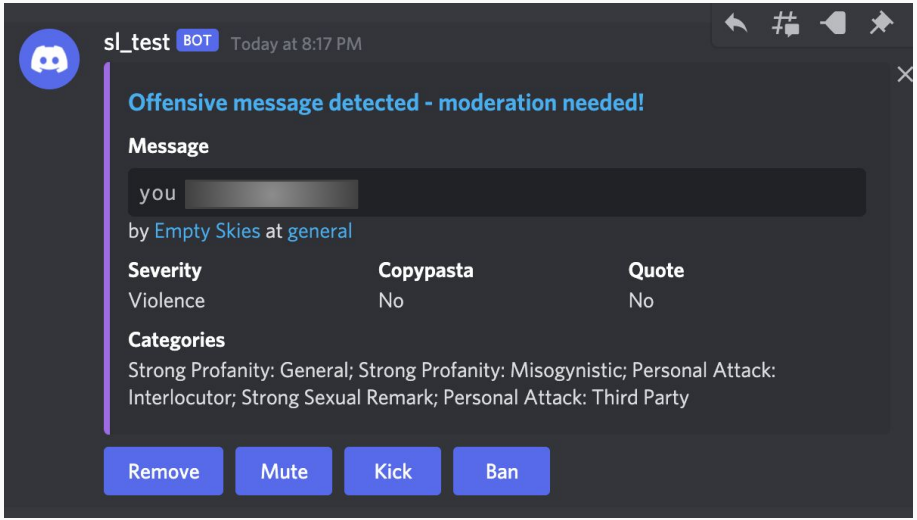


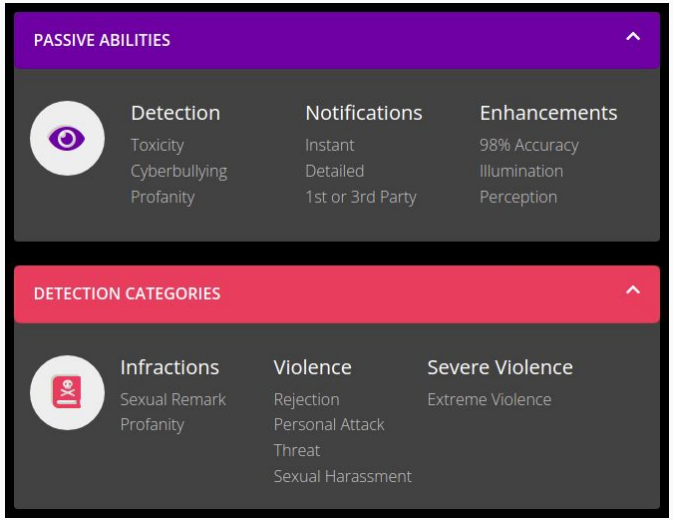
Podział ze względu na odbiorcę: druga osoba, trzecia osoby, grupy mniejszościowe, politycy, itd.

