Privacy-Preserving Natural Language Processing

Lecture 2 – Text anonymization

Prof. Dr. Ivan Habernal

April 17, 2025

www.trusthlt.org
Chair of Trustworthy Human Language Technologies (TrustHLT)
Ruhr University Bochum & Research Center Trustworthy Data Science and Security

Anonymization and the GDPR

1 Anonymization and the GDPR

2 Case studies in anonymization

3 Discussion

Personal data

Data is deemed personal if the information relates to an identified or identifiable individual

Art. 4 No. 1 GDPR.

General Data Protection Regulation — a European Union regulation on information privacy in the European Union (EU) and the European Economic Area (EEA)

Enhance individuals' control and rights over their personal information and to simplify the regulations for international business

P. Voigt and A. von dem Bussche (2017). **The EU General Data Protection Regulation (GDPR): A Practical Guide.** Springer International Publishing

Personal data

P. Voigt and A. von dem Bussche (2017). **The EU General Data Protection Regulation (GDPR): A Practical Guide.** Springer International Publishing

Individual does not need to be identified already.

The mere possibility of identification, **identifiability**, will render data **personal** under the GDPR

Identification — by combining different information that by themselves would not have traced back to the person but does so in combination

Personal data

"In October 2016, the ECJ ruled that the risk of identification appears insignificant in reality if it requires a disproportionate effort in terms of time, cost and manpower, the aforementioned being relative criteria."

"Hence, a person can be considered as identifiable if the missing information that would allow identification is (easily) accessible, for instance, because it is published on the Internet or in a (commercial) information service."

(Voigt and Bussche, 2017, p. 12)

P. Voigt and A. von dem Bussche (2017). **The EU General Data Protection Regulation (GDPR): A Practical Guide.** Springer International Publishing

Anonymization

A way of modification of personal data with the result that there remains no connection of data with an individual

Anonymised data is personal data that was rendered anonymous in such a manner that the person is no longer **identifiable**

“In case of an effective anonymisation, the GDPR does not apply”

(Voigt and Bussche, 2017, p. 13)

P. Voigt and A. von dem Bussche (2017). **The EU General Data Protection Regulation (GDPR): A Practical Guide.** Springer International Publishing

Anonymization: Practical Advice

As the EU does not provide for a standard of successful anonymisation, a combination of **randomisation** and **generalisation** techniques should be considered for stronger privacy guarantees

As a risk factor is always inherent to anonymisation, this must be considered when assessing possible techniques corresponding to the severity and likelihood of the identified risk. As a consequence, the optimal solution needs to be determined on a case-by-case basis

(Voigt and Bussche, 2017, p. 14)

P. Voigt and A. von dem Bussche (2017). **The EU General Data Protection Regulation (GDPR): A Practical Guide.** Springer International Publishing

HIPAA

The U.S. Health Insurance Portability and Accountability Act (HIPAA, 1996)

- Specifically to personal information in medical data
- “Safe Harbor” method — requires removal of 18 types of protected health information (PHI), including names, location, phone numbers

One the method is applied and PHI removed, the data might be published and are no longer subject to the HIPAA privacy rules

Anonymization and the GDPR

Text anonymization in NLP

Text anonymization

Complete and irreversible removal of personally identifiable information (PII) from text data

PII

- Directly (name, passport, number, phone number, social media ID)
- Indirectly (gender, nationality, date of birth, etc.)

O. Yermilov, V. Raheja, and A. Chernodub (2023). **"Privacy- and Utility-Preserving NLP with Anonymized data: A case study of Pseudonymization".** In: *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*. Toronto, Canada: Association for Computational Linguistics, pp. 232–241

Text anonymization

O. Yermilov, V. Raheja, and A. Chernodub (2023). **“Privacy- and Utility-Preserving NLP with Anonymized data: A case study of Pseudonymization”.** In: *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*. Toronto, Canada: Association for Computational Linguistics, pp. 232–241

Is complete anonymization possible?

Two sub-tasks of anonymization

De-identification

Process of removing specific, predefined direct identifiers from a dataset

Pseudonymisation

Process of replacing direct identifiers with pseudonyms or coded values (such “John Doe” → “Patient 3”). The mapping between coded values and the original identifiers is then stored separately.

