Pseudonymization Categories across Domain Boundaries

Maria Irena Szawerna, Simon Dobnik, Ricardo Muñoz Sánchez, Therese Lindström Tiedemann, Xuan-Son Vu, and Elena Volodina

LREC-COLING 2024
PRESENTED IN GOTHENBURG, 30.04.2024

Outline

- Background
- Questions
- Materials and methods
- Results
- Can a tagset be universal?
- Future work

Background

- **Personal Identifiable Information (PII)** complicates sharing linguistic data.
- We explore existing methods mitigating this issue with a focus on **pseudonymization**.
- We also identify types of PIIs in corpora representing various domains.

The really bad day for me when I ***lost*** my ***sister***, **Natanya**, when my ***sister died***. She was **four** years only, she was ***little*** when she ***died***.

The really bad day for me when I **XXX** my **XXXX**, **XXXXX**, when my **XXXX XXXXX**. She was **XXX** years only, she was **XXXX** when she **XXXXX**.

The really bad day for me when I ***lost*** my ***sister***, **Natanya**, when my ***sister died***. She was **four** years only, she was ***little*** when she ***died***.

The really bad day for me when I @**event** my @**fam**, @**name**, when my @**fam** @**event**. She was @**age** years only, she was @**other** when she @**event**.

The really bad day for me when I ***lost*** my ***sister***, **Natanya**, when my ***sister died*****ed**. She was **four** years only, she was ***little*** when she ***died*****ed**.

The really bad day for me when I ***lost*** my ***cousin***, **Frankie**, when my ***cousin died***. She was **six** years only, she was ***little*** when she ***dieded***.

Questions

1. What are the differences in PIIs between tagsets and domains?
2. Is a universal tagset possible?

Materials and methods

We have chosen to investigate eight different tagsets to compare the kinds of tags appearing in them:

ID	Paper	Domain
Anonymization		
1	Adams et al. (2019)	Chat
2	Pilán et al. (2022)	Legal
3	Accorsi et al. (2012)	SMS
4	Bråthen et al. (2021)	Medical
Pseudonymization		
5	Megyesi et al. (2018, 2021)	L2 essays
6	Eder et al. (2019, 2020, 2022)	E-mail
7	Alfalahi et al. (2012)	Medical
8	Dalianis (2019)	Medical

Materials and methods, part 2

When attempting to determine the applicability of one of those tagsets we chose to try to annotate data from the following domains and sources:

ID	Source	Domain	Language
A	Private	Medical	Swedish
B	Enron Corp and Cohen	E-mails	English
C	Pilán et al. (2022)	Legal	English
D	Szawerna (2023)	Memoir	Polish
E	Ahrenberg et al. (2020)	Blogs	Swedish
F	Twitter Mix, n/a, (2020)	Tweets	Swedish
G	Ahrenberg et al. (2020)	Web news	Swedish
H	Volodina et al. (2022)	Learner essays	Swedish
I	McAuley and Leskovec (2013)	Reviews	English

Results

- Variety of coverage by tagsets (including level of detail)
- PIIs can be universal, domain-specific, or simply rare
- A heterogenous *miscellaneous* category
  - What about sufficiently frequent PIIs without a category?
- The distribution of PII categories varies across domains