Podział ze względu na siłę: poważne, silne, łagodne

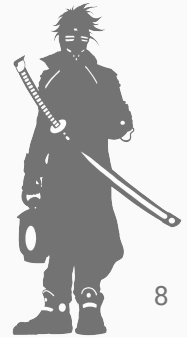


S4m — Discord Bot

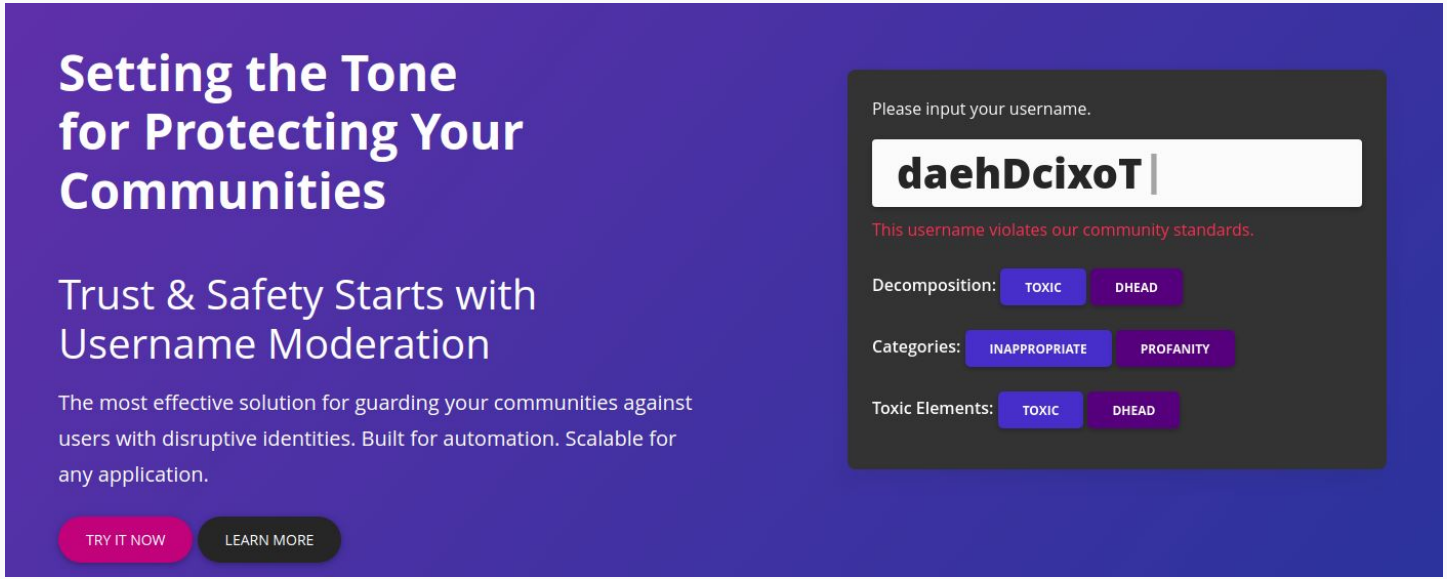




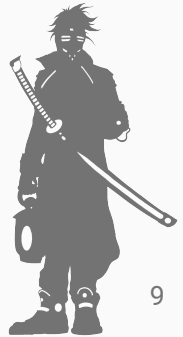
https://s4m.ai



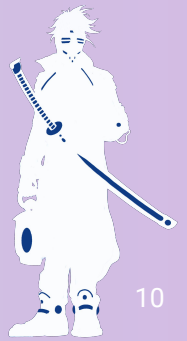
Moderacja nazw użytkowników



https://username.samurailabs.ai

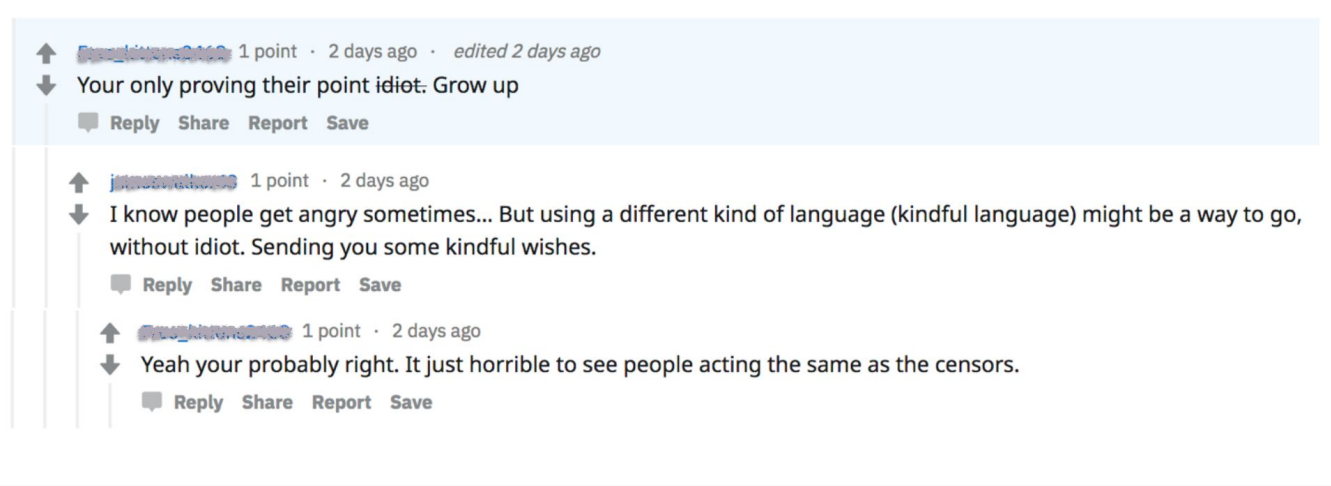


Badania



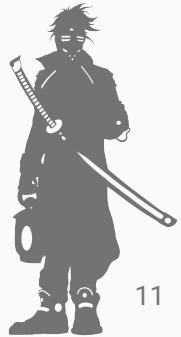
samurai LABS

- Bot działający na reddicie **r/MensRights**
- Wykrywał i reagował na przemoc słowną
- Reakcją była odpowiedź na post, której celem było “ostudzenie” emocji

