SCIENCE
NASK

# BAN-PL

A Polish Dataset of Banned Harmful and Offensive Content from Wykop.pl web service

# Datasets

## Source

- **Twitter**
  (Waseem and Hovy 2016; Davidson et al. 2017; Founta et al. 2018; Zampieri et al. 2019)
- Wikipedia talk page
  (Wulczyn et al. 2017)
- Facebook and Youtube
  (Hammer 2017; Salminen et al. 2018)
- Reddit
  (Caselli etl a. 2020)

## Labels and coverage

- varied labels and modes of annotation, such as hate speech, offensiveness, aggression, racism, sexism, and toxicity;
- distinct targets of attacks, including personal attacks, attacks on women, and attacks on migrants;
- differing numbers of classes, ranging from binary to multi-label classification;
- varying degrees of class balance, where, in many cases, the neutral class significantly outweighs the harmful class or classes

# Polish datasets

**hate_speech_pl**
(Troszyński et al. 2017)

- 12 676 samples
- data source: online forums (2012)
- scope: hate speech
- annotation: sentiment, irony/sarcasm, call for action

**KLEJ CBD Task**
(Ptaszynski et al. 2019)

- 11 041 samples
- 8.9% harmful vs. 91.1% non-harmful
- data source: Twitter
- scope: cyberbullying and hate speech

**Deepsense.ai**
(Szmyd et al. 2023)

- 1 000 samples
- 10.6% harmful vs. 89.4% non-harmful
- data source: Twitter
- scope: cyberbullying and hate speech

SCIENCE
NASK

# BAN-PL dataset

Collaboration with **Wykop.pl** web service

Content **banned** and labeled by **professional moderators**

**Equal classes** of harmful and neutral content

SCIENCE
NASK

# BAN-PL dataset



Collaboration with **Wykop.pl** web service

Content **banned** and labeled by **professional moderators**

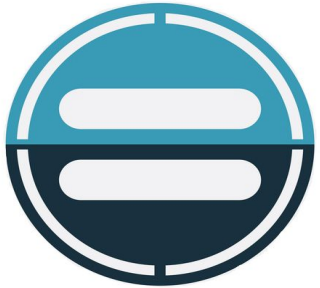
**Equal classes** of harmful and neutral content

# BAN-PL dataset



Collaboration with **Wykop.pl** web service

Content **banned** and labeled by **professional moderators**



**Equal** classes of **harmful and neutral** content

# Wykop.pl web service

## Platform

- launched in **2005** as a digg.com equivalent and often referred to as the **"Polish Reddit"**
- app. **30,000** new **posts and comments daily** (wirtualnemedia.pl 2023)
- **wide range of content**, e.g. sports, economy, political affairs, and viral Internet phenomena

## Features

- based on **user-generated content**, allowing users to submit and share news stories, articles, and other forms of media
- **voting system** enables users to participate in the curation of content, either by "digging" or "burying" it

## Users

- one of the top 10 most popular SNSs in Poland, with app. **3 million users**
- the user base is predominantly in the **18-45 age group**, with a significant contribution from **15-24 year olds**
- a higher proportion of **male users**

# Moderation scheme

- the online content is under constant monitoring by trained moderators from the user community
- additionally, every user can independently flag any piece of content
- internal taxonomy of **21 ban reasons (web service's policy violations)**
- app. 2–5% of the content is being reported→ **30 000–60 000 submissions per month**
- **14 professionally trained moderators** recruited from the Wykop.pl community
- **5 independent votes** → majority voting
  (the consensus is typically represented by **5:0** and **4:1** vote split)
  - app. 15% – „advertising content, spam"
  - app. 10% – „inappropriate content"
  - app. 7% – „incorrect tags"
  - app. 49% – "incorrect reports"

SCIENCE
NASK

# Data collection process

The person is irritating (flood, hinders the use of the service)