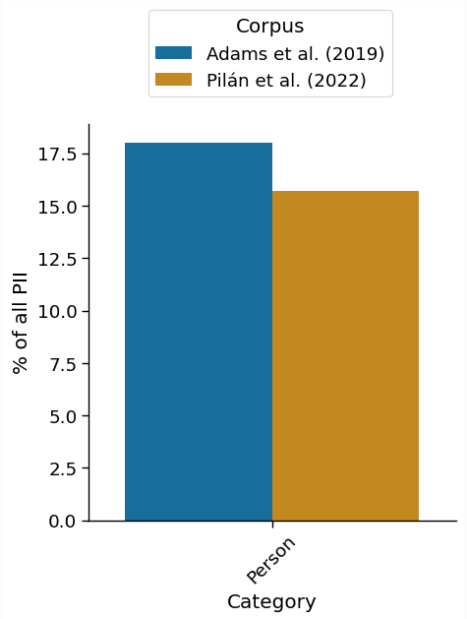
Tag	Domains
firstname_male	**1, 2, 3, 4, 5, 6, 7, 8**
firsname_female	**1, 2, 3, 4, 5, 6, 7, 8**
firstname_unknown	**1, 2, 3, 4, 5,** 6, **7, 8**
initials	**1, 2,** 3, 4, **5,** 6, **7, 8**
middlename	**1, 2,** 3, 4, **5,** 6, **7, 8**
surname	**1, 2, 3, 4, 5, 6, 7, 8**
foreign	1, 2, 3, 4, **5,** 6, 7, 8
area	**1, 2, 3, 4, 5,** 6, **7, 8**
city	**1, 2, 3, 4, 5, 6, 7, 8**
geo	**1, 2, 3, 4, 5,** 6, **7, 8**
country	**1, 2, 3, 4, 5,** 6, **7, 8**
place	**1, 2, 3, 4, 5, 6, 7, 8**
region	**1, 2, 3, 4, 5,** 6, **7, 8**
street_nr	**1, 2, 3, 4, 5, 6, 7, 8**
zip_code	**1, 2, 3, 4, 5, 6, 7, 8**
school	**1, 2,** 3, 4, **5,** 6, 7, 8
work	**1, 2,** 3, 4, **5,** 6, 7, 8
other_institution	**1, 2, 3,** 4, **5, 6,** 7, 8
transport_name	1, 2, 3, 4, **5,** 6, 7, 8
transport_nr	1, 2, 3, 4, **5,** 6, 7, 8
age_digits	1, **2,** 3, **4, 5,** 6, **7, 8**
age_string	1, **2,** 3, **4, 5,** 6, **7, 8**
date_digits	**1, 2,** 3, **4, 5, 6, 7, 8**
day	**1, 2,** 3, **4, 5, 6, 7, 8**
month_digit	**1, 2,** 3, **4, 5, 6, 7, 8**
month_word	**1, 2,** 3, **4, 5, 6, 7, 8**
year	**1, 2,** 3, **4, 5, 6, 7, 8**
phone_nr	**1, 2, 3, 4, 5, 6, 7, 8**
email	1, 2, **3,** 4, **5, 6,** 7, 8
url	**1,** 2, **3,** 4, **5, 6,** 7, 8
personid_nr	**1, 2,** 3, **4, 5, 6,** 7, 8
account_nr	1, **2,** 3, 4, **5, 6,** 7, 8
license_nr	1, **2,** 3, 4, **5, 6,** 7, 8
other_nr_seq	1, **2,** 3, 4, **5,** 6, 7, 8
extra	**1, 2, 3,** 4, **5,** 6, 7, 8
prof	1, 2, 3, 4, **5,** 6, 7, 8
edu	1, 2, 3, 4, **5,** 6, 7, 8
fam	1, 2, 3, 4, **5,** 6, 7, 8
sensitive	1, **2,** 3, 4, **5,** 6, 7, 8