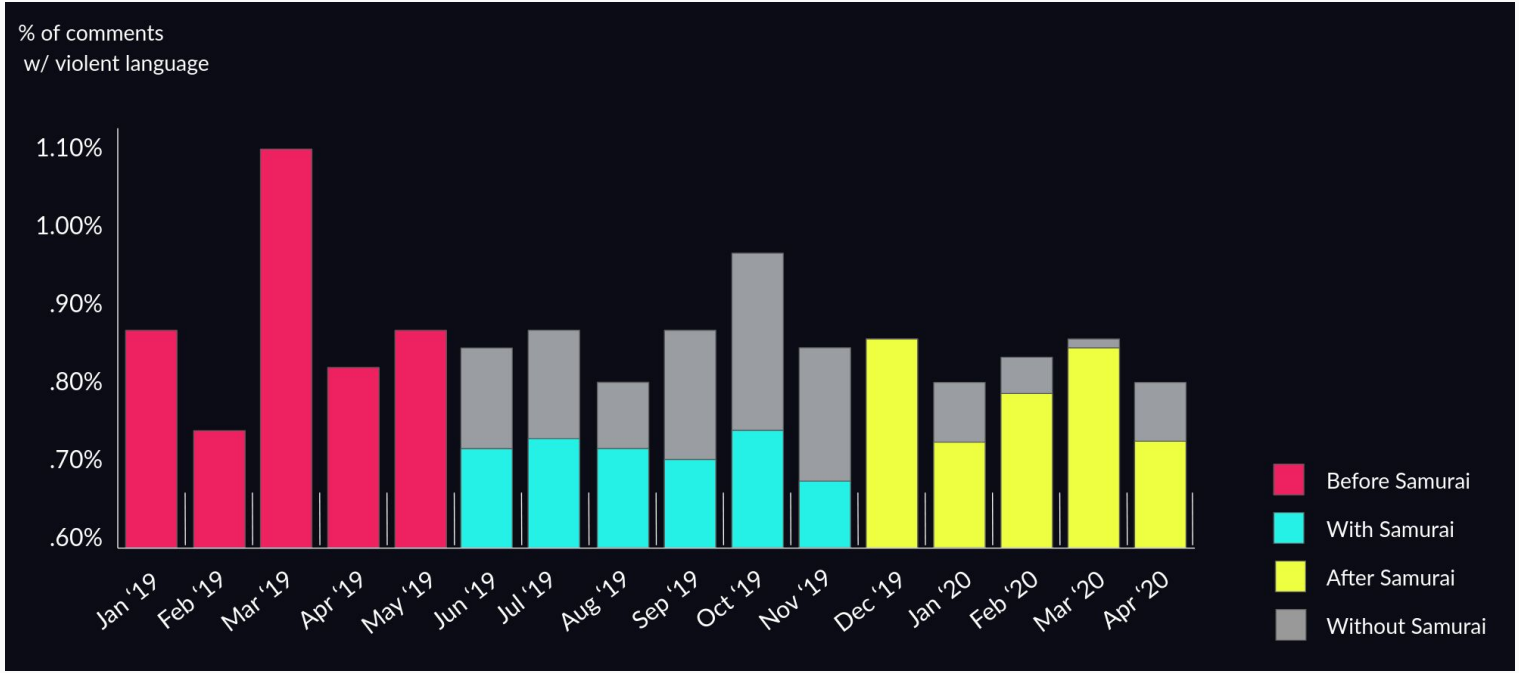


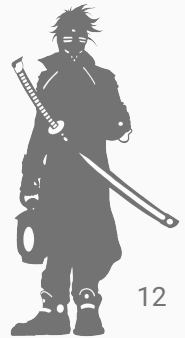
Michał Bilewicz, Patrycja Tempska, Gniewosz Leliwa, Maria Dowgiałło, Michalina Tańska, Rafał Urbaniak, Michał Wroczyński
Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment, Aggressive Behavior. 2021;47:260–266

https://onlinelibrary.wiley.com/doi/abs/10.1002/ab.21948



James Walker cd.





Treści suicydalne na Redditcie

- Treści suicydalne na Reddit-cie

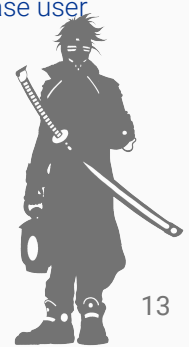
Michal Ptaszynski, Monika Zasko-Zielinska, Michal Marcinczuk, Gniewosz Leliwa, Marcin Fortuna, Kamil Soliwoda, Ida Dziublewska, Olimpia Hubert, Pawel Skrzek, Jan Piesiewicz, Paula Karbowska, Maria Dowgiallo, Juuso Eronen, Patrycja Tempska, Maciej Brochocki, Marek Godny and Michal Wroczynski. Looking for Razors and Needles in a Haystack: Multifaceted Analysis of Suicidal Declarations on Social Media—A Pragmalinguistic Approach, Int. J. Environ. Res. Public Health 2021, 18(22), 11759

https://www.mdpi.com/1660-4601/18/22/11759

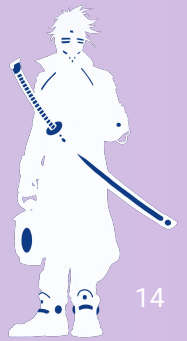
- Wpływ ataków personalnych na aktywność użytkowników Reddit-a

Rafal Urbaniaka, Michał Ptaszyński, Patrycja Tempska, Gniewosz Leliwa, Maciej Brochocki, Michał Wroczyński. Personal attacks decrease user activity in social networking platforms. Computers in Human Behavior Volume 126, January 2022, 106972

https://www.sciencedirect.com/science/article/abs/pii/S0747563221002958



Modele symboliczne i maszynowe uczenie



Metody symboliczne vs. maszynowe uczenie

Zalety

- Pełna kontrola nad działaniem modelu,
- Przewidywalne zachowanie,
- Wsparcie dla wyjaśniania decyzji,
- Możliwość dostrajania modeli

Wady

- Czasochłonny proces budowy systemu,
- Ryzyko utraty przejrzystości reguł ze względu na ich wzrastającą złożoność,

Zalety

- Relatywna “łatwość” budowy modelu,
- Możliwości automatyzacji procesu,
- W przypadku pretrenowanych modeli dostęp do wiedzy ogólnej wykraczającej poza zbiór uczący,

Wady

- Brak lub ograniczone wsparcie dla wyjaśniania decyzji, szczególnie w przypadku głębokich sieci neuronowych,

Dlaczego wyjaśnianie decyzji jest ważne?

Please input your username.

Ash Hull

Please input your username.

Ash Hull

This username violates our community standards.

Dlaczego?

Please input your username.

Ash Hull

This username violates our community standards.

English: OFFENSIVE PROFANITY

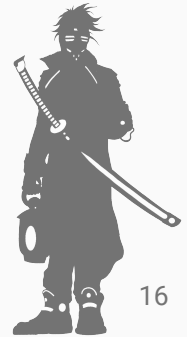
Due to similarity to “**asshole**”

Detailed output

Decomposition of the username - DigBildo - *Big D*ldo*

Categories - *Offensive, Sexual*

Toxic Element - *D*ldo*



Dlaczego wyjaśnianie decyzji jest ważne?

Przykład dla modelu do wykrywania gróźb

"ok then we will kill those black nationalists, terrorists and pedophile, but you will need to die with them"

groźba

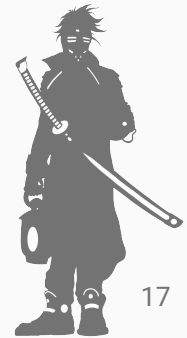
Dlaczego?

