Annotating Personal Information in Swedish Texts with SPARV

Maria Irena Szawerna[*], **David Alfter**[†‡], **Elena Volodina**[*]
[*]Språkbanken Text, SFS, University of Gothenburg, Sweden
[†]GRIDH, LIR, University of Gothenburg, Sweden
[‡]InfraVis, Sweden
mormor.karl@svenska.gu.se
{maria.szawerna,david.alfter,elena.volodina}@gu.se

Abstract

Digital Humanities (DH) research, among many others, relies on data, a subset of which comes in the form of language data that contains personal information (PI). Working with and sharing such data has ethical and legal implications. The process of removing (anonymization) or replacing (pseudonymization) of personal information in texts may be used to address these issues, and often begins with a PI detection and labeling stage. We present a new tool for personal information detection and labeling for Swedish, SBX-PI-DETECTION (henceforth SBX-PI), alongside a visualization interface, (IM)PERSONAL DATA, which allows for the comparison of outputs from different tools. A valuable feature of SBX-PI is that it enables the users to run the annotation locally. It is also integrated into the text annotation pipeline SPARV, allowing for other types of annotation to be performed simultaneously and contributing to the privacy by design requirement set by the GDPR. A novel feature of (IM)PERSONAL DATA is that it allows researchers to assess the extent of detected PI in a text and how much of it will be manipulated once anonymization or pseudonymization are applied. The tools are primarily aimed at researchers within Digital Humanities and Natural Language Processing and are linked to CLARIN's Virtual Language Observatory.[1]

1 Introduction and prior work

Personal information (PI) [2] is ubiquitous in many kinds of language data – data which oftentimes is the basis for research in fields such as Digital Humanities (DH), Natural Language Processing (NLP), or linguistics. When working with this kind of data, one may choose to employ privacy-protection measures in order to comply with appropriate legislation (e.g. GDPR (Official Journal of the European Union, 2016)) or out of ethically motivated concerns for data subject privacy. Commonly employed privacy-protection methods at the text level are the removal (*anonymization*) or replacement (*pseudonymization)* of PI. The difference is illustrated in (1), where *the original sentence* is anonymized using [redacted], pseudonymized using `tags` and pseudonymized using **replacement entities**.

(1) *Mitt namn är Sonja och jag är 29*
Mitt namn är [redacted] och jag är [redacted]
Mitt namn är `name` och jag är `age`
Mitt namn är **Anna** och jag är **31**

'My name is Sonja and I am 29'

While anonymization and pseudonymization can be carried out manually, it is very time-consuming. As with many other types of annotation, this can be sped up with the help of automated methods.

Both pseudonymization and anonymization can be construed of as two-stage processes, with the first one — PI detection and labeling, necessary in order to know what elements need to be handled — being shared by both, and succeeded by, respectively, the removal or replacement of detected spans.[3] Automatic PI detection and labeling is a task closely related to Named Entity Recognition and Classification (NER, NERC) (Lison et al., 2021). The key distinction is that while the overlap between PI and Named Entities (NEs) is large,

[1] https://spraakbanken.gu.se/en/analyses/sbx-swe-pi_detection-sparv (DOI) https://vlo.clarin.eu/record/Other_47_doi_10_23695_6wp0_ds77.xml?1&count=3&index=0&q=sparv-plugin (VLO)

[2] In this paper this term is used to refer to both information that on its own or in combination with other pieces of information can be used to re-identify a natural person, i.e. Personally Identifiable Information (PII), and sensitive information (e.g. sexuality, religious beliefs).

[3] There exist approaches where this distinction is not made, e.g. seq2seq or LLM prompting to directly return a sanitized text, cf. Yermilov et al. (2023).

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 155–163, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-015

not all PIs are NEs and not all NEs are personal in nature; whether a piece of information in a text is personal is highly dependent on the context.

