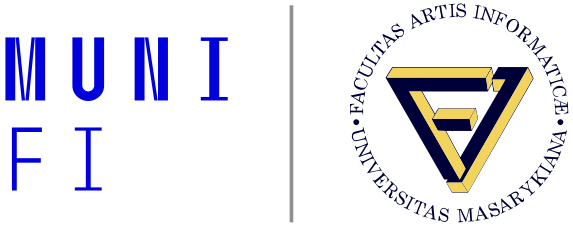


Building a Corpus for Personal Data Detection

For Social Messaging

Ondřej Sotolář
xsotolar@fi.muni.cz

Faculty of Informatics, Masaryk University

May 18, 2021

Overview

Motivation

- Ethical
  - Open Science:
    - transparency, reproducibility, and reusability,
    - sharing data between institutions.
  - Principles: purpose and storage limitation, confidentiality, data minimisation.
- Legal
  - GDPR:
    - Terms: personal data, identifiable person, anonymous data,
    - *Recital 26*: applies to any information concerning an identifiable person by **reasonably likely means**.
  - HIPAA:
    - Terms: sensitive data, de-identification, sanitization.
- Project FUTURE [1]:
  - comparing self-reports to real-time metrics including texts.

Privacy-Utility balance (1/3)

Utility

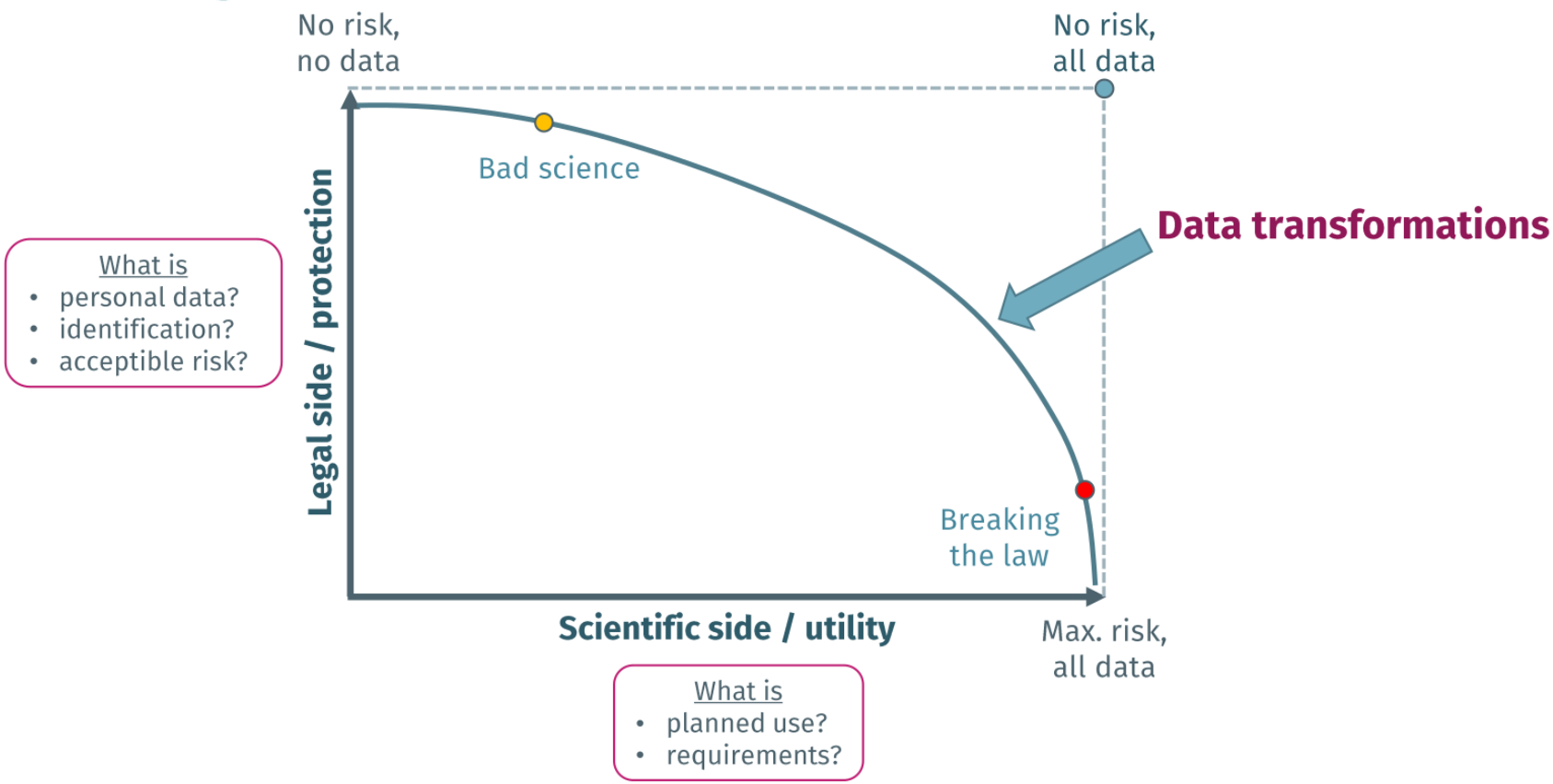


Figure: Balance between utility and privacy [1]

[1] Barth-Jones, Brussels Privacy Symposium [2]

Privacy-Utility balance (2/3)