“NLP research on text anonymisation has focused to a large extent on the tasks of de-identification” (Lison, Pilán, Sanchez, Batet, and Øvrelid, 2021, p. 4190)

P. Lison, I. Pilán, D. Sanchez, M. Batet, and L. 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)*. Online: Association for Computational Linguistics, pp. 4188–4203

Two sub-tasks of anonymization

P. Lison, I. Pilán, D. Sanchez, M. Batet, and L. 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)*. Online: Association for Computational Linguistics, pp. 4188–4203

De-identification generally modelled as a sequence labelling task, similar to Named Entity Recognition

Case studies in anonymization

1 Anonymization and the GDPR

2 Case studies in anonymization

3 Discussion

Case-study: Job postings

- Task: detecting privacy-related entities in job posts on StackOverflow
- Intention: cannot identify which company posted the job (name of the employees who posted and their contact information)
- New data: 22k EN sentences annotated with person names, contact details, and profession
- Five entities (Org, Loc, Contact, Name, Profession)
- Annotation study: three annotators, good kappa, well-done study
- Models: entity taggers (LSTM, BERT) + auxiliary tasks

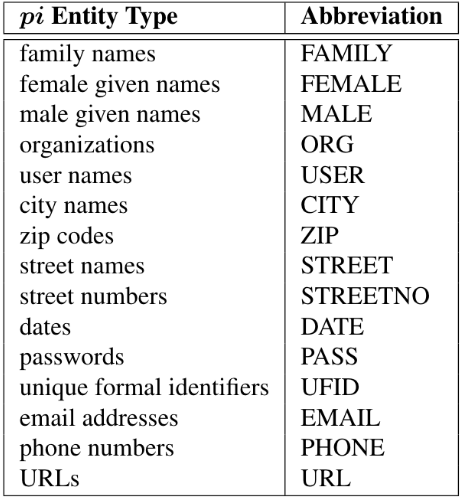
K. N. Jensen, M. Zhang, and B. Plank (2021). **"De-identification of Privacy-related Entities in Job Postings".** In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa).* Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, pp. 210–221

Case-study: Geman e-mails

- Focuses on PII recognition and substitution with surrogates
- Dataset: "CodE Alltag", German e-mails from a) USENET and b) donations
- Entity types: more than NER - names, orgs, city names, zip codes, street names, street numbers, dates, passwords, e-mails, URLs, phone numbers (see the next slides)
- Experiments: detect PIIs (BIO tagging)

E. Eder, M. Wiegand, U. Krieg-Holz, and U. Hahn (2022). **""Beste Grüße, Maria Meyer" — Pseudonymization of Privacy-Sensitive Information in Emails".** In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference.* Marseille, France: European Language Resources Association, pp. 741–752

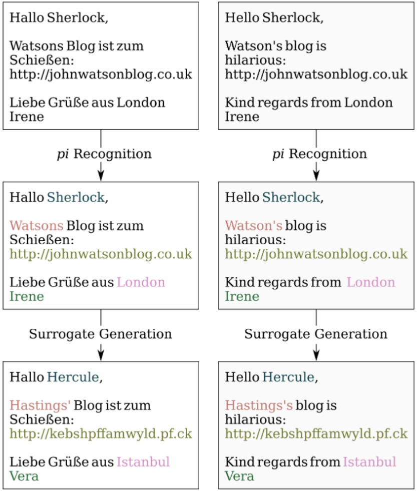
Case-study: Geman e-mails — PII types



pi **Entity Type**	**Abbreviation**
family names	FAMILY
female given names	FEMALE
male given names	MALE
organizations	ORG
user names	USER
city names	CITY
zip codes	ZIP
street names	STREET
street numbers	STREETNO
dates	DATE
passwords	PASS
unique formal identifiers	UFID
email addresses	EMAIL
phone numbers	PHONE
URLs	URL

E. Eder, M. Wiegand, U. Krieg-Holz, and U. Hahn (2022). **““Beste Grüße, Maria Meyer” — Pseudonymization of Privacy-Sensitive Information in Emails”.** In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 741–752