The approaches to PI detection and labeling range from rule-based systems (Accorsi et al., 2012; Dalianis, 2019; Blokland et al., 2020; Volodina et al., 2020) through machine learning approaches using e.g. Conditional Random Fields (Berg and Dalianis, 2019, 2020; Adams et al., 2019; Eder et al., 2020), Recurrent Neural Networks (Adams et al., 2019; Eder et al., 2020; Jensen et al., 2021; López-García et al., 2023), or Transformer-based classifiers (Johnson et al., 2020; Jensen et al., 2021; Eder et al., 2022; Meaney et al., 2022; López-García et al., 2023; Ngo et al., 2024; Szawerna et al., 2024, 2025), to Large Language Model (LLM) prompting (Yang et al., 2023; Ilinykh and Szawerna, 2025) or knowledge distillation from LLMs (Deußer et al., 2025), and combined approaches (Jensen et al., 2021; Eder et al., 2022; Cabrera-Diego and Gheewala, 2024). Notably, the ability to conduct automated PI detection locally is advantageous, since sending potentially sensitive data to external tools or APIs increases the chance of a security breach and information leakage. Ease of use also plays a role, since the more complicated a tool is to set up or use, the smaller its userbase is likely to become.

In this paper we present a new, flexible plugin SBX-PI for Personal Information detection and labeling in Swedish for the SPARV text annotation pipeline, alongside a novel visualization tool, (IM)PERSONAL DATA, intended for analyzing and comparing the performance of such systems. The tools stand out in two ways: on the one hand, SBX-PI embeds ethics into SPARV through addressing the g the ‘privacy by design’ requirement imposed by the GDPR (Official Journal of the European Union, 2016); on the other hand, (IM)PERSONAL DATA is the only tool known to us allowing the users to assess how much of the research data will be altered once anonymization or pseudonymization are applied, making it possible to assess the value and validity of the data after the applied manipulations. We showcase both tools on a sample text[4] and compare the performance of the PI detection plugin to that of the commercially available tool MICROSOFT PRESIDIO,[5], the web-based tool for Swedish texts HB DEID (Berg and Dalianis, 2019, 2021),[6] and the results of prompting GEMMA 2 9B (Gemma Team et al., 2024),[7] with the help of the visualization tool. GEMMA 2 9B has previously been used by Ilinykh and Szawerna (2025) to detect and label personal information and performed best out of the tested models; due to its size it could potentially be run locally. We discuss the differences in performance and the advantages and disadvantages of the aforementioned approaches. Finally, we present plans and suggestions for further development of both of our tools.

2 SPARV plugin

SPARV (v5.3.0, Hammarstedt et al., 2022)[8] is a Python-based modular command line tool for text annotation designed primarily for Swedish, created and maintained by Språkbanken Text. It can be run locally and is designed to handle importing the data, annotating it, and exporting it. The choice of formats and annotations is controlled using a corpus configuration file. SPARV’s design makes it also very easy to extend it with new modules or plugins, which has enabled the addition of Personal Information detection and labeling, meaning that this task can be performed together with other kinds of annotation, e.g. part-of-speech tagging.

2.1 System Design

Our plugin[9] makes use of six PI detection classifiers for Swedish (Szawerna et al., 2025) based on `KB/bert-base-swedish-cased` (Devlin et al., 2019; Malmsten et al., 2020), hosted on Språkbanken Text’s HuggingFace page.[10] The models’ performance reported in Szawerna et al. (2025) is shown in Table 1. The models differ in terms of the tags that they can assign to the detected spans, as outlined in Table 2. It is very important to highlight that, as per their HuggingFace model cards, these models “[...] perform best on [...] second-language learner essays,” the type of texts that they were trained on. By not including the models in the plugin itself but accessing them

[4] While we acknowledge that evaluation measures would be valuable, there is no openly available dataset for PI detection in Swedish. The performance of the models used in our tool on their test set was reported in Szawerna et al. (2025).