poważna groźba względem 2 osoby

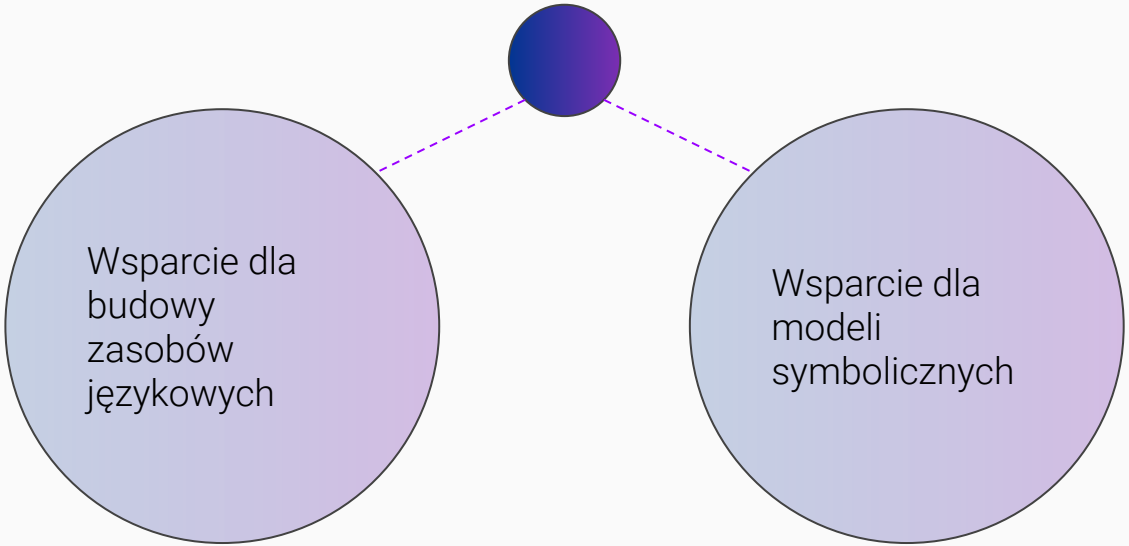
you will need to die with them

@you @future @must @threat_verbs_die_suffer @preposition @third_object_pronoun

```
[14:15] you                          you
[15:16] future                       will
[16:18] must                         need to
[18:19] threat_verbs_die_suffer      die
[19:20] preposition                  with
[20:21] third_object_pronoun         them
```

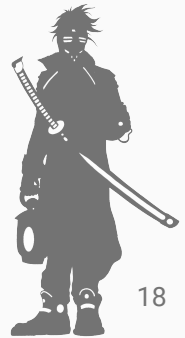


Sposoby wykorzystania metod ML

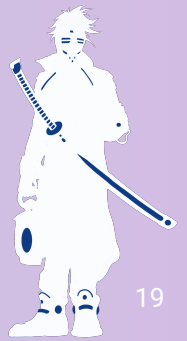


- Odkrywanie danych do anotacji;
- Rozszerzanie zasobów językowych;
- Weryfikacja ręcznej anotacji.

- Realizacja określonych zadań na potrzeby modelu symbolicznego;
- Filtrowanie wyników modelu symbolicznego.

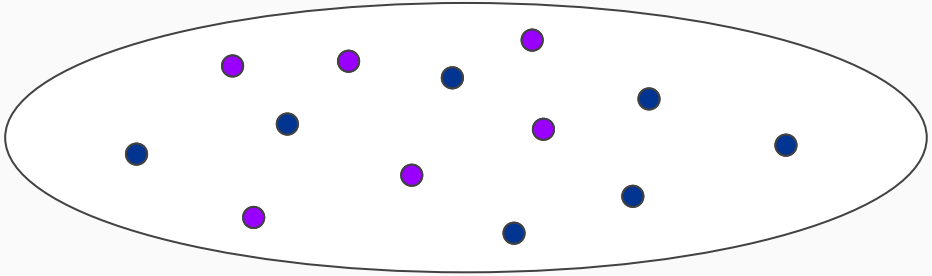


Odkrywanie danych



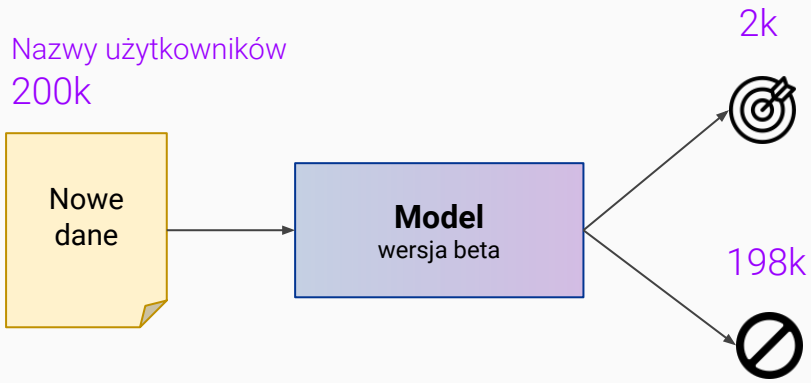
samurai LABS

Budowa zbioru referencyjnego — preferowane (idealne) podejście



- **Zaleta** — rzeczywiste odzwierciedlenie populacji
- **Wada** — problem z danymi o nierównomiernej dystrybucji klas
  - zjawiska, które chcemy wykrywać są rzadkie, np. 1 na 100/1000,
  - wydłuża to proces anotacji,
  - z czasem coraz trudniej znaleźć nowe, unikalne przykłady.

Procedura pracy z nowymi danymi



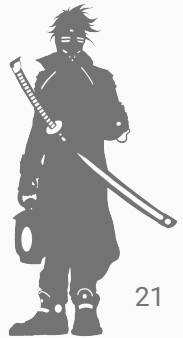
True Positives — OK

False Positives — do poprawy precyzji

True Negatives — OK

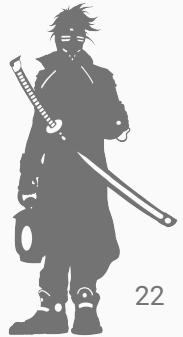
False Negatives — do poprawy kompletności

Tylko **1 na 100** nazw jest toksyczna



Wykorzystanie pretrenowanych modeli ML

- **Modele MLM** (masked-language model) uczone w sposób nienadzorowany na dużych kolekcjach danych — *BERT*, *RoBERTa*, *DistilBERT*, *BigBird*, *Muppet*, itp. potrafią uchwycić wiedzę ogólną, znaczenie sekwencji znaków (słów).
- **Docelowy model ML** jest dostrajany na danych referencyjnych dla danego zadania.
- **Przykład**:
  - *Zadanie:* wykrywanie toksycznych nazw użytkowników
  - *Zbiór referencyjny:* ~25k toxic, ~60k non-toxic,
  - *Wynik na eval*: **precyzja i kompletność powyżej 80%**,
  - *Wynik na nowym zbiorze*: **precyzja 38%**, **kompletność 80%**,
  - *Częstość toxic w próbce*: **17%** vs **1%** — 17 x więcej istotnych nowych przykładów.