Disclosure types

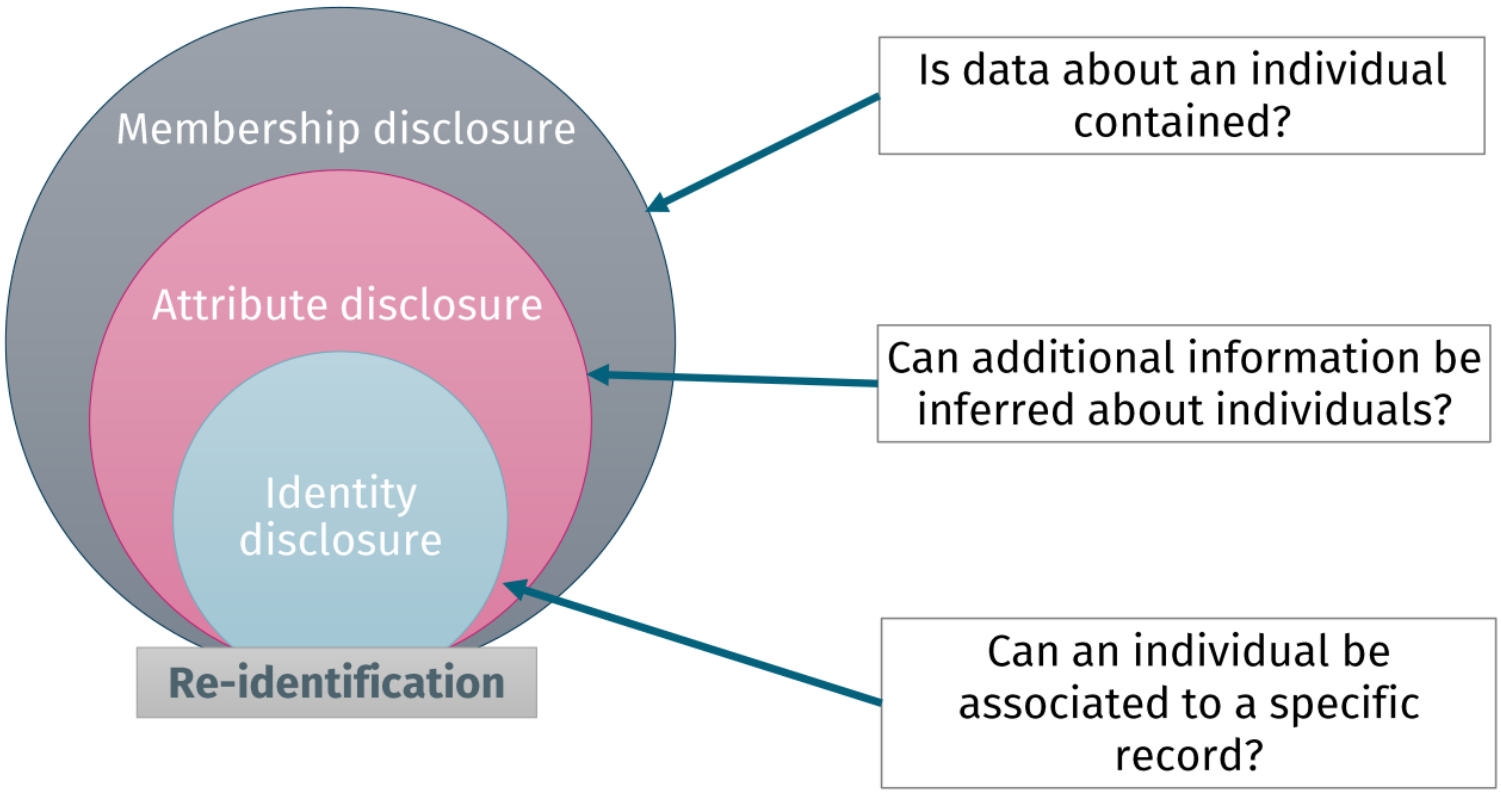


Figure: Types of attacks [2]

[2] Fabian Praßer [3]

Privacy-Utility balance (3/3)