[5] https://microsoft.github.io/presidio/

[6] https://hbdeid.dsv.su.se/

[7] https://huggingface.co/google/gemma-2-9b

[8] https://spraakbanken.gu.se/sparv/

[9] https://github.com/spraakbanken/sparv-sbx-pi-detection

[10] https://huggingface.co/sbx

Model	F2
detailed_iob	0.519 ± 0.085
detailed	0.558 ± 0.063
general_iob	0.720 ± 0.054
general	0.763 ± 0.059
basic_iob	0.800 ± 0.045
basic	0.824 ± 0.038

Table 1: Mean results ± standard deviation for each type of model, courtesy of the authors (Szawerna et al., 2025).

via HuggingFace, we make it possible to access newer versions of the same models, should they ever be released. It also makes it relatively simple to expand the plugin with additional models by modifying very little of the code.

We follow the general recommended structure for SPARV plugins.[11] The code is accompanied by a number of required or recommended files specifying the functionality or behavior of the plugin for both SPARV itself and the user. The plugin's requirements are `Sparv 5.0` or higher, `Transformers 4.51.3` or higher (Wolf et al., 2020), and `PyTorch` (Ansel et al., 2024).

The core of the plugin's functionality lies in the `pi_detection.py` file, which defines the functions called when the annotations provided by this plugin are requested by the user. In such a case the input is first tokenized at word level using a user-defined or SPARV's default tokenizer. The appropriate classifier model and corresponding tokenizer are loaded in according to the corpus configuration using `Transformers`. Since BERT-based models use sub-word tokenization and have a maximum input length, each input text is chunked if it were to exceed the length of 512 sub-word tokens, with the boundaries following the word-level tokenization. Next, predictions are obtained from the model for each chunk. Finally, these are mapped back to the word-level tokens. In cases where multiple sub-word tokens constituting one word have received different tags, the following heuristics are applied: i) if at least one sub-word token is tagged as personal information, the entire word is tagged as that and ii) if two different personal information tags were assigned to two sub-word tokens of one word, the one closest to the beginning of the word is selected, as we consider that to be more likely to be the meaning-bearing element of the word. These tags are then forwarded to the export method defined by the user in the corpus configuration.

We also provide a sample corpus with our plugin, which consists of two text files, one with an example essay and one intended for the user to edit, alongside a corpus configuration file.

2.2 Functionality

Once the plugin is installed following the instructions that come with it, the user can request the PI annotation in the `config.yaml` configuration file for their corpus. First of all, in `annotations`, one has to specify the annotation type as `<token>:sbx_pi_detection.pi`. Next, the specific tagset (and, consequently, classifier) has to be specified, e.g. `pi_detection: general`. The names of the available tagsets in the configuration are the same as in Table 2, and more detailed user instructions are included in the plugin's README file. Both the input and output format for the data are independent from our plugin and depend on the user choice defined in the corpus configuration.

A key advantage that our plugin has is its integration into SPARV, as that allows for other types of annotation to be carried out simultaneously, according to what is defined in the corpus configuration. This also makes it easy for current SPARV users to incorporate PI annotation in their workflow.

3 (Im)Personal Data visualization

In order to visualize the system output, a custom visualization[12] was commissioned with INFRAVIS[13], the Swedish National Research Infrastructure for Data Visualization.

3.1 System Design

The interface is realized as a Vue 3 frontend and builds on two modules. The first module uses Texty (Nualart and Pérez-Montoro, 2013), "an icon that represents the physical distribution of keywords of a text as a flat image," to give an overall impression of the distribution of labels in the text (see Figure 1). The second module aligns the input texts on the word level and allows for the comparison of labels across different methods (see Figure 2).

In order to allow for new text additions, the interface is written in such a way as to adapt to new

[11] `https://spraakbanken.gu.se/sparv/developers-guide/writing-sparv-plugins/`

[12] `https://github.com/spraakbanken/impersonaldata`

[13] `infravis.se`

Model	Tags
detailed	O, firstname_male, firstname_female, firstname_unknown, initials, middlename, surname, school, work, other_institution, area, city, geo, country, place, region, street_nr, zip_code, transport_name, transport_nr, age_digits, age_string, date_digits, day, month_digit, month_word, year, phone_nr, email, url, personid_nr, account_nr, license_nr, other_nr_seq, extra, prof, edu, fam, sensitive
detailed_iob	O, B-firstname_male, I-firstname_male, ...
general	O, personal_name, institution, geographic, transportation, age, date, other
general_iob	O, B-personal_name, I-personal_name, B-institution, ...
basic	O, S
basic_iob	O, B, I

Table 2: Tagsets in the models available in the plugin. O appears in all of them and marks the non-PI tokens. IOB models have the same semantic categories as their base versions, but with the addition of marking the beginning and inside of the span. See Megyesi et al. (2018) and Szawerna et al. (2025) for more details on the tagsets and models.

data automatically. New data is added by adding the annotated texts in a specific folder, running the Texty Python script, and, finally, running a custom script that calculates word alignment and copies all the relevant data to the frontend folder.

3.2 Functionality

In the interface, the user can choose one of the preselected texts, which is then displayed. The user can then visualize the high-level label distribution of different methods with Textys, and inspect the labels more closely in detail view.

The current interface loads static pre-computed files, but this behavior may be changed in the future — adding a proper backend could allow users to test their own texts dynamically.

4 Case study

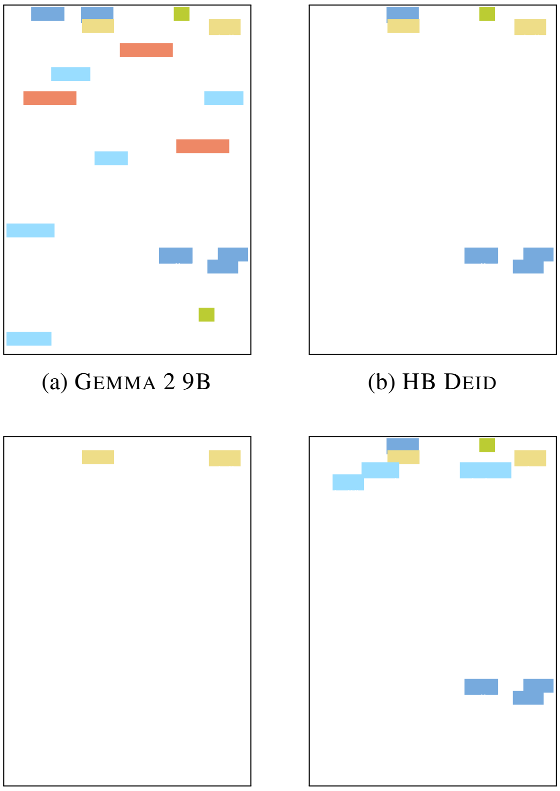
We use a sample text in order to better illustrate the performance of our plugin and the visualization tool. The text is a fictive personal story in Swedish, i.e. it contains information that would be personal if it referred to any natural person, and is structured like a personal story.[14]

We obtained PI annotations from four different tools: (a) Gemma 2 9B (Gemma Team et al., 2024), (b) HB Deid (Berg and Dalianis, 2019, 2021), (c) Microsoft Presidio, (d) our plugin. In the case of Microsoft Presidio and HB Deid we mapped these tools' tagsets to the one used by the general model in our plugin, and we instructed Gemma 2 9B to follow the same type of annotation (one-shot prompting adapted from Ilinykh and Szawerna (2025) with alterations for a different tagset and enforcing a JSON output). Importantly, using Microsoft Presidio for a language other than English requires additional coding to enable the use of NER models for the language in question. While Microsoft Presidio can be further customized (e.g. by adding rules), we opted for trying to use it as ”out of the box” as possible; the same is true in the case of prompts for Gemma 2 9B, which we did not engineer beyond including our tagset in the aforementioned prompt structure. We unified the output formats to follow the requirements for inputs to the visualization tool.

Figure 1 shows the generated Texty images for the annotated text. It is immediately visible that Gemma 2 9B predicts the most diverse categories, followed by our plugin; upon closer inspection, though, it can be noted that the models disagree on which entities should be marked as other. HB Deid only detects three categories, and Microsoft Presidio just one; in the web interface of the visualization tool these are identified as personal_name, age and geographic for the former and only geographic for the latter.