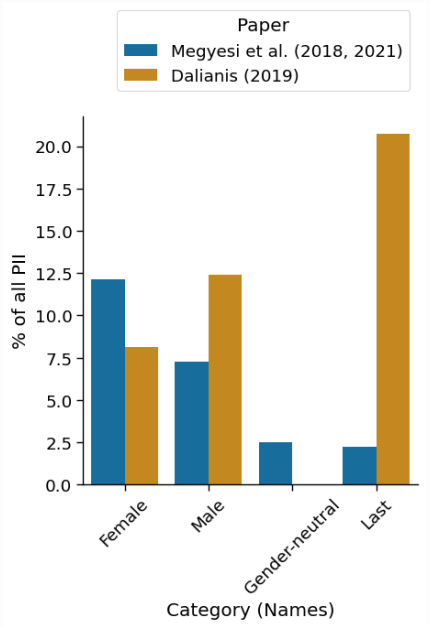
Tag	Domains
username	1, **2,** 3, 4, 5, **6,** 7, 8
password	1, 2, **3,** 4, 5, **6,** 7, 8
IP address	**1,** 2, 3, 4, 5, 6, 7, 8
product	**1,** 2, **3,** 4, 5, 6, 7, 8
facility	**1,** 2, 3, 4, 5, 6, 7, 8
nationality	**1, 2,** 3, 4, 5, 6, 7, 8
work of art	**1,** 2, 3, 4, 5, 6, 7, 8
language	**1, 2,** 3, 4, 5, 6, 7, 8
unit	**1,** 2, 3, 4, 5, 6, 7, 8
med/chem entity	**1,** 2, 3, 4, 5, 6, 7, 8
sports team	**1,** 2, 3, 4, 5, 6, 7, 8
known group	**1,** 2, 3, 4, 5, 6, 7, 8
known figure	**1,** 2, 3, 4, 5, 6, 7, 8
fictional figure	**1,** 2, 3, 4, 5, 6, 7, 8
healthcare unit	1, 2, 3, **4,** 5, 6, **7, 8**
demographic attribute	1, **2,** 3, 4, 5, 6, 7, 8
duration	1, **2,** 3, 4, 5, 6, 7, 8
quantity, value	1, **2,** 3, 4, 5, 6, 7, 8
nickname	1, **2, 3,** 4, 5, 6, 7, 8
belief	1, **2,** 3, 4, 5, 6, 7, 8
political views	1, **2,** 3, 4, 5, 6, 7, 8
sexuality, gender identity	1, **2,** 3, 4, 5, 6, 7, 8
ethnicity	1, **2,** 3, 4, 5, 6, 7, 8
health	1, **2,** 3, 4, 5, 6, 7, 8
patronymic/other name	1, 2, 3, 4, 5, 6, 7, 8

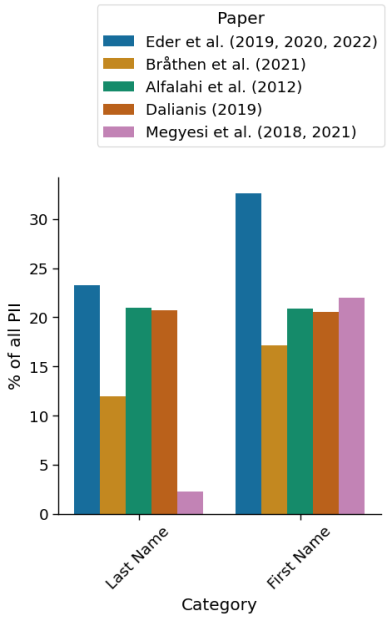
Results: a comparison of tagsets

ID	Paper
Anonymization	
1	Adams et al. (2019)
2	Pilán et al. (2022)
3	Accorsi et al. (2012)
4	Bråthen et al. (2021)
Pseudonymization	
5	Megyesi et al. (2018, 2021)
6	Eder et al. (2019, 2020, 2022)
7	Alfalahi et al. (2012)
8	Dalianis (2019)

Results: towards a comparison of domains







Tag	Domains
firstname_male	A, **B, C, D, E, F, G, H,** I
firsname_female	A, **B,** C, **D, E, F, G, H,** I
firstname_unknown	A, B, C, D, E, **F, G, H,** I
initials	A, **B, C, D, E,** F, G, **H,** I
middlename	A, B, C, **D,** E, F, **G, H,** I
surname	A, **B, C, D, E, F, G, H,** I
foreign	A, B, **C,** D, E, F, **G, H,** I
area	A, B, C, **D,** E, F, G, **H,** I
city	**A, B, C, D, E, F, G, H, I**
geo	A, B, C, D, E, F, **G, H,** I
country	A, **B, C,** D, E, **F, G, H, I**
place	**A, B,** C, D, E, **F,** G, **H,** I
region	A, **B,** C, **D, E, F, G, H, I**
street_nr	**A, B,** C, D, E, **F,** G, **H,** I
zip_code	**A, B,** C, D, E, F, G, **H,** I
school	**A,** B, C, **D,** E, **F,** G, **H,** I
work	A, **B, C,** D, E, **F, G, H,** I
other_institution	**A, B, C,** D, **E, F, G, H, I**
transport_name	A, B, C, D, E, **F, G, H,** I
transport_nr	A, B, C, D, E, **F,** G, **H,** I
age_digits	A, B, C, **D,** E, **F, G, H, I**
age_string	A, B, C, D, E, F, G, **H,** I
date_digits	**A, B,** C, D, E, F, G, **H,** I
day	A, **B, C, D,** E, **F, G, H,** I
month_digit	A, **B,** C, D, E, F, G, **H,** I
month_word	A, **B, C, D,** E, **F, G, H,** I
year	A, **B, C, D,** E, **F, G, H,** I
phone_nr	**A, B,** C, D, E, F, G, **H,** I
email	A, **B,** C, D, E, **F,** G, **H,** I
url	A, **B,** C, D, **E, F,** G, **H,** I
personid_nr	A, B, C, D, E, F, G, **H,** I
account_nr	A, B, C, D, E, F, G, **H,** I
license_nr	A, B, C, D, E, F, G, **H,** I
other_nr_seq	A, B, **C,** D, E, F, G, **H,** I
extra	A, **B, C, D, E, F, G, H, I**
prof	A, B, **C, D,** E, **F, G, H,** I
edu	A, B, **C,** D, E, F, G, **H,** I
fam	**A,** B, **C, D, E, F, G, H, I**
sensitive	**A,** B, **C,** D, E, F, G, **H,** I