Re-identification types

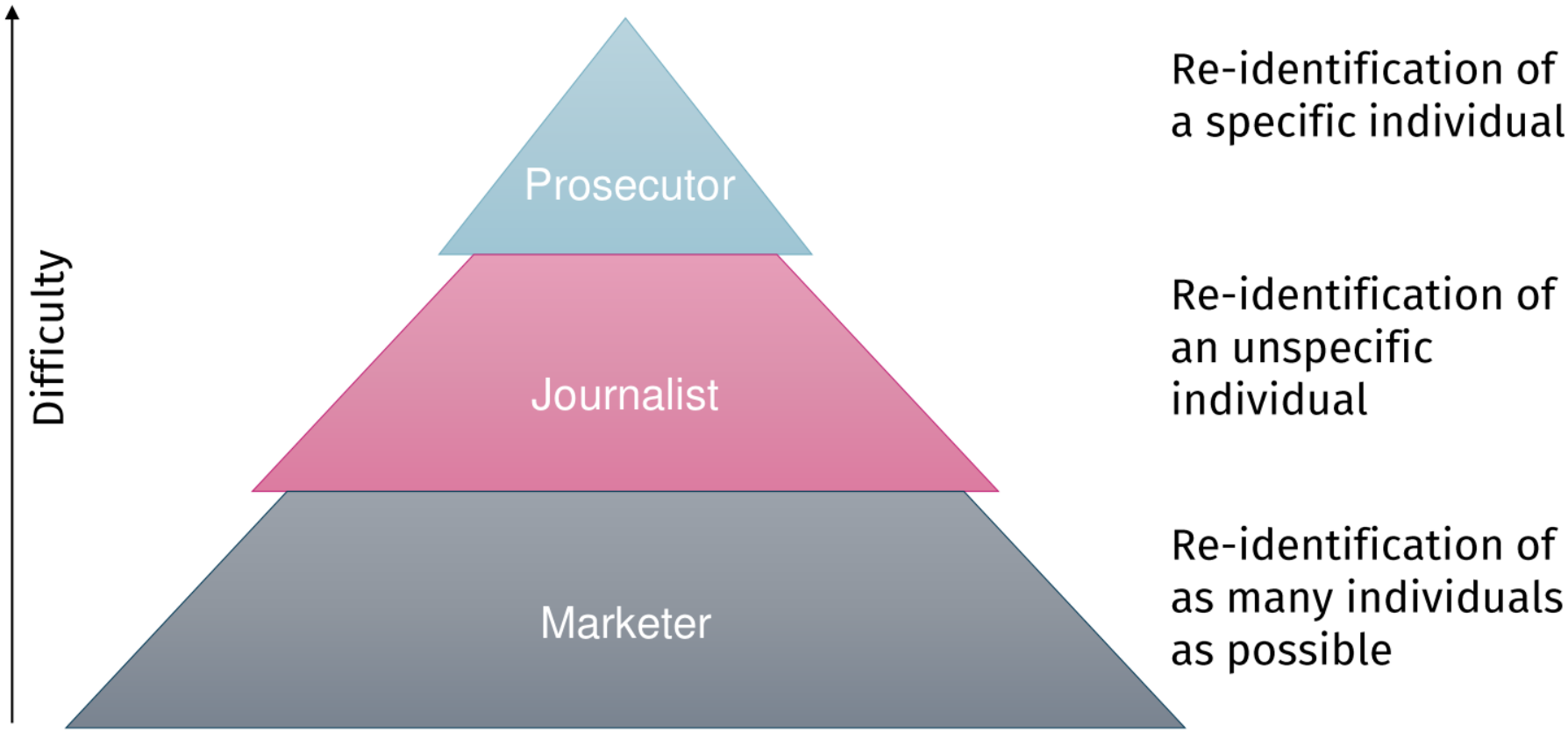


Figure: Re-identification types [3]

[3] Fabian Praßer [3]

Related Work

- Other types of data:
  - Terms: *Identifier, Quasi-identifier*,
  - Models: *k*-anonymity[4], ε-differential privacy [5], and many derivative models.
- Text:
  - text-2-text:
    - Named Entity Recognition (NER), [6, 7, 8]
    - semantical relatedness (C-Sanitized[9] and derivative models[10, 11]).
  - Noisy representation:
    - ε-differential noise [12, 13],
    - TensorFlow Privacy.
  - Measuring utility-privacy balance:
    - masking, generalizations [14, 15, 16],
    - side-by-side privacy evaluation [17].

Target Domain: Social Messaging

Language Style:

- informal, composed of short messages (avg. line length is 30 chars),
- simple-structure sentences or fragments,