The detailed view — shown in Figure 2 — is required to properly assess the performance of the tools against each other, as that is where the anno-

[14] The text can be found here: https://github.com/mormor-karl/annotating-PI-with-SPARV



(a) GEMMA 2 9B

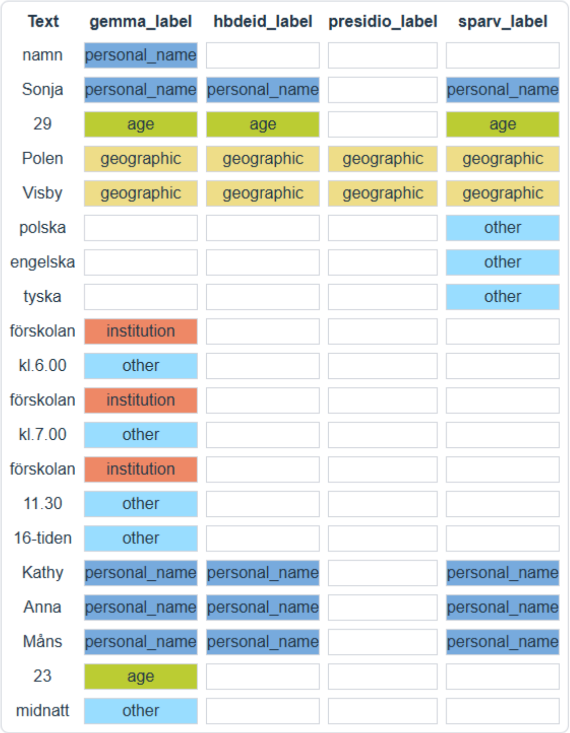
(b) HB DEID

(c) MICROSOFT PRESIDIO

(d) SPARV plugin

Figure 1: Texty visualizations for the sample essay for each annotation tool. Colors represent PI of different categories across the running text.

tated tokens are displayed. All of the tools agree on the annotation of the `geographic` entities, and all but MICROSOFT PRESIDIO correctly identify the ages and personal names present in the text; here GEMMA 2 9B returns two false positives, marking *namn* 'name' and *23* (which in the context clearly refers to a time). Further differences between our Sparv plugin and GEMMA 2 9B concern the `institution` and `other` categories. The Sparv plugin is the only one to mark the foreign languages *polska* 'Polish', *engelska* 'English', and *tyska* 'German' as `other`, which is an expected behavior. GEMMA 2 9B instead assigns this tag to four time points (including *midnatt* 'midnight') marking the daily routine of the person in the text; while this was not overtly stated in the prompt, it is interesting to see the LLM make this generalization, as this type of information could in some cases lead to re-identification. Finally, GEMMA 2 9B also tags the three mentions of *förskolan* 'the kindergarten' as `institution`. This is another justified generalization on the LLM's part, as the Sparv plugin tends to only mark specific institutions with that tag due to the way it was used in the training data (i.e. these would likely have been tagged by the plugin if they mentioned the name of the kindergarten). It is, however, worth pointing out that a big part of the text describes the activities at the kindergarten, meaning that if the type of workplace were to lead to reidentification, more than just 'kindergarten' would have to be handled. Interestingly, none of the models marked *förskollärare* 'kindergarten teacher' as `other` (which is meant to include professions).



Text	gemma_label	hbdeid_label	presidio_label	sparv_label
namn	personal_name			
Sonja	personal_name	personal_name		personal_name
29	age	age		age
Polen	geographic	geographic	geographic	geographic
Visby	geographic	geographic	geographic	geographic
polska				other
engelska				other
tyska				other
förskolan	institution			
kl.6.00	other			
förskolan	institution			
kl.7.00	other			
förskolan	institution			
11.30	other			
16-tiden	other			
Kathy	personal_name	personal_name		personal_name
Anna	personal_name	personal_name		personal_name
Måns	personal_name	personal_name		personal_name
23	age			
midnatt	other			

Figure 2: Detailed view of annotation differences from the visualization tool.