Case-study: Geman e-mails — Overview



Hallo Sherlock,

Watsons Blog ist zum Schießen: http://johnwatsonblog.co.uk

Liebe Grüße aus London
Irene

pi Recognition

Hallo Sherlock,

Watsons Blog ist zum Schießen: http://johnwatsonblog.co.uk

Liebe Grüße aus London
Irene

Surrogate Generation

Hallo Hercule,

Hastings' Blog ist zum Schießen: http://kebshpffamwyld.pf.ck

Liebe Grüße aus Istanbul
Vera

Hello Sherlock,

Watson's blog is hilarious: http://johnwatsonblog.co.uk

Kind regards from London
Irene

pi Recognition

Hello Sherlock,

Watson's blog is hilarious: http://johnwatsonblog.co.uk

Kind regards from London
Irene

Surrogate Generation

Hello Hercule,

Hastings's blog is hilarious: http://kebshpffamwyld.pf.ck

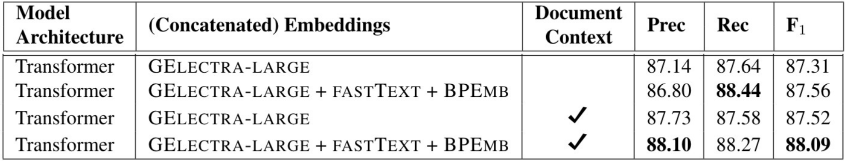
Kind regards from Istanbul
Vera

E. Eder, M. Wiegand, U. Krieg-Holz, and U. Hahn (2022). **““Beste Grüße, Maria Meyer” — Pseudonymization of Privacy-Sensitive Information in Emails”.** In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference.* Marseille, France: European Language Resources Association, pp. 741–752

Case-study: Geman e-mails — Evaluation

BIO scheme: ‘B’ — Beginning of an entity, ‘I’ (Inside) for its continuation, and ‘O’ (Outside) for tokens that do not belong to any pi entity)

Weighted average of precision, recall and F1 score



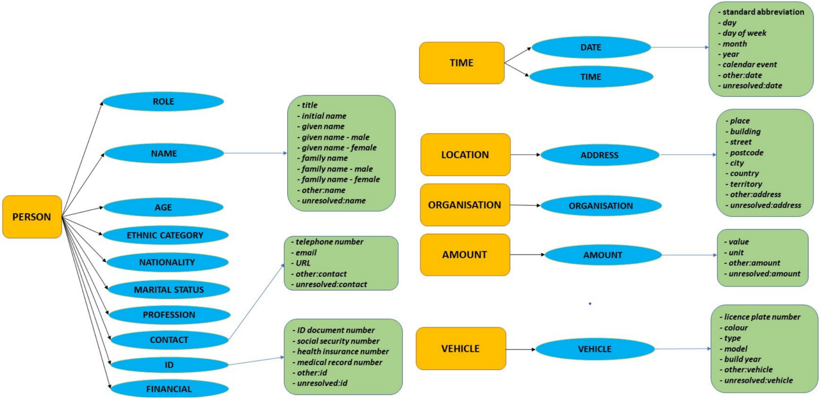
Model Architecture	**(Concatenated) Embeddings**	**Document Context**	**Prec**	**Rec**	**F$_1$**
Transformer	GElectra-Large		87.14	87.64	87.31
Transformer	GElectra-Large + fastText + BPEmb		86.80	**88.44**	87.56
Transformer	GElectra-Large	✓	87.73	87.58	87.52
Transformer	GElectra-Large + fastText + BPEmb	✓	**88.10**	88.27	**88.09**

How robustly does the system protect privacy?