- capital letters are used less or inirregular places,
- typos are common,
- words and phrases from different languages (mostly English and Slovak),
- nicknames, both related to a personal name and not.

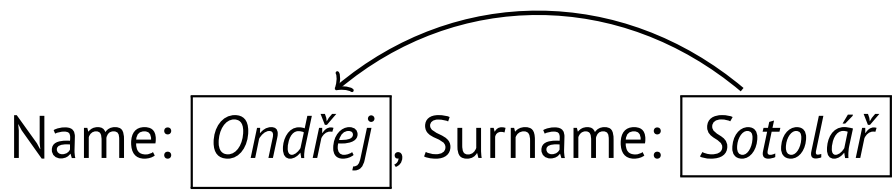
Table: The importance of case, diacritics, and typos in CNEC for NER.

Solution	Precision	Recall	F1
CNEC	71.36	68.25	69.44
CNEC lowercase, no diacritics	62.35	57.96	59.83
above + 1 character typo	60.02	54.0	54.38

Old Annotation Schema

- *Name:* any name that contains a surname, or a first name belonging to a surname in close proximity.
- *ID:* birth number, tax identification, bank account, credit card, ID card, passport, driver's license number, etc.
- *Location:* street name with house number, GPS coordinates, IP address.
- *Contact Information:* email, phone number, URL, etc.