This is a multi-account

Spam account

Unauthorised avatar, description or profile data

Impersonates me or a friend of mine

**Attacks me**

Manipulates content or votes

Pornographic content

Fraudulent, misleading content, false information

**Propagation of hatred or violence, drastic content**

**Breach of regulations – inappropriate content**

Violation of personal rights

Advertising content, spam

Missing 18+

Incorrect tags

Manipulation of votes

Advertising content, spam, flood

**Attacks others**

Duplicate

**Attacks me or violates my personal rights**

Other

SCIENCE
NASK

# Data collection process

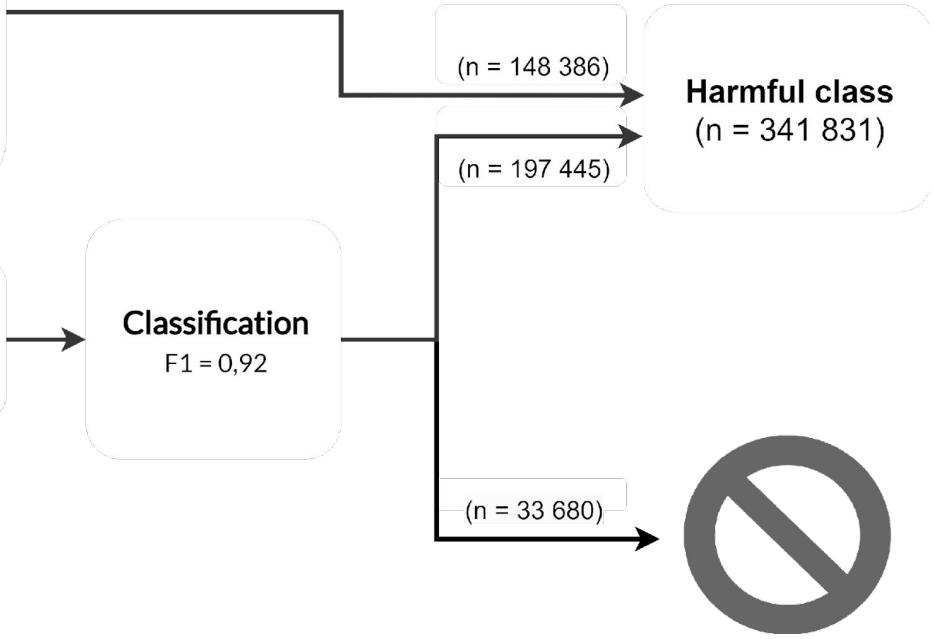
Attacks me

Attacks me or violates my personal rights

Attacks others

Propagation of hatered or violence, drastic content

Rules violation – inappropriate content




SCIENCE
NASK

# Data anonymization

**Stage 1.:** regex → ***username***, ***URL***

**Stage 2.**: regex → ***phonenumber***, ***mail***, ***number*** [PrivMasker]

**Stage 3.:** NER → ***surname***, ***address*** [PolDeepNer2]

**Stage 4.:** dictionaries of historical/fictional figures and pseudonyms → ~~***surname***~~, ***pseudonym***

**Stage 5.:** double manual review of **24,000 samples** by 7 annotators and 2 super-annotators (linguists with experience in annotating social media data)

# Data anonymization

| Original sample after anonymization (in Polish) | English translation |
|---|---|
| będziesz pan wisiał, panie [surname] | you will be hanging, mr [surname] |
| [surname] to glupia c**a i nie nalezy sie jej ten nobel | [surname] is a stupid c**t and she doesn't deserve this nobel prize |
| {USERNAME}: pedała ci żal że się nad ty[m] [pseudonym] użalasz? | {USERNAME}: do you pitty a f****t since you feel sorry for this [pseudonym]? |
| #[pseudonym] k***o pedophilu. Nie jesteś streamerem. Nie liczysz się w świecie YouTube. Płace 5.000 zł za adres tej k***y #patostreamy | #[pseudonym] you wh**e pedophile. You're no streamer. You don't count in the YouTube world. I'm paying 5,000 PLN for this wh***'s address #patostreamy |
| Do tego świra-pedalarza w lateksach z [address] w Poznaniu - wyk*****j cwelu i sam patrz do tyłu. Sorry za zj***nie średniej idioto. #rower #pedalarze | To this kook-pedal pusher in latexes from [address] in Poznan - f**k off w****r and look back yourself. Sorry to f*** up your average idiot. #bike #pedalpushers |
| {URL} c**j Wam na stale, śmiecie z jutuba, czyli urywek ostatniego streamu! #[pseudonym] #[pseudonym] #patostreamy #[pseudonym] #patostreamy | {URL} f**k you forever, jutube trash, that is a snippet from the last stream! #[pseudonym] #[pseudonym] #patostreamy #[pseudonym] #patostreamy |
| [phonenumber] napalona agatka lat 16 lubi BDSM | [phonenumber] horny agatka aged 16 likes BDSM |