E. Eder, M. Wiegand, U. Krieg-Holz, and U. Hahn (2022). **““Beste Grüße, Maria Meyer” — Pseudonymization of Privacy-Sensitive Information in Emails”.** In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 741–752

Case-study: MAPA project

- 3-level hierarchy of entities: level 1: person, time, location, organization, amount, vehicle



TIME
DATE
TIME
- standard abbreviation
- day
- day of week
- month
- year
- calendar event
- other:date
- unresolved:date

PERSON
ROLE
NAME
AGE
ETHNIC CATEGORY
NATIONALITY
MARITAL STATUS
PROFESSION
CONTACT
ID
FINANCIAL

- title
- initial name
- given name
- given name - male
- given name - female
- family name
- family name - male
- family name - female
- other:name
- unresolved:name

- telephone number
- email
- URL
- other:contact
- unresolved:contact

- ID document number
- social security number
- health insurance number
- medical record number
- other:id
- unresolved:id

LOCATION
ADDRESS
- place
- building
- street
- postcode
- city
- country
- territory
- other:address
- unresolved:address

ORGANISATION
ORGANISATION

AMOUNT
AMOUNT
- value
- unit
- other:amount
- unresolved:amount

VEHICLE
VEHICLE
- licence plate number
- colour
- type
- model
- build year
- other:vehicle
- unresolved:vehicle

V. Arranz, K. Choukri, M. Cuadros, A. García Pablos, L. Gianola, C. Grouin, M. Herranz, P. Paroubek, and P. Zweigenbaum (2022). **"MAPA Project: Ready-to-Go Open-Source Datasets and Deep Learning Technology to Remove Identifying Information from Text Documents".** In: *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference.* Marseille, France: European Language Resources Association, pp. 64–72

Case-study: MAPA project

- NER model: BERT
- Classification only in 2 levels of hierarchy
- Rntity replacement: random or using a LM
- Datasets: EUR-LEX (2k sentences per lang)
- Medical data: created “fake ones” in French by replacing “he”, “she”, “the patient” with fake names
- Also some legal judgment in Spanish
- Annotated data machine-translated

V. Arranz, K. Choukri, M. Cuadros, A. García Pablos, L. Gianola, C. Grouin, M. Herranz, P. Paroubek, and P. Zweigenbaum (2022). **“MAPA Project: Ready-to-Go Open-Source Datasets and Deep Learning Technology to Remove Identifying Information from Text Documents”.** In: *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference.* Marseille, France: European Language Resources Association, pp. 64–72

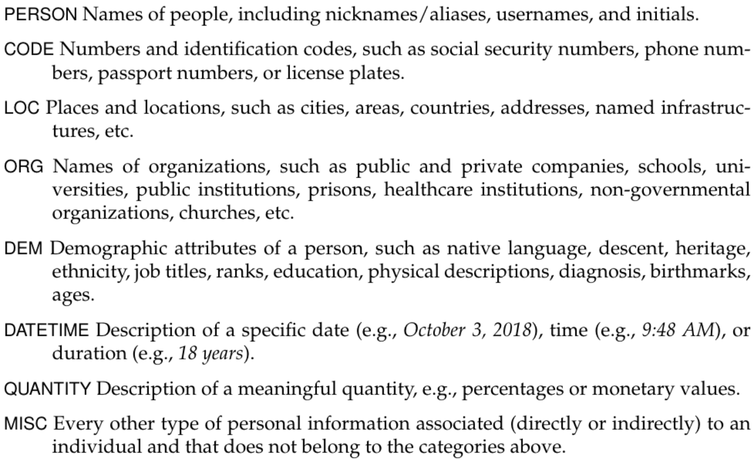
Case study: Pseudonymization with LLMs

- Pseudonymization adopted from Eder, Wiegand, Krieg-Holz, and Hahn (2022) — recognize and replace entities
- Tested 3 approaches: NER, seq2seq (enc-dec), LLM with zero shot
- 3 classes of entities: PER, LOC, ORG
- Spacy for NER; for LLM they use GPT-3 and ChatGPT
- Downstream test on pseudoanonym. data -> summarization (CNN daily mail, standard) and classification (IMDB, standard)
- Results: on summarization drop in BLEU (quite a big one), on sentiment very little (why?)

O. Yermilov, V. Raheja, and A. Chernodub (2023). **“Privacy- and Utility-Preserving NLP with Anonymized data: A case study of Pseudonymization”.** In: *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*. Toronto, Canada: Association for Computational Linguistics, pp. 232–241

Case study: TAB Project

1,268 En cases from ECHR (Introduction and Facts sections)



PERSON Names of people, including nicknames/aliases, usernames, and initials.