Identifier + quasi-identifiers = Personal Data Entity (PDE)



Old Annotation Tool Detection Efficiency

Solution	Precision	Recall	F1
NER	8.05	71.36	14.46
above + rules	8.78	**86.5**	15.94
above + PDE composition	**29.92**	78.7	**43.36**

Old Annotation Tool Masking

Type	Original value	Replacing value (gazetteer)
Name	Ondřeje Sotoláře	Tomáše Klusáka
Address	potkáme se v Praze	potkáme se v Zádveřicích

Type	Original value	Replacing value (generalize)
Email	xsotolar@fi.muni.cz	dfadf6541@mail.cz
Car plate	1B2 9806	2D6 2398

Old Annotation Tool Utility

Measure change on downstream tasks:

- Named Entity Recognition

Solution	**Precision**	**Recall**	**F1**
Original text	100	100	100
Suppression [14]	80.21	58.19	67.45
Type tags [15, 16]	94.29	68.4	79.28
Our solution	**95.63**	**88.47**	**91.91**

- Classification: supervised classification task (label Risky Behavior)

  -> **No Difference**:

  - three-layer feed-forward neural network with an embedding layer on top,
  - Deep Open Classification model [18] with FastText embeddings (web-crawl).

Annotation schema (1/5)

Superclass	Subclass	Tag
Osobní jména	křestní jména, přezdívky založené na nich	pf
	příjmení, i zkomolené	ps
	celé jméno (víceslovné)	P
	data narození, úmrtí, svatby: pouze ke konkrétní osobě	pb
	ostatní jména, přezdívky ne-odvozené od jména	p_
Kontaktní údaje	tel, email, URL, jméno YT kanálu, nick ve hře	cnt
Identifikátory	čísla dokladů, karet, smluv	idf
Lokace	adresy, GPS, ulice, města, země	loc
Nemoc	nemoci	dis
Organizace	Instituce, firma, škola, politická strana	i_
Klasifikováno do špatné třídy	**neměnit rozsah!**	mis
Není entita	**neměnit rozsah!**	not
Neznámá třída	**neměnit rozsah!**	unk

Figure: Annotation Schema Overview

Annotation schema (2/5)

Osobní jména: celá jména by měla být označena pouze jednou značkou, kromě případů, kdy jsou oddělena jinými slovy. Přezdívky, například „Ritchie“ pro Richarda nebo „Tom“ pro Tomáše, by měly být označeny. Přezdívky nezaložené na jménu pouze takové, kterými lze oslovovat.

- jména historických, fiktivních postav, známých osobností
  - musí se vztahovat ke konkrétní osobě
  - musí být odvozená od skutečných jmen osob
- přezdívky (p_)
  - ustálené unikátní oslovení konkrétní osoby ne-odvozené od jména
  - neoznačujeme oslovení (zlato, miláčku...)
- jména zvířat
  - lidská jména stejně jako by šlo o člověka
  - běžná jména domácích mazlíčků neoznačujeme
  - konkrétní, unikátní, zvíře, např. kůň, který vyhrál velkou cenu, označíme “p_”
- přivlastňovací označovat běžně

Anotace	Třída
Jméno: Ondřej, Příjmení: Sotolář	pf, ps
Mam jeden navrh od Michala dedka na ...	P
Půjde s námi i tomík a godar?	pf, p_
Karel IV, narozen 4.10.1289 byl králem.	P, pb
Sněhurka a 7 trpaslíků, Sherlock Holmes	–,P
miláčku, zlato, bobíku, kámo, jarunko, Smíšek, micka, Argael	–, –, –, –, pf, p_, –, p_
Jardovo kolo	pf

Figure: Annotation Schema Detail (1/4)

Annotation schema (3/5)

Kontaktní údaje: e-mail, telefonní číslo, url apod. Měly by být označeny pouze jako celek. Přezdívky takové, kterými se neoslovuje.

Anotace	Třída
Tel na mě je +420 777 236 598 a na bráchu 608956659	cnt,cnt
Mail na kolegu je martin.novak@firma.cz	cnt
Můj Skype nick je hunter1235	cnt

Identifikátory: rodné číslo; čísla kreditních karet, občanských průkazů, pasů a řidičských průkazů, SPZ. Pokud jsou vícedílné, měly by být označeny jako jeden celek.