Overall, the Sparv plugin and GEMMA 2 9B are the clear forerunners for this text. Sparv is somewhat more conservative and does not make the same kinds of generalizations as GEMMA 2 9B, but it does not return false positives either. Both of these tools potentially miss some additional personal information, which highlights the importance of using these to assist de-identification, but not completely automatize it, as it is a high-stakes task and no existing tool can guarantee 100% accuracy.

Another relevant point for comparison is the time it takes the tools to annotate the data. Table 3 shows the times we have measured, although they are not fully comparable. HB DEID is a web-based demonstrator and while the annotation seemed instantaneous, it then had to be manually transfered into a machine-readable format. While Sparv and MI-

Tool	Time
Sparv plugin	15s
Microsoft Presidio	10s
HB Deid	-
Gemma 2 9B	59s

Table 3: A comparison of time it takes to run the different annotation tools.

crosoft Presidio were run on one of our local machines, Gemma 2 9B was run on a server with two GeForce RTX 2080 Ti GPUs. With that in mind, our plugin seems to strike a good balance between performance, speed, and hardware requirements, and therefore a clear winner when it comes to sustainability and eco-friendliness, as the other well-performing tool takes nearly four times longer on better hardware.

5 Discussion and conclusions

We have presented SBX-PI, a new tool for personal information detection and labeling for Swedish texts which empowers researchers in Digital Humanities, linguistics, Natural Language Processing and other research domains dependent on access to language data. SBX-PI functions as a plugin for the text annotation tool Sparv. As such, its functionality can be combined with a range of other annotations. Our plugin is relatively lightweight and fast for its performance. It does not require extensive programming knowledge to use, but with some knowledge of Python it can be easily modified. Such modifications may include allowing the model to use different PI classifier models, potentially extending its use beyond Swedish. Additionally, framing this tool as a Sparv plugin makes it easy to combine it in the future with plugins that would carry out the second stage of anonymization or pseudonymization (i.e. remove or replace the personal information in the text), effectively completing the de-identification pipeline, which we consider as our future goals.

We have also introduced the (Im)Personal Data visualization tool which can be used to illustrate and qualitatively analyze the output of our plugin and other PI detection and labeling models, as well as to visualize the extent of research data that needs to be manipulated before it is shared with other researchers. This interface, which currently operates by displaying static files, has the potential to be expanded with a backend to display the performance of specific models on the go.

We have also performed a case study to demonstrate our plugin's performance and the usefulness of the (Im)Personal Data visualization tool. We have shown that our plugin performs noticeably better on the text we tested than one of the most popular openly available tools, Microsoft Presidio. It can detect a wider range of personal information than the HB Deid tool, and seems to be less prone to false positives than the LLM Gemma 2 9B, in comparison to which it is also much faster and less resource-intensive.

We believe that our plugin and the visualization tool can contribute to language resource construction by facilitating the de-identification procedures, indirectly contributing to research in a variety of fields which rely on such data, including but not limited to linguistics, Digital Humanities, or Natural Language Processing. At the same time, our tools will hopefully enable further research on NLP methods for anonymization and pseudonymization.

The future directions include, among others, (1) developing models for pseudonymization step (i.e. for the generation of replacement equivalents for the detected spans), and (2) developing solutions that would offer an option for customizing which subset of PI categories to replace, thus protecting research data from being over-manipulated. The two future steps address the duality of the problem: ethical requirement to protect people in the data versus the need of valid research data that is as close to the original as it is legally and ethically allowed.

Acknowledgments

This work was possible thanks to the funding of several grants from the Swedish Research Council. It has been funded by Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2025–2028; project id 2023-00161). It has also been funded by the research environment project *Grandma Karl is 27 years old: Automatic pseudonymization of research data* (2023-2029, funding number 2022-02311). The authors acknowledge support from InfraVis for providing application expertise for visualization through the Swedish Research Council grant 2021-00181.

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.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM.

Hanna Berg and Hercules Dalianis. 2019. Augmenting a de-identification system for Swedish clinical text using open resources and deep learning. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 8–15, Turku, Finland. Linköping Electronic Press.