CODE Numbers and identification codes, such as social security numbers, phone numbers, passport numbers, or license plates.

LOC Places and locations, such as cities, areas, countries, addresses, named infrastructures, etc.

ORG Names of organizations, such as public and private companies, schools, universities, public institutions, prisons, healthcare institutions, non-governmental organizations, churches, etc.

DEM Demographic attributes of a person, such as native language, descent, heritage, ethnicity, job titles, ranks, education, physical descriptions, diagnosis, birthmarks, ages.

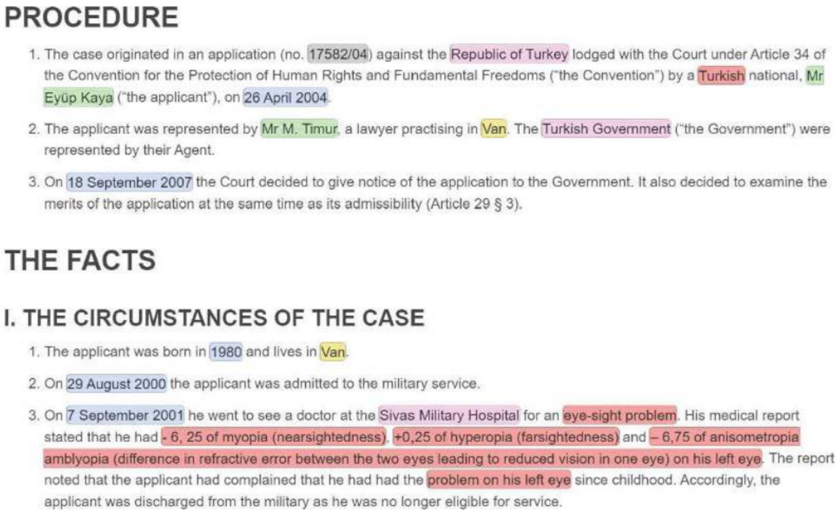
DATETIME Description of a specific date (e.g., *October 3, 2018*), time (e.g., *9:48 AM*), or duration (e.g., *18 years*).

QUANTITY Description of a meaningful quantity, e.g., percentages or monetary values.

MISC Every other type of personal information associated (directly or indirectly) to an individual and that does not belong to the categories above.

I. Pilán, P. Lison, L. Øvrelid, A. Papadopoulou, D. Sánchez, and M. Batet (2022). **“The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization”.** In: *Computational Linguistics* 48.4, pp. 1053–1101

Case study: TAB Project



PROCEDURE

1. The case originated in an application (no. 17582/04) against the Republic of Turkey lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms ("the Convention") by a Turkish national, Mr Eyüp Kaya ("the applicant"), on 26 April 2004.

2. The applicant was represented by Mr M. Timur, a lawyer practising in Van. The Turkish Government ("the Government") were represented by their Agent.

3. On 18 September 2007 the Court decided to give notice of the application to the Government. It also decided to examine the merits of the application at the same time as its admissibility (Article 29 § 3).

THE FACTS

I. THE CIRCUMSTANCES OF THE CASE

1. The applicant was born in 1980 and lives in Van.

2. On 29 August 2000 the applicant was admitted to the military service.

3. On 7 September 2001 he went to see a doctor at the Sivas Military Hospital for an eye-sight problem. His medical report stated that he had - 6, 25 of myopia (nearsightedness), +0,25 of hyperopia (farsightedness) and – 6,75 of anisometropia amblyopia (difference in refractive error between the two eyes leading to reduced vision in one eye) on his left eye. The report noted that the applicant had complained that he had had the problem on his left eye since childhood. Accordingly, the applicant was discharged from the military as he was no longer eligible for service.