This table contains examples of hate speech and vulgar personal attacks. The authors do not support the use of harmful language, nor any of the harmful representations quoted above

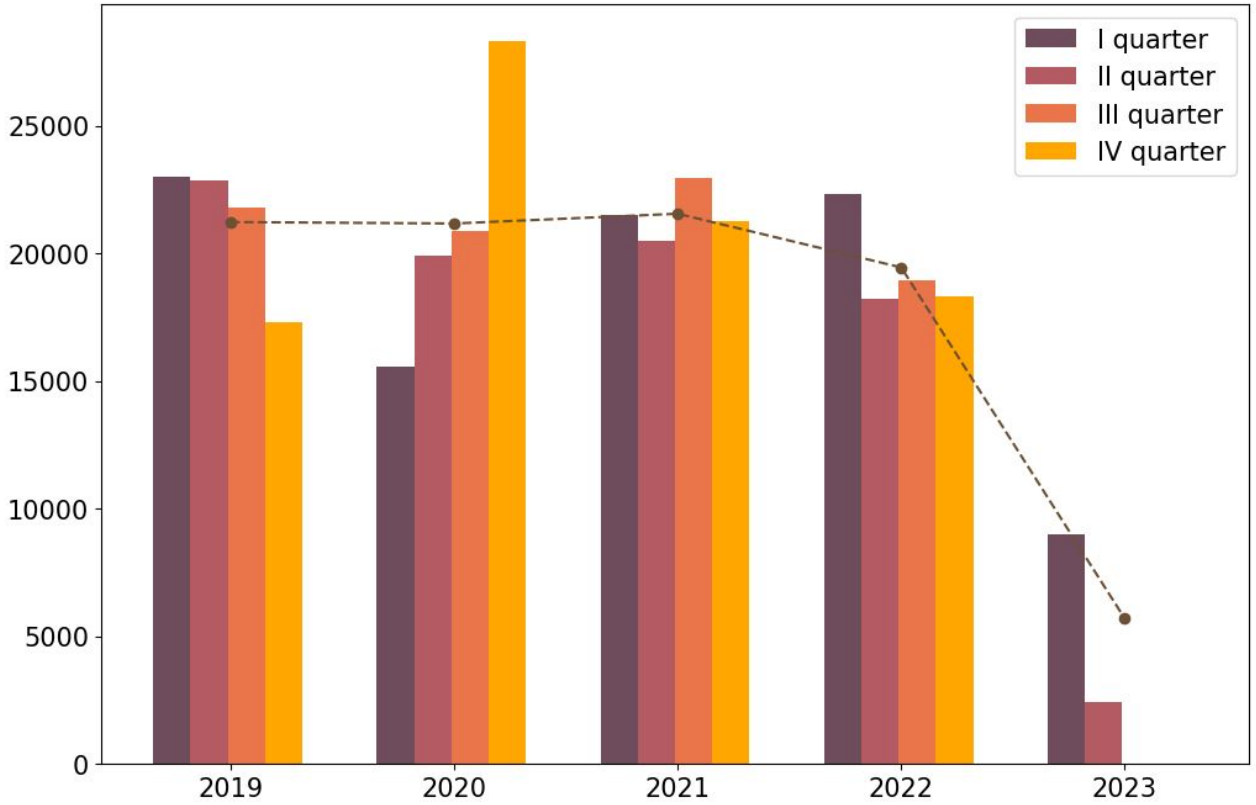
SCIENCE
NASK

# Number of tokens by class

| | N | M | SD | Q1 | Mdn | Q3 |
|---|---|---|---|---|---|---|
| Harmful total (n = 345,831) | 12,064,249 | 34.88 | 98.48 | 10 | 17 | 35 |
| Harmful flagged (n = 148,386) | 7,392,680 | 37.44 | 118.05 | 10 | 18 | 35 |
| Harmful predicted (n = 197,445) | 4,671,569 | 31.48 | 63.55 | 11 | 18 | 34 |
| Neutral (n = 345,831) | 14,143,837 | 40.90 | 55.87 | 13 | 24 | 47 |

SCIENCE
NASK

# Harmful content by quarters




SCIENCE
NASK

# Topical coverage

# Folksonomy


#bekazlewactwa
#bekazpisu
#gry
#spierdolenie
#starebabyjebacpradem
#highleague
#zamach
#muzyka
#patostreamy
#pytanie
#divyzwykopem
#humorobrazkowy
#covid19
#wykop
#protest
#polityka
#afera
#wosp
#suchodolski
#sejm
#rosja
#jp2gmd
#bekazpodludzi
#fut
#gielda
#oswiadczenie
#warszawa
#przegryw
#papiez
#aborcja
#lgbt
#bekazkatoli
#seks
#gdansk
#polska
#loveisland
#gorzkiezale
#wroclaw
#leagueoflegends
#moderacjacontent
#4konserwy
#takaprawda
#pdk
#kanalsportowy
#youtubeshore
#mecz
#bigbrother
#kiciochpyta
#moderacja
#sylwesterzwykopem
#depresja
#tinder
#jebacpis
#ukraina
#skoki
#tfwnogf
#rozowepaski
#pilkanozna
#przemyslenia
#stulejacontent
#famemma