Hanna Berg and Hercules Dalianis. 2020. A semi-supervised approach for de-identification of Swedish clinical text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4444–4450, Marseille, France. European Language Resources Association.

Hanna Berg and Hercules Dalianis. 2021. HB Deid - HB de-identification tool demonstrator. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 467–471, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Rogier Blokland, Niko Partanen, and Michael Rießler. 2020. A pseudonymisation method for language documentation corpora: An experiment with spoken Komi. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 1–8, Wien, Austria. Association for Computational Linguistics.

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 rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 16–23, Turku, Finland. Linköping Electronic Press.

Tobias Deußer, Max Hahnbück, Tobias Uelwer, Cong Zhao, Christian Bauckhage, and Rafet Sifa. 2025. Resource-efficient anonymization of textual data via knowledge distillation from large language models. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 243–250, Abu Dhabi, UAE. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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 Krieg-Holz, 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.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A.

Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size.

Martin Hammarstedt, Anne Schumacher, Lars Borin, and Markus Forsberg. 2022. Sparv 5 user manual. Technical report, Institutionen för svenska, flerspråkighet och språkteknologi, Göteborgs universitet, Göteborg.

Nikolai Ilinykh and Maria Irena Szawerna. 2025. "I need more context and an English translation": Analysing how LLMs identify personal information in Komi, Polish, and English. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 165–178, Tallinn, Estonia. University of Tartu Library, Estonia.

Kristian Nørgaard Jensen, Mike Zhang, and Barbara Plank. 2021. De-identification of privacy-related entities in job postings. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 210–221, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Alistair E. W. Johnson, Lucas Bulgarelli, and Tom J. Pollard. 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, pages 214–221, New York, NY, USA. Association for Computing Machinery.

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.

Guillermo López-García, Francisco J. Moreno-Barea, Héctor Mesa, José M. Jerez, Nuria Ribelles, Emilio Alba, and Francisco J. Veredas. 2023. Named Entity Recognition for De-identifying Real-World Health Records in Spanish. In *Computational Science – ICCS 2023*, pages 228–242, Cham. Springer Nature Switzerland.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT.

Christopher Meaney, Wali Hakimpour, Sumeet Kalia, and Rahim Moineddin. 2022. A Comparative Evaluation Of Transformer Models For De-Identification Of Clinical Text Data. ArXiv:2204.07056 [cs, stat].

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 from the 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.

Phuong Ngo, Miguel Tejedor, Therese Olsen Svenning, Taridzo Chomutare, Andrius Budrionis, and Hercules Dalianis. 2024. Deidentifying a Norwegian clinical corpus - an effort to create a privacy-preserving Norwegian large clinical language model. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 37–43, St. Julian's, Malta. Association for Computational Linguistics.

Jaume Nualart and Mario Pérez-Montoro. 2013. Texty, a visualization tool to aid selection of texts from search outputs. *Information Research*, 18(2).

Official Journal of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). *Official Journal*, (Document 02016R0679-20160504).

Maria Irena Szawerna, Simon Dobnik, Ricardo Muñoz Sánchez, Therese Lindström Tiedemann, and Elena Volodina. 2024. Detecting personal identifiable information in Swedish learner essays. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 54–63, St. Julian's, Malta. Association for Computational Linguistics.

Maria Irena Szawerna, Simon Dobnik, Ricardo Muñoz Sánchez, and Elena Volodina. 2025. The devil's in the details: the detailedness of classes influences personal information detection and labeling. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 697–708, Tallinn, Estonia. University of Tartu Library.

Elena Volodina, Yousuf Ali Mohammed, Sandra Derbring, Arild Matsson, and Beata Megyesi. 2020. Towards privacy by design in learner corpora research: A case of on-the-fly pseudonymization of Swedish learner essays. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 357–369, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jianliang Yang, Xiya Zhang, Kai Liang, and Yuenan Liu. 2023. Exploring the application of large language models in detecting and protecting personally identifiable information in archival data: A comprehensive study*. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2116–2123.

Oleksandr Yermilov, Vipul Raheja, and Artem 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)*, pages 232–241, Toronto, Canada. Association for Computational Linguistics.