Wykorzystanie pewności klasyfikacji do sortowania wyników

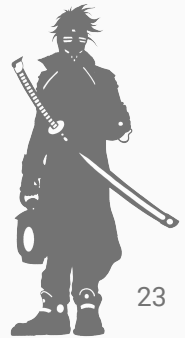
— im wyższa pewność modelu, tym wyższa precyzja klasyfikacji

Precyzja 30-40%

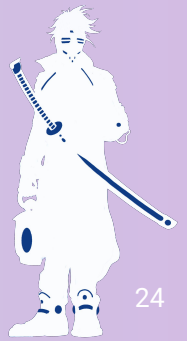
User
----------_---
-----_------_---
---_-__---_--_
--Dopamine--
--_T_T_--
--bigassboi--
--bite_me--
--wellDAM--
-4LPH40M3G4-
-B0ring_Usernam3
-BONK
-Benjamin_Dover-
-Billy_Butcher-
-Danky_Kang-
-El_Gallo_Negro-
-Fapologist-
-I-am-a-real-Aspie-
-IGottaPeeOnHer-

Precyzja 70-80% **dla top 10%**

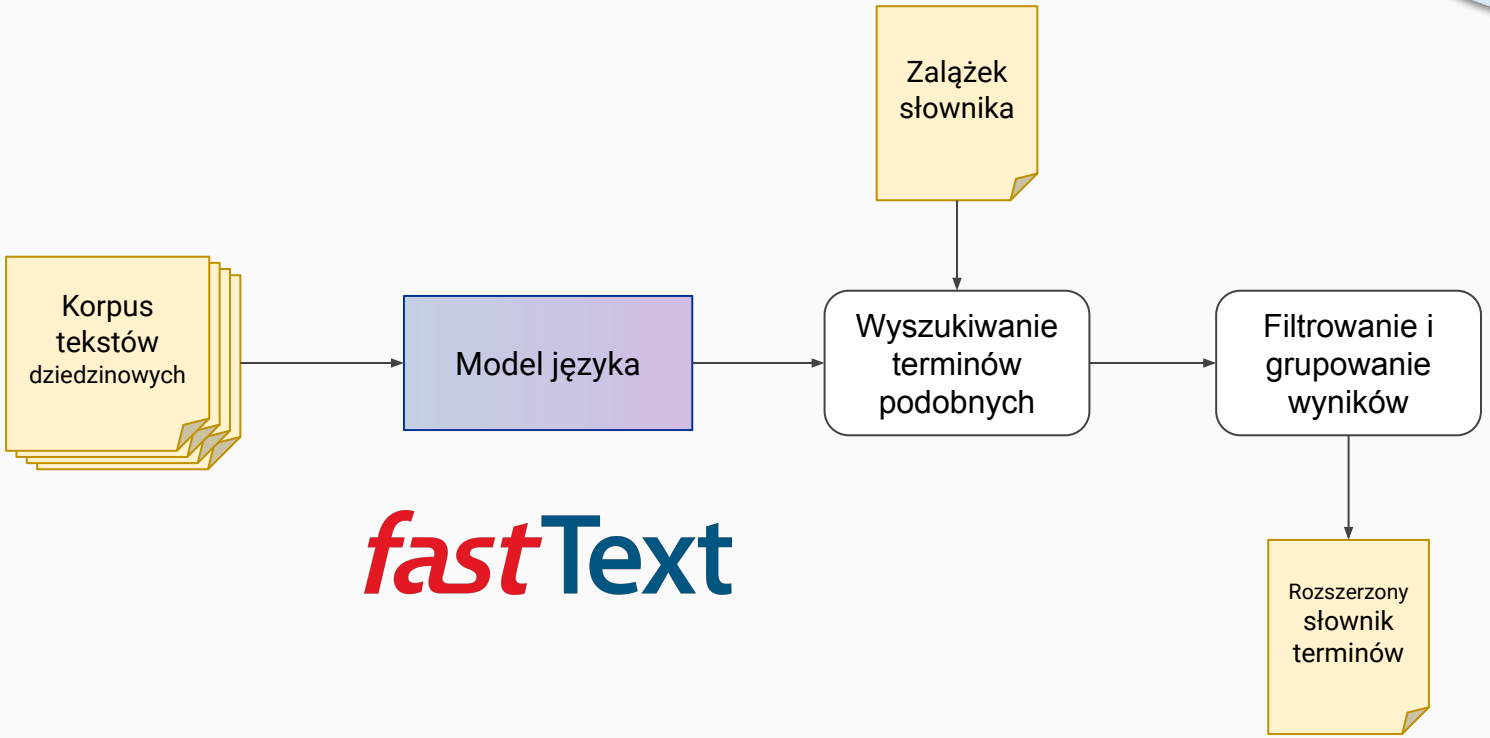
User	Waga *toxic*
ILoveAnimeTiddes	3.269531
Love_asweetbooty	3.207031
BootsySubwayAlien	3.201172
Pm_me_yourdadnaked	3.191406
bigtiddieeyelash	3.189453
I_RATE_YOUR_TITTIES_	3.189453
butdemtiddies	3.189453
BamPlsPenetrateMe	3.183594
LittleBootsy	3.179688
Pushytushy	3.175781
ScrotieMcTickler	3.175781
RandyButteSavage	3.173828
ranch_onmy_titties	3.171875
titties4uandme	3.169922
WholeLottaSEX	3.169922
TITTIES	3.166016
butsbutts	3.164062
YouretheballLickers	3.162109