Anotace	Třída
Posílám to číslo občanky: 5051648452	idf
Číslo karty je 6548 1238 4569 7894	idf
... je Američan, jeho social security number je 123-65-7894	idf

Figure: Annotation Schema Detail (2/4)

Annotation schema (4/5)

Lokace: název ulice s číslem domu, IP adresa, GPS pozice. Adresy by měly být označeny jako celek jednou značkou (na jednom řádku), pokud nejsou rozděleny slovy. I zkomolené.

- jméno označovat zvlášť

Anotace	Třída
Jan Novák Bezručova 180, Malenovice, 763 02 Zlín	P, loc
Vyzvedni mě na táborské na kolejích tvrdkách	loc, loc
Brno je České město, metropole Moravy, části České Republiky	loc,loc,loc,loc

Nemoci: pojmené nemoci skutečné i nejisté, ale ne pouze příznaky, pocity nebo nálady.

Anotace	Třída
dostala jsem covid a je mi blbě	dis
ze své anorexie mám už fakt depku	dis
jsem fakt hrozně nachlazený, myslíš že můžu mít chřipku?	dis

Figure: Annotation Schema Detail (3/4)

Annotation schema (5/5)

Organizace: konkrétní instituce, školy, firmy, kapely, události apod.

- ne afiliace

Anotace	Třída
pracuje na úřadě práce, chodím na gympl v Brně	–, loc
Hustopečské hody, jdeš na majáles?	i_, -
jsem student na gymnáziu juraja hronca	i_

Afiliace: příslušnost ke konkrétní skupině, etnicitě, klubu označujeme jako “unk”

- ne obecné pohlaví, sexualita apod.

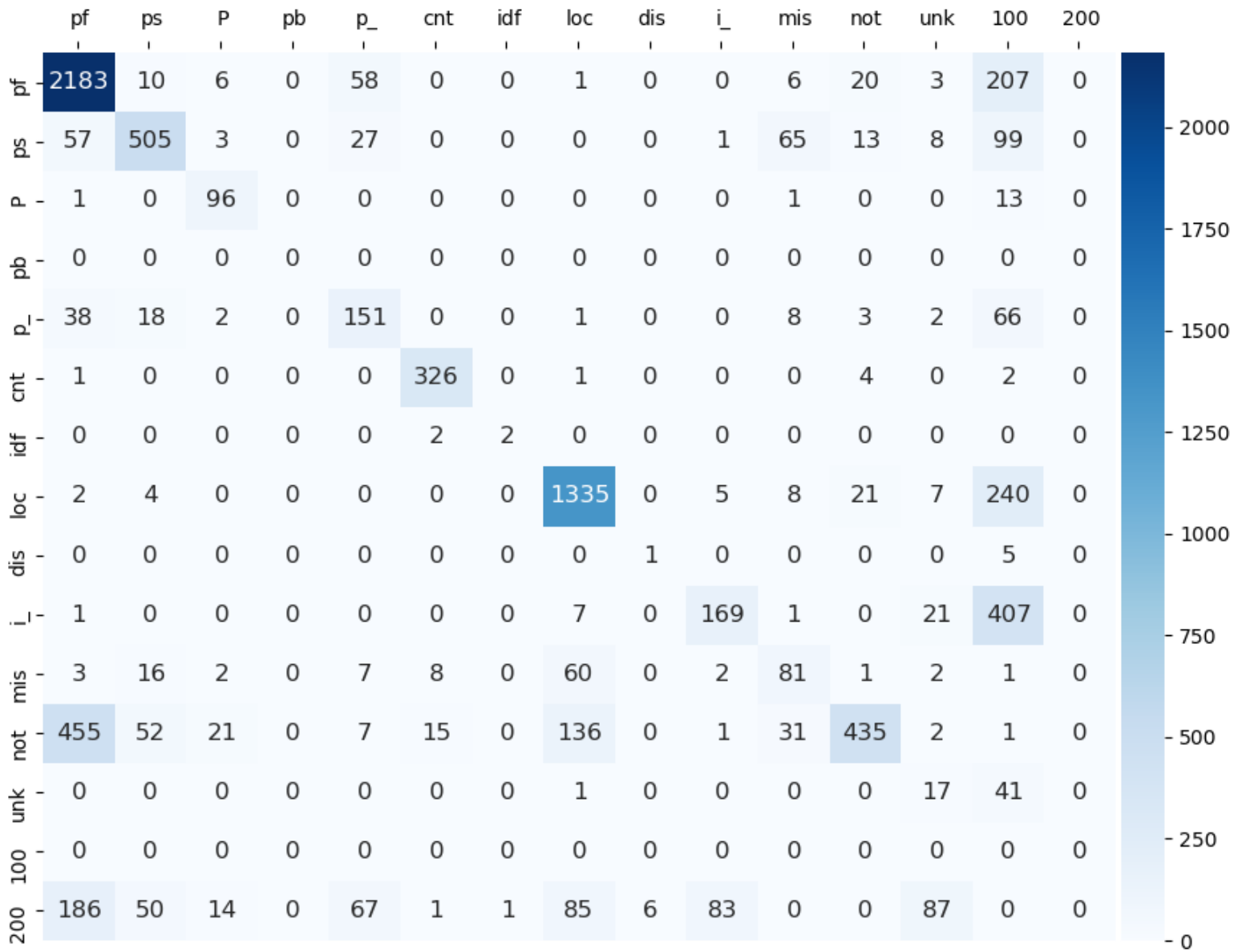
Anotace	Třída
čech, rom, černoch, cigán, vietnamec, američan	unk,unk,unk,unk,unk, unk
muž, homosexuál, gay, bisexuál, slepec, manžel, metalista	-
sparťan, fotbalista, komunista, volič pirátů, pirátský předseda	unk, –, –, i_, unk