Results: a comparison of domains

ID	Source
A	Private
B	Enron Corp and Cohen
C	Pilán et al. (2022)
D	Szawerna (2023)
E	Ahrenberg et al. (2020)
F	Twitter Mix, n/a, (2020)
G	Ahrenberg et al. (2020)
H	Volodina et al. (2022)
I	McAuley and Leskovec (2013)

Can a tagset be universal?

- We strive towards a **universal tagset**, acknowledging that while none of the existing tagsets comprehensively covers all types of PIIs found in texts, it is not unreasonable to pursue such a standard.
- Ideally, the universal tagset would have a hierarchical structure
- Such a tagset would have to be regularly revised, in sync with new kinds of potentially personal information emerging.

Future work

- Do we need to classify PIIs?
  - Are the categories necessary for the detection step, the generation step, both, or neither?
- What is the best balance between detailed and general classification?
- Pracitcal future work: initializing work on a universal tagset

Thank you for your attention!

References

- Pierre Accorsi, Namrata Patel, Cédric Lopez, Rachel Panckhurst, and Mathieu Roche. 2012. Seek&Hide: Anonymising a French SMS corpus using natural language processing techniques. *Linguisticae Investigationes*, 35:163–180.
- Allison Adams, Eric Aili, Daniel Aioanei, Rebecca Jonsson, Lina Mickelsson, Dagmar Mikmekova, Fred Roberts, Javier Fernandez Valencia, and Roger Wechsler. 2019. AnonyMate: A Toolkit for Anonymizing Unstructured Chat Data. *In Proceedings of the Workshop on NLP and Pseudonymisation*, pages 1–7, Turku, Finland. Linköping Electronic Press.
- Alyaa Alfalahi, Sara Brissman, and Hercules Dalianis. 2012. Pseudonymisation of Personal Names and other PHIs in an Annotated Clinical Swedish Corpus. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2012) held in conjunction with LREC 2012.*
- Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*. Istanbul: ELRA, volume Accepted, page 474–478.
- Synnøve Bråthen, Wilhelm Wie, and Hercules Dalianis. 2021. Creating and Evaluating a Synthetic Norwegian Clinical Corpus forDeIdentification. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 222–230, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.
- Luis Adrián Cabrera-Diego and Akshita Gheewala. 2024. PSILENCE: A pseudonymization tool for international law . In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 25–36, St. Julian's, Malta. Association for Computational Linguistics.
- Hercules Dalianis. 2019. Pseudonymisation of Swedish electronic patient records using a rulebased approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 16– 23, Turku, Finland. Linköping Electronic Press.
- George Danezis, Josep Domingo-Ferrer, Marit Hansen, Jaap-Henk Hoepman, Daniel Le Metayer, Rodica Tirtea, and Stefan Schiffner. 2015. Privacy and data protection by designfrom policy to engineering. *arXiv preprint arXiv:1501.03726*.
- Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2019. De-identification of emails: Pseudonymizing privacy-sensitive data in a German email corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 259– 269, Varna, Bulgaria. INCOMA Ltd.
- Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2020. CodE alltag 2.0 — a pseudonymized German-language email corpus. *In Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4466–4477, Marseille, France. European Language Resources Association.
- Elisabeth Eder, Michael Wiegand, Ulrike KriegHolz, and Udo Hahn. 2022. "Beste Grüße, Maria Meyer" — Pseudonymization of Privacy Sensitive Information in Emails. *In Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 741–752, Marseille, France. European Language Resources Association.
- EU EU Commission. 2016. *General data protection regulation.* Official Journal of the European Union, 59, 1-88.
- Langdon Holmes, Scott Crossley, Harshvardhan Sikka, and Wesley Morris. 2023. Piilo: an opensource systemfor personally identifiable information labeling and obfuscation. *Information and Learning Sciences.*
- Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation Models for Text Data: State of the art, Challenges and Future Directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. Association for Computational Linguistics.
- Beáta Megyesi, Lena Granstedt, Sofia Johansson, Julia Prentice, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, Mats Wirén, and Elena Volodina. 2018. Learner Corpus Anonymization in the Age of GDPR: Insights fromthe Creation of a Learner Corpus of Swedish. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 47–56, Stockholm, Sweden. LiU Electronic Press.