Rozszerzanie zasobów językowych



Rozszerzanie zasobów językowych



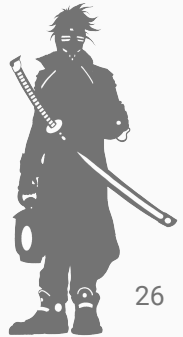
samurai LABS

Terminy ofensywne z uwzględnieniem morfologii

Liczba pojedyncza, wołacz, rodzaj męski

- babsztylu
- bałwanie
- baranie
- bęcwale
- bezmózgowcu
- bezmózgowcze
- bezmózgu
- biedaku
- bolszewiku
- bucu
- …

Top listy	Poniżej *top-u*
50.10 skurwysynie	20.83 pierdolisz
43.59 gówniarzu	19.89 skórwysynu
36.23 bydlaku	19.78 matołku
35.12 przygłupku	18.54 spierdalaj
33.95 zasrańcu	18.15 pierdolcu
33.87 wszarzu	18.04 chłystku
32.87 gnojku	17.80 skurwysynek
32.61 popaprańcu	17.64 kanciarzu
31.34 kmiocie	17.61 przygłupasie
31.15 pętaku	17.15 padalcu
31.05 skurwysynku	17.05 pojebańcu
30.91 pojebie	
29.79 wypierdku	
28.80 smarkaczu	
27.89 śmieciu	
27.33 głupolu	
27.22 kmiotku	
26.00 idioto	
25.51 cwaniaczku	
25.16 debilku	
24.19 skurczybyku	
22.97 pojebusie	



samurai LABS

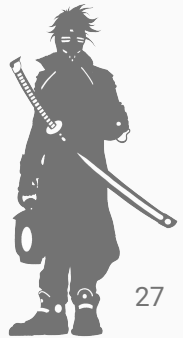
Terminy do wykrywania intencji samobójczych

kill

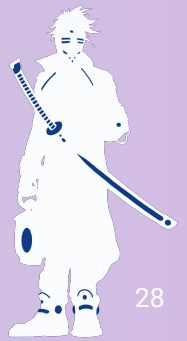
kms	0.654263
die	0.618894
asap	0.616186
commit	0.611897
iam	0.610663
killing	0.608998
guts	0.607888
offing	0.606651
punish	0.604843
ima	0.603541

kms

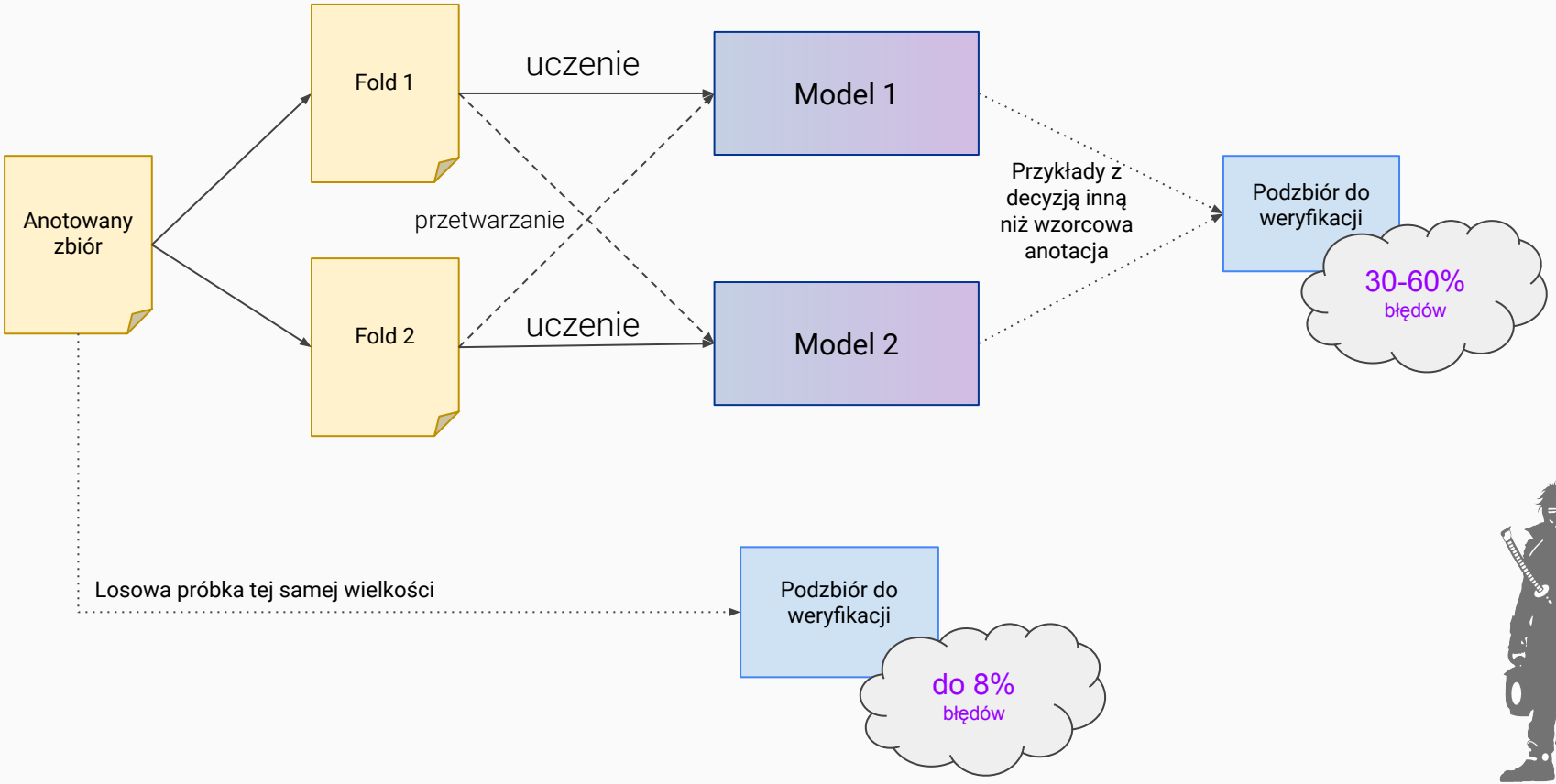
1. cant wait to **kms** on heroin
2. i already tried to **kms** once when i was like but failed
3. i tried to **kms** once but failed
4. im about to turn and that will most likely be the day that i will **kms** th april
5. i set a date to **kms**