Figure: Annotation Schema Detail (4/4)

Annotation Process

1. Text authors labeled personal data with the old annotation schema:
   1.1 anonymize with the existing tool.
2. Create a batch of approx. 20K rows (avg. 30 chars per row):
   2.1 pre-label entities: NameTag + rules,
   2.2 each of the 2 annotators annotates it (takes 6-15h) in Brat [19].
3. Follow up with a fix-batch of non-matching entities:
   3.1 Each annotator annotates it.
4. What next?
   4.1 Reach perfect agreement, or:
   4.2 Measure Cohen-Kappa and set good-enough threshold.
   4.3 Boost minority classes by scoring and selecting documents.
   4.4 Validate schema with a preliminary model.

Inter-Annotator Confusion Matrix (so far)



Inter-Annotator Agreement (so far)

key	precision	recall	F1-measure	support
pf	0.75	0.88	0.81	2927 (2494)
ps	0.77	0.65	0.70	655 (778)
P	0.67	0.86	0.75	144 (111)
pb	1.00	1.00	1.00	0 (0)
p_	0.48	0.52	0.50	317 (289)
cnt	0.93	0.98	0.95	352 (334)
idf	0.67	0.50	0.57	3 (4)
loc	0.82	0.82	0.82	1627 (1622)
dis	0.14	0.17	0.15	7 (6)
i_	0.65	0.28	0.39	261 (606)
mis	0.40	0.44	0.42	201 (183)
not	0.88	0.38	0.53	497 (1156)
unk	0.11	0.29	0.16	149 (59)
100	0.00	1.00	0.00	1082 (580)

Conclusion and Future Work

The current detection method deserves dramatic improvement. We are going to achieve it by building a new corpus and training a new model.
Task 1: design new detection architecture:

1. ensemble of a neural model and rules,
2. prefer recall over precision -> keep rules [20, 21],
3. try rule generation, first name derivation,
4. compare with NameTag 2 (BERT + Flair embeddings), retrain + tune parameters.