I. Pilán, P. Lison, L. Øvrelid, A. Papadopoulou, D. Sánchez, and M. Batet (2022). **"The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization".** In: *Computational Linguistics* 48.4, pp. 1053–1101

Case study: TAB Project — Evaluation

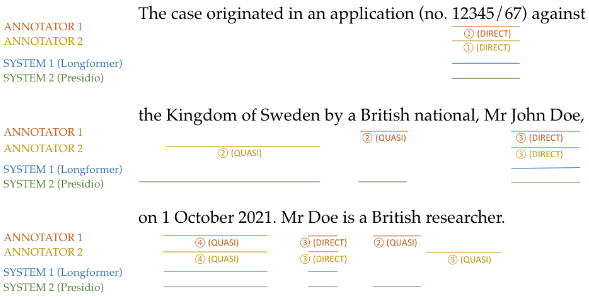
Evaluation - recall of PII = privacy protection, precision of PII = utility

Absolute recall — relies on a uniform weight over all annotated identifiers and thus fails to account for the fact that some (quasi-)identifiers have a much larger influence on the disclosure risk than others

Evaluation of anonymization methods is the need to compute the recall at the level of entities rather than mentions. Whenever an entity is mentioned several times in a given document, it only makes sense to view this entity as “protected” if all of its mentions are masked.

I. Pilán, P. Lison, L. Øvrelid, A. Papadopoulou, D. Sánchez, and M. Batet (2022). **“The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization”.** In: *Computational Linguistics* 48.4, pp. 1053–1101

Case study: TAB Project — New metrics



The case originated in an application (no. 12345/67) against the Kingdom of Sweden by a British national, Mr John Doe, on 1 October 2021. Mr Doe is a British researcher.

I. Pilán, P. Lison, L. Øvrelid, A. Papadopoulou, D. Sánchez, and M. Batet (2022). **"The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization".** In: *Computational Linguistics* 48.4, pp. 1053–1101

Entity is correctly masked if and only if the anonymization model manages to completely mask all of its mentions

Separate recall measures for the direct id (person's names, case numbers, etc.) and the quasi-id (dates, locations, etc.)

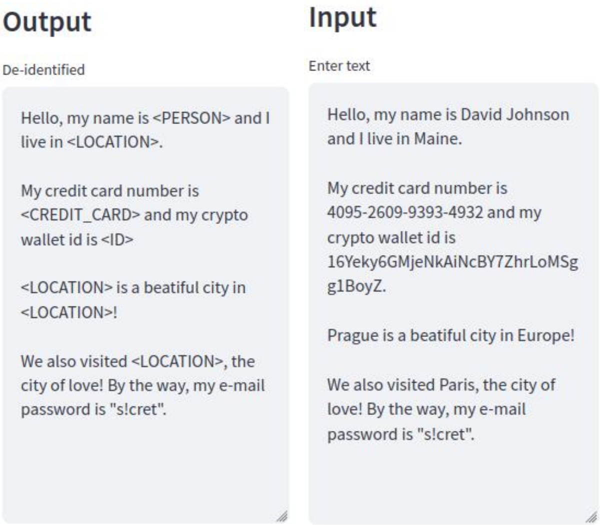
Discussion

1 Anonymization and the GDPR

2 Case studies in anonymization

3 Discussion

Presidio



Output

De-identified

Hello, my name is <PERSON> and I live in <LOCATION>.

My credit card number is <CREDIT_CARD> and my crypto wallet id is <ID>

<LOCATION> is a beatiful city in <LOCATION>!

We also visited <LOCATION>, the city of love! By the way, my e-mail password is "s!cret".

Input

Enter text

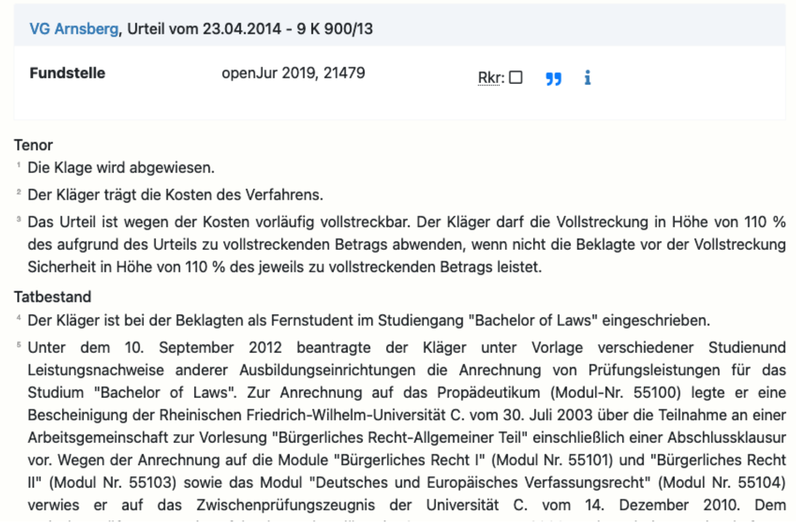
Hello, my name is David Johnson and I live in Maine.