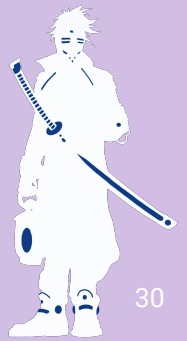
Weryfikacja anotacji



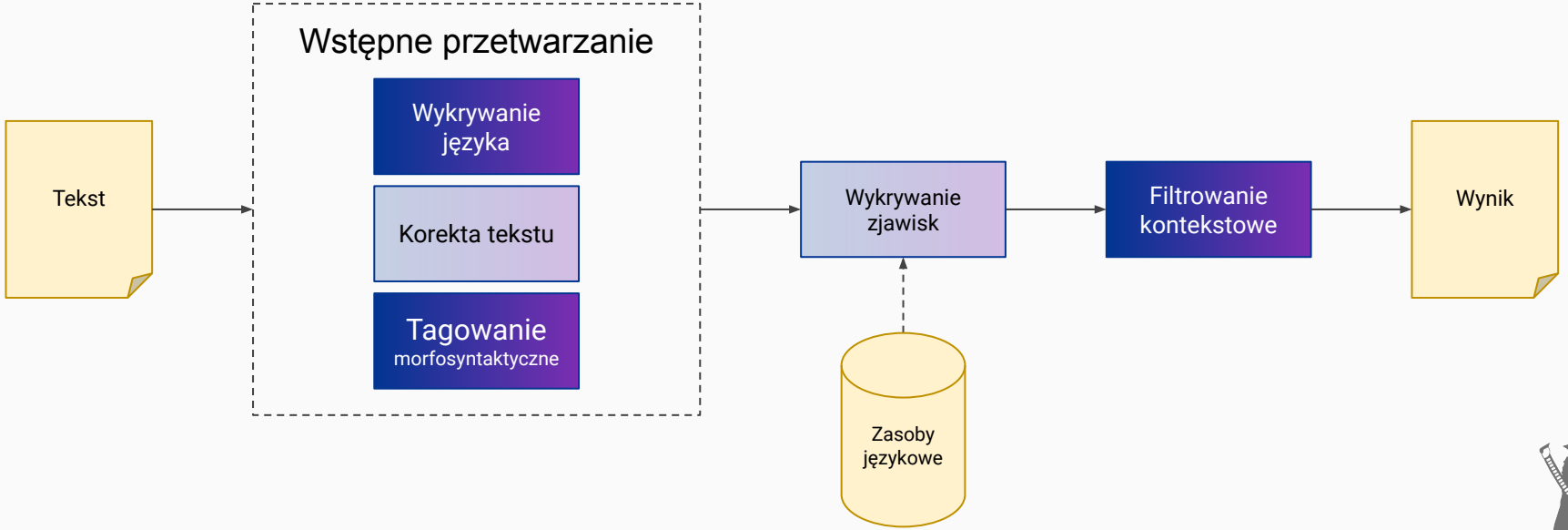
Weryfikacja anotacji



Realizacja zadań pomocniczych



Potok przetwarzania



Moduł ML

Moduł symboliczny

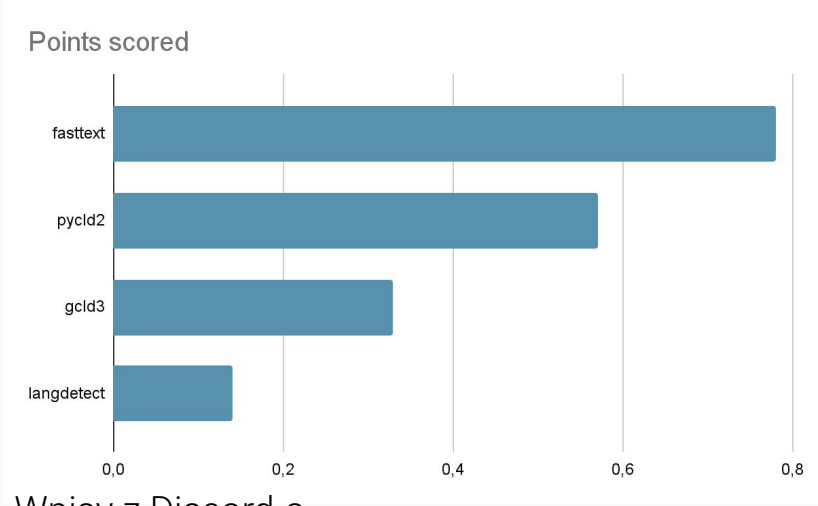
samurai LABS

Narzędzia

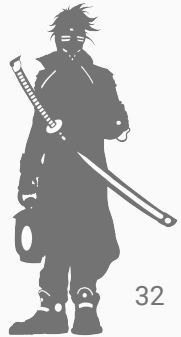
- fasttext
- langid
- langdetect
- spacy-langdetect
- pycld3
- gcld3

	fasttext	langid	langdetect	spacy-langdetect
speed	129ms	8.07s	1min 4s	2min 2s
accuracy	0.860	0.827	0.845	0.845