#gownowpis
#wykopobrazapapieza
#marszniepodleglosci
#bialorus
#bekazprawakow
#mikrokoksy
#narkotykizawszespoko
#pis
#zwiazki
#heheszki
#p0lka
#niebieskiepaski
#logikarozowychpaskow
#usa
#wojna
#policja
#blackpill
#antykapitalizm
#ksw
#zalesie
#patologiazewsi
#tvpis
#neuropa
#f1
#pasta
#wislakrakow
#pracbaza
#czarnyhumor
#krakow
#cyberpunk2077
#cenzopapa
#primemma
#nsfw
#wybory
#2137
#koronawirus
#konfederacja
#czarnolisto
#logikaniebieskichpaskow

# Linguistic features

## Word formation

- borrowings from English social media language, e.g. *simp*, *blackpill*, *kukold (cuckold)*
- Polish-rooted equivalents to English vocabulary, e.g. *przegryw* (an equivalent to 'incel')
- original Polish-rooted expressions, e.g. *beciak*, *normik*
- creative eponyms, e.g. *Mirek*, *Chad*, *Julka*, *Seba*, *Oskar*, *Alvaro*, *Carlos*

## Spelling and use of graphical characters

- intentionally crafted grammatical and syntactical errors, e.g. *chłop -> huop*
- emojis and textual emoticons, incl. lenny faces, e.g. *¯_(ツ)_/¯*

## Obfuscation strategies

- character substitution, e.g. *flower -> fl0w3r*
- using a string of identical symbols, e.g. *flower -> flow***
- using random symbols, e.g. *flower -> fl*%#$$r*
- phonetic spelling, e.g. *duck -> duq*
- extra character insertion or deletion, e.g. *coffee -> coff&ee*
- word splitting or merging, e.g. *coffee -> c o f f e e*