- Beáta Megyesi, Lisa Rudebeck, and Elena Volodina. 2021. Swe LL pseudonymization guidelines.
- Tuan Minh Nguyen and Xuan-Son Vu. 2023. Privacy and trust in iot ecosystems with big data: A survey of perspectives and challenges. In *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 215–222. IEEE.
- Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. *In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.
- Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framew ork for text anonymization. *Computational Linguistics*, 48(4):1053–1101.
- Maria Sierro, Begoña Altuna, and Itziar Gonzalez-Dios. 2024. Automatic detection and labelling of personal data in case reports from the ECHR in Spanish: Evaluation of tw o different annotation approaches. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 18–24, St. Julian's, Malta. Association for Computational Linguistics.
- Elena Volodina, Yousuf Ali Mohammed, Sandra Derbring, Arild Matsson, and Beata Megyesi. 2020. Tow ards Privacy by Design in Learner Corpora Research: A Case of On-the-fly Pseudonymization of Sw edish Learner Essays. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 357– 369, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Elena Volodina, Simon Dobnik, Therese Lindström Tiedemann, and Xuan-Son Vu. 2023. Grandma Karl is 27 years old – research agenda for pseudonymization of research data. In *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*, Athens, Greece, 2023, Los Alamitos. IEEE Computer Society

LR References

- Ahrenberg, Lars and Frid, Johan and Olsson, Leif-Jöran. 2020. *Swe-NERC*. Språkbanken Text, University of Gothenburg. PID https://hdl.handle.net/10794/121.
- Enron Corp and Cohen, William W. *Enron Email Dataset*. PID https://hdl.loc.gov/loc.gdc/gdcdatasets.2018487913.
- Julian J. McAuley and Jure Leskovec. 2013. *From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews*.
- n/a. 2022. *Twitter Mix*. Språkbanken Text. Distributed via SBX/CLARIN. PID https://hdl.handle.net/10794/869.
- Pilán, Ildikó and Lison, Pierre and Øvrelid, Lilja and Papadopoulou, Anthi and Sánchez, David and Batet, Montserrat. 2022. *The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization*. MIT Press.
- Maria Irena Szaw erna. 2023. *IŻ SWÓJ JĘZYK MAJĄ! An exploration of the computational methods for identifying language variation in Polish*.
- Volodina, Elena and Granstedt, Lena and Matsson, Arild and Megyesi, Beáta and Pilán , Ildikó and Prentice, Julia and Rosén, Dan and Rudebeck, Lisa and Schenström, Carl-Johan and Sundberg, Gunlög and Wirén, Mats. 2022. *SweLL-gold*. Språkbanken Text. Distributed via SBX/CLARIN. PID https://hdl.handle.net/10794/846.
- Wirén, Mats and Matsson, Arild and Rosén, Dan and Volodina, Elena. 2019. *Svala: Annotation of second-language learner text based on mostly automatic alignment of parallel corpora*. Linköping University Electronic Press.