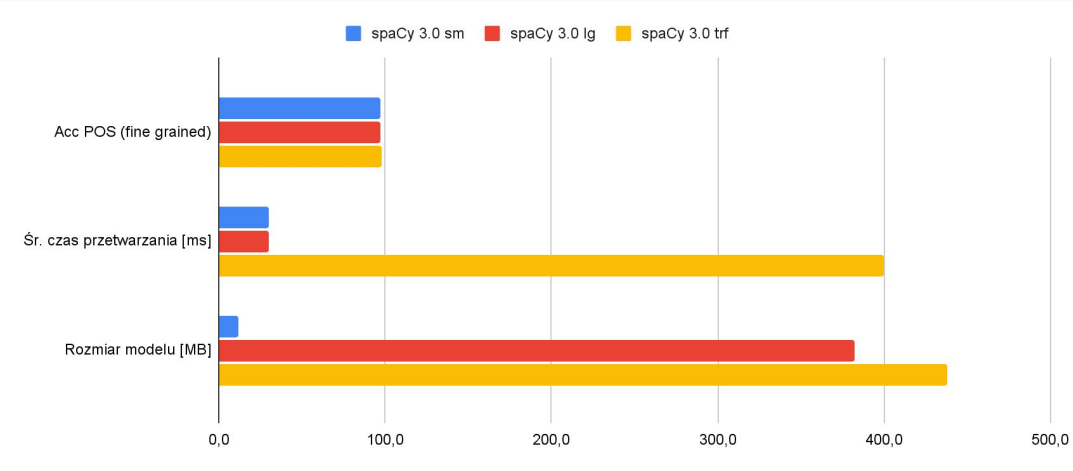
Tweet set (https://www.kaggle.com/rtatman/the-umass-global-english-on-twitter-dataset)



Wpisy z Discord-a

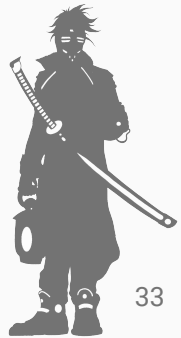


- **spaCy 3.0** dla j. angielskiego (https://github.com/explosion/spaCy)



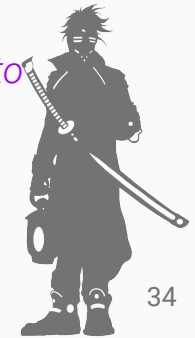
- **KRNNT** dla j. polskiego (https://github.com/kwrobel-nlp/krnnt)

  KFTT (https://github.com/kwrobel-nlp/kftt) — 97% vs. 93% (KRNNT)

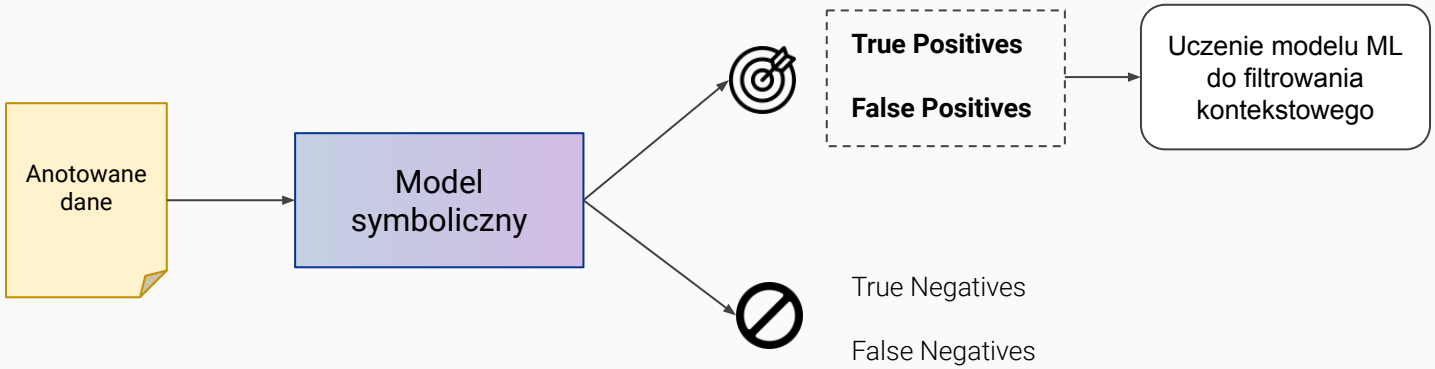


- *"I am 32 and my gf is 13 (...)"*
  - pedofilia?
  - *"(...) Can we still do quests together or are our levels too far apart? Sorry kind of new to WoW."*
  - Źródło — forum gry World of Warcraft

- "I tried to kill myself and I couldn't. (...)"
  - próba popełnienia samobójstwa?
  - "(...) *I also noticed I had no effect on objectives. It was so weird. The next match I was back to normal."*
  - Źródło — forum gry Call of Duty: Modern Warfare



Filtrowanie kontekstowe cd.



Dziękuję za uwagę

samurai LABS

Customers

Khoros · INACH · webex by cisco · HEROO MOBILE

Partners

NVIDIA · FAIR PLAY ALLIANCE · PUBLIC · FACTMATA

Press

Forbes · BBC · BUSINESS INSIDER

Technology Awards

CBINSIGHTS · Gartner

Neuro-symboliczna sztuczna inteligencja w walce z cyberprzemocą w Internecie

dr inż. Michał Marcińczuk (Samurai Labs)