My credit card number is 4095-2609-9393-4932 and my crypto wallet id is 16Yeky6GMjeNkAiNcBY7ZhrLoMSgg1BoyZ.

Prague is a beatiful city in Europe!

We also visited Paris, the city of love! By the way, my e-mail password is "s!cret".

Limits of anonymization?



VG Arnsberg, Urteil vom 23.04.2014 - 9 K 900/13

Fundstelle openJur 2019, 21479 Rkr: □

Tenor

1 Die Klage wird abgewiesen.

2 Der Kläger trägt die Kosten des Verfahrens.

3 Das Urteil ist wegen der Kosten vorläufig vollstreckbar. Der Kläger darf die Vollstreckung in Höhe von 110 % des aufgrund des Urteils zu vollstreckenden Betrags abwenden, wenn nicht die Beklagte vor der Vollstreckung Sicherheit in Höhe von 110 % des jeweils zu vollstreckenden Betrags leistet.

Tatbestand

4 Der Kläger ist bei der Beklagten als Fernstudent im Studiengang "Bachelor of Laws" eingeschrieben.

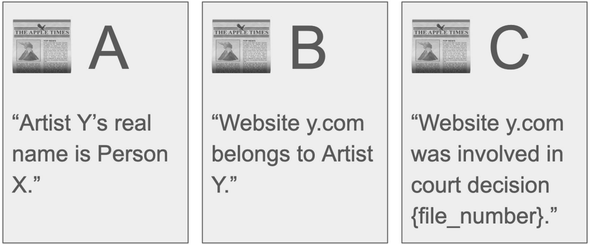
5 Unter dem 10. September 2012 beantragte der Kläger unter Vorlage verschiedener Studienund Leistungsnachweise anderer Ausbildungseinrichtungen die Anrechnung von Prüfungsleistungen für das Studium "Bachelor of Laws". Zur Anrechnung auf das Propädeutikum (Modul-Nr. 55100) legte er eine Bescheinigung der Rheinischen Friedrich-Wilhelm-Universität C. vom 30. Juli 2003 über die Teilnahme an einer Arbeitsgemeinschaft zur Vorlesung "Bürgerliches Recht-Allgemeiner Teil" einschließlich einer Abschlussklausur vor. Wegen der Anrechnung auf die Module "Bürgerliches Recht I" (Modul Nr. 55101) und "Bürgerliches Recht II" (Modul Nr. 55103) sowie das Modul "Deutsches und Europäisches Verfassungsrecht" (Modul Nr. 55104) verwies er auf das Zwischenprüfungszeugnis der Universität C. vom 14. Dezember 2010. Dem

Attacks on anonymized documents

Explored reconstructing Swiss court decisions (cases from the federal court)

Auxiliary data: Created a dataset by manually linking court rulings and newspaper articles using keywords, e.g., “4A_375/2021”, ”10 years in prison”, etc.

A. Nyffenegger, M. Stürmer, and J. Niklaus (2024). **“Anonymity at Risk? Assessing Re-Identification Capabilities of Large Language Models in Court Decisions”.** In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Mexico City, Mexico: Association for Computational Linguistics, pp. 2433–2462



A: “Artist Y’s real name is Person X.”

B: “Website y.com belongs to Artist Y.”

C: “Website y.com was involved in court decision {file_number}.”

Exercise

Build your own text anonymizer

- Get existing dataset (see the case studies)
- Fine-tune a small BERT transformer model (or RNN tagger)
- Evaluate quantitatively and explore errors manually

License and credits

Credits

Ivan Habernal

Content from ACL Anthology papers licensed under CC-BY
`https://www.aclweb.org/anthology`

Partly inspired by lectures from Antti Honkela, Aurélien Bellet, Gautam Kamath