The replacement methods and their evaluation are solid, but can be improved further.
Task 2: improve PDE tracking:

1. use co-reference resolution,
2. improve form retrieval,
3. adapt *k*-anonymity or another model for measuring privacy.

Bibliography I

[1] IRTIS. Modelling the future understanding the impact of technology on adolescent's well-being (future). 2021. URL: https://irtis.muni.cz/research/projects/future.

[2] Bert J. Tolkamp, Marie J. Haskell, Fritha M. Langford, David J. Roberts, and Colin A. Morgan. Are cows more likely to lie down the longer they stand? *Applied Animal Behaviour Science*, 124(1-2):1–10, 2010.

[3] Fabian Prasser. Workshop on anonymization of research data. 2021. URL: https://www.ceitec.cz/workshop-anonymization-of-research-data/a4021.

Bibliography II

[4] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. *Technical report, SRI International*, 1998.

[5] Shuchi Chawla, Cynthia Dwork, Frank McSherry, Adam Smith, and Hoeteck Wee. Toward privacy in public databases. In *Theory of Cryptography Conference*, pages 363–385. Springer, 2005.

[6] Filip Graliński, Krzysztof Jassem, Michał Marcińczuk, and Paweł Wawrzyniak. Named entity recognition in machine anonymization. *Recent Advances in Intelligent Information Systems*:247–260, 2009.

Bibliography III

[7] Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodrguez, JA Lopez Martin, Marta Villegas, and Martin Krallinger. Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, 2019.

[8] Fadi Hassan, Josep Domingo-Ferrer, and Jordi Soria-Comas. Anonymization of unstructured data via named-entity recognition. In *International conference on modeling decisions for artificial intelligence*, pages 296–305. Springer, 2018.

Bibliography IV

[9] David Sánchez and Montserrat Batet. C-sanitized: a privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163, 2016.

[10] Fadi Hassan, David Sánchez, Jordi Soria-Comas, and Josep Domingo-Ferrer. Automatic anonymization of textual documents: detecting sensitive information via word embeddings. In *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 358–365. IEEE, 2019.

Bibliography V

[11] Fadi Hassan, David Sanchez, and Josep Domingo-Ferrer. Utility-preserving privacy protection of textual documents via word embeddings. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[12] Ghazaleh Beigi, Kai Shu, Ruocheng Guo, Suhang Wang, and Huan Liu. Privacy preserving text representation learning. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, pages 275–276, 2019.

[13] Ahmadreza Mosallanezhad, Ghazaleh Beigi, and Huan Liu. Deep reinforcement learning-based text anonymization against private-attribute inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2360–2369, 2019.

Bibliography VI

[14] United Kingdom Data Service. Text anonymization helper tool. April 27, 2016. URL: `https://bitbucket.org/ukda/ukds.tools.textanonhelper` (visited on 03/01/2021).

[15] Ishna Neamatullah, Margaret M Douglass, H Lehman Li-wei, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, 8(1):1–17, 2008.

[16] Horacio Vico and Daniel Calegari. Software architecture for document anonymization. *Electronic Notes in Theoretical Computer Science*, 314:83–100, 2015.

Bibliography VII

[17] Bennett Kleinberg, Maximilian Mozes, Yaloe van der Toolen, et al. Netanos-named entity-based text anonymization for open science. 2017.

[18] Davis Liang and Yan Shu. Deep automated multi-task learning. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 55–60, Taipei, Taiwan. Asian Federation of Natural Language Processing, November 2017. URL: https://www.aclweb.org/anthology/I17-2010.

Bibliography VIII

[19] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics, 2012.

[20] Aktualne. Machat podminecne propusteni. 2021. URL: `https://zpravy.aktualne.cz/domaci/machat-vezeni-podminene-propusteni/r~c1e8f8ee9cf711ebb9860cc47ab5f122/`.

[21] UFAL. Nametag demo. 2021. URL: `http://lindat.mff.cuni.cz/services/nametag/`.

Thank You for Your Attention!

MUNI
FACULTY
OF INFORMATICS