# Preliminary experiments

| Model | Recall | Precision | F1 |
|---|---|---|---|
| **RoBERTa base v2** | 0.84 | 0.83 | 0.83 |
| **TrelBERT** | 0.76 | 0.82 | 0.79 |
| **Polbert-CB** | 0.83 | 0.78 | 0.81 |
| **HerBERT-HS** | 0.79 | 0.80 | 0.79 |

The results of the preliminary experiments on the publicly available BAN-PL dataset (24,000 samples)

# Discussion

## post-moderation bias

- data collection process
- harmful data is usually obtained from publicly available social media content (most notably Twitter) through APIs
- obtained data might have already undergone some initial moderation resulting in limited number of explicit hateful posts.

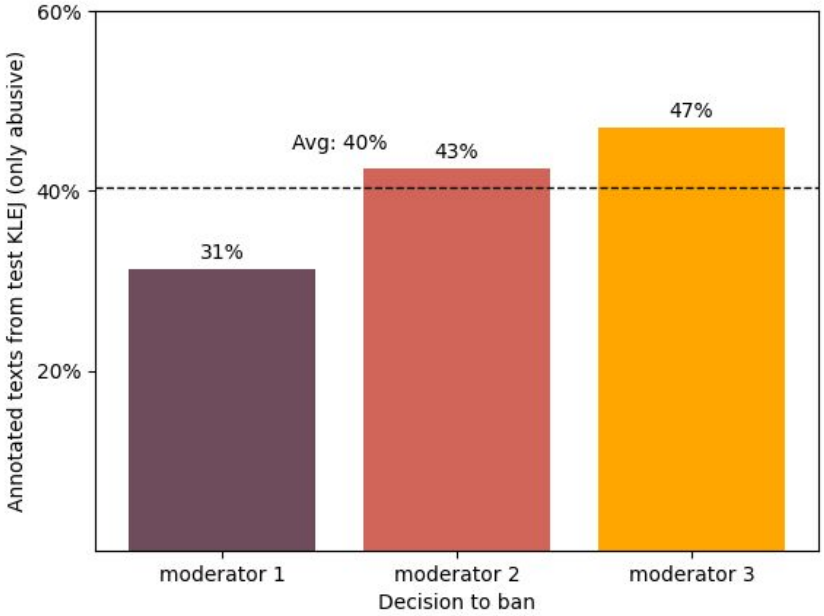
## pre-selection bias

- data collection process typically involves pre-selection of users, key words or hashtags
- this approaches undermines the possibility of creating a genuinely representative corpus of abusive content by collecting relatively homogeneous data centered around specific topics or targets of hate (Ludwig et al. 2022)

## annotation bias

- covers a range of challenges resulting from varying annotation guidelines, which are rooted in the lack of consensus on definitions of harmful content and subjective notions of what constitutes hate speech

# Re-annotation task




Percentage of samples from the KLEJ test set (Cyberbullying detection) labeled as offensive by three Wykop.pl moderators

# Contributions

- introduction of a **new dataset** for offensive language detection in the Polish language comprised of **content banned by professional moderators**
- analysis of **significant linguistic features** of the content posted on Wykop.pl
- **identification of biases** that the dataset avoids and those that still persist and discussion of potential strategies for addressing them
- publication of the **anonymized open available subset of the dataset** along with preprocessing scripts that can be readily applied in real-life scenarios

# Future work

- developing **further preprocessing** for text **normalization and profanity unmasking**
- releasing a **subcorpus of manually annotated hate speech**
- providing a **dictionary of hate speech and offensive language**
- fine-tuning NER and **anonymizing of the whole BAN-PL dataset**, using a refined automatic pipeline and involving decreasing manual correction

Anna Kołos
Inez Okulska
Kinga Głąbińska
Agnieszka Karlińska
Emilia Wiśnios
Paweł Ellerik
Andrzej Prałat

ziliat@nask.pl

BAN-PL dataset available at:
https://github.com/ZILiAT-NASK

/NASKpl
/NASK_pl
/nask
/nask.pl
/NASKPIB
nask.pl

SCIENCE
NASK