Birds of a Feather: Do Embedding Representations of Personal Information Flock Together?

Maria Irena Szawerna[1], Simon Dobnik[2]
[1]Språkbanken Text, SFS, University of Gothenburg
Gothenburg, Sweden
[2]CLASP, FLoV, University of Gothenburg
Gothenburg, Sweden
{maria.szawerna,simon.dobnik}@gu.se

Abstract

Personally identifiable information (PII or PI) can appear in a wide variety of linguistic data, posing both ethical and legal challenges for conducting research and developing applications involving such texts. In this paper, we investigate the alignment between automatic clustering of FastText and Transformer embedding representations of personal information spans sourced from essays written by adult learners of Swedish as a second language and the general and detailed personal information labels assigned to these spans by expert annotators. Our goals are to assess the extent of overlap between the semantic categories and evaluate the semantic coherence of the human-assigned classes, which may have implications for de-identification procedures. We observe that while contextual embeddings, especially ones from a specialized word-in-context model, produce relatively good clustering results, they only partly map to the human understanding of how to classify personal information.

Keywords: personal information, PII, de-identification, pseudonymization, anonymization, clustering

1. Introduction and Prior Research

The presence of personally identifiable information (PII, PI)[1] in language data poses undeniable ethical and legal challenges. There is a need for the development of tools aimed at automatizing the time-consuming task of personal information detection, followed by redaction or labeling and replacement. PI detection and labeling is a ubiquitous step in Named Entity Recognition-like approaches to de-identification of such data (Lison et al., 2021; Volodina et al., 2025). Many such approaches rely on (contextual) embedding representations of the tokens in the text to carry out the classification (cf. Grancharova and Dalianis, 2021 or Pilán et al., 2022), as rule-based methods can only capture some PI types that show less diversity in terms of form (Volodina et al., 2020). However, it has been shown that such systems can be sensitive to how internally consistent personal information classes are (Sierro et al., 2024; Szawerna et al., 2025).

In this paper, we set out to address the question to what degree do embedding representations of PI words and phrases capture the semantic knowledge of humans, where that semantic knowledge is approximated by a PI taxonomy developed by human annotators (i.e., division of spans identified as PI into classes salient for humans)? Our goal is to improve the semantic understanding of labels used in PI annotation and to assess how representations used by models align with a human-devised taxonomy, which could help improve both PI detection methods and the taxonomies themselves. Investigating the semantic alignment between humans and language models in this specific domain may also hint at some more general patterns. In that sense, our work is reminiscent of language games pioneered by Steels and Belpaeme (2005), where they evaluate the similarity between natural language categories and categories in a an automatically-induced language emergent from situational grounding of two artificial agents.

We address our research question through clustering embedding representations of personal information. Clustering embeddings to understand the distributional properties of language data has previously been employed by e.g. Hertzberg et al. (2022) in the domain of political dogwhistles; while there are several differences in our approaches, the main one is our comparison being conducted against ground truth labels, whereas theirs tries to determine whether the successfulness of clustering correlates with inter-annotator agreement.

2. Materials

In our experiments, we use 947 texts (totaling 301095 tokens) from SweLL-pilot and SweLL-gold (Volodina, 2024; Språkbanken Text, 2025b).[2] These two SweLL corpora are collections of essays written by learners of Swedish as a second language (L2). Many of these texts contain various kinds of PI, which are pseudonymized in the

[1]Henceforth often simply 'personal information.'

[2]SweLL access can be requested at `https://sunet.artologik.net/gu/swell`

released versions of the corpora; however, we use the essays with the original PI intact.

The PI spans in the SweLL data are annotated with PI categories (see Megyesi et al. (2021) and Volodina et al. (2020)). This taxonomy is hierarchical, with 7 overarching general categories and 37 possible detailed PI categories. For instance, in the fictitious example of *mitt namn är Sonja och jag är 29* 'my name is Sonja and I am 29', *Sonja* would be labeled by an expert annotator as the detailed class `firstname_female` (which belongs to the general category `personal_name` together with surnames, masculine names, etc.), and *29* would be marked as `age_digits` (belonging to the general category `age`). In our data only 32 of those detailed categories are present.[3] Both singular tokens and multi-word expressions may be annotated as PI; 3348 tokens constituting 3076 spans are annotated as PI in our data.[4] As that table shows, some classes are much more frequent in the data than others, which is likely to negatively affect the discriminability of the infrequent classes.

As we are working with Swedish data, we chose three models trained for this language to obtain embedding representations of the PI spans: I) `Kubord-fasttext - Dagens Nyheter 2010-2024 - token` (Språkbanken Text, 2025a): one of the FastText embedding models for Swedish (see Bojanowski et al. (2016)). Embedding size: 300. Henceforth FastText; II) `KB/bert-base-swedish-cased` (Malmsten et al., 2020): the Swedish version of the original BERT model (Devlin et al., 2019). Embedding size: 768. Henceforth KB-BERT; III) `pierluigic/xl-lexeme` (Cassotti et al., 2023): a specialized multilingual word-in-context (WiC) model based on XLM-RoBERTa-large (Conneau et al., 2020). Embedding size: 1024. Henceforth XL-LEXEME.

The FastText embeddings serve as a non-contextual baseline. KB-BERT has previously been used in many token classification tasks for Swedish, including PI detection and labeling applications (e.g., by Grancharova and Dalianis (2021), Vakili et al. (2022), or Szawerna et al. (2024)). XL-LEXEME belongs to a similar language model category as KB-BERT, but as it is specialized for word-in-context tasks, it may capture more of the nuances of personal information. Embeddings for each token or subword token in a PI span were obtained from each model. Maximum input size of 100 KB-BERT subword tokens was used for the masked language models to ensure that a comparable context was provided for the phrase in question. For KB-BERT, the last-layer representations were obtained, as those are more sensitive to semantics and context (Jawahar et al., 2019). In the cases of multi-token spans, a mean of the embeddings was obtained for FastText and BERT to preserve dimensionality; it was possible to directly obtain an embedding for the whole span from XL-LEXEME. These three sets of embeddings will henceforth be referred to as embedding types.

3. Methods

In order to evaluate the alignment between the embeddings of different PI spans and the human-assigned labels, we perform automatic clustering on the embeddings. We first reduce the embedding size using Uniform Manifold Approximation and Projection (UMAP, McInnes et al., 2020). This step already helps capture the underlying patterns and speeds up computation. We then perform a parameter search for four clustering algorithms: Hierarchical Density-Based Clustering (Campello et al., 2013), Affinity Propagation (Frey and Dueck, 2007), Mean Shift (Fukunaga and Hostetler, 1975), and Agglomerative Clustering,[5] in their scikit-learn implementations (Pedregosa et al., 2011). These four algorithms all permit varied cluster size, which is essential given the uneven distribution of human-annotated PI classes in our data. We evaluate the intrinsic quality of the clustering using the silhouette score (Rousseeuw, 1987) (which is a measure of how well data points fit their clusters and how well-bounded those clusters are on a scale between -1 and 1) and select the best clustering algorithm and parameters for each embedding type.

We calculate extrinsic measures for the selected results, comparing the emergent clusters to the human annotation. We focus on completeness (data points from one ground truth class being grouped in one cluster), homogeneity (the internal purity of clusters relative to the ground truth), and the combined v-score (Rosenberg and Hirschberg, 2007) as interpretable measures of specific properties of the clustering relative to the human annotation at both the detailed and general label level. We consider homogeneity to be more important than the other two in understanding how the machine clustering relates to the human-identified classes; it is clear that completeness will be much lower in clustering outcomes which result in hundreds of clusters, but as long as those are internally ho-

[3] `initials, area, url, personid_nr, account_nr, license_nr` are absent.

[4] Table 2 in Appendix A shows the detailed counts of the annotated tokens and phrases alongside information as to which detailed categories correspond to which general ones. The latter is also explained better in Megyesi et al. (2021); Volodina et al. (2020); Szawerna et al. (2025).

[5] To the best of our knowledge, there is no single citation for hierarchical agglomerative clustering, and only various linkage methods have standard references, see Müllner (2011).

Embedding	Clustering	Homogeneity	Completeness	V-score	N clusters	N outliers
FastText	HDBSCAN	0.69 (0.67)	0.44 (0.24)	0.54 (0.35)	66	1132
KB-BERT	HDBSCAN	0.72 (0.84)	0.53 (0.35)	0.61 (0.50)	41	1304
XL-LEXEME	HDBSCAN	0.77 (0.86)	0.50 (0.31)	0.60 (0.46)	70	561

Table 1: Metrics per embedding type for the best clustering results. Scores in black are compared against human-assigned detailed labels, whereas the (gray scores) are relative to the general labels.

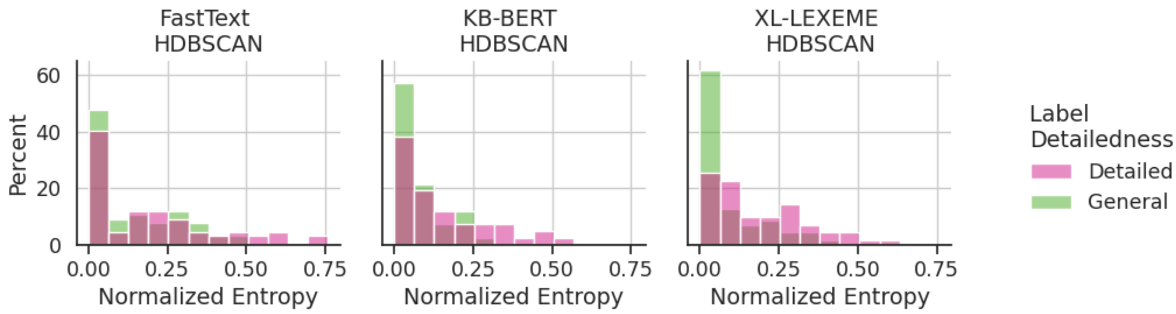


Figure 1: Histograms of entropy distribution, in percent for comparability across embedding types.

mogeneous, one can conclude that the clustering simply splits a human-assigned category into even more granular ones. Additionally, we calculate entropy (Shannon, 1948) per cluster and normalize it[6] to further inspect how pure the specific clusters are, analogous to how Dobnik and Kelleher (2013) or Dobnik and Kelleher (2014) use this measure to assess the purity of semantic categories against a set of labels.

4. Results and Discussion

The best silhouette scores were obtained for all embedding types by the HDBSCAN algorithm when the outlier category that it predicts was excluded from the calculation (ilhouette score of 0.83 for FastText, 0.67 for KB-BERT, and 0.72 for XL-LEXEME).[7]

Given that the silhouette score ranges from -1 (very bad) to 1 (perfect), these scores are good, and it is unsurprising to see clustering algorithms that eliminate outliers perform well on the intrinsic metric. However, as shown in Table 1, between 18 and 42% of the data was excluded as outliers, indicating that a large part of the data is hard to cluster cleanly. Across all embedding types, nearly all detailed PI categories are represented among the outliers. When inspecting the items identified as outliers, some trends can be noted, such as classes that are generally infrequent being more likely to have a large proportion of outliers, personal names and dates being hard to cluster with FastText embeddings, or KB-BERT struggling with geographic and transportation classes.[8] Results for XL-LEXEME stand out here, with the lowest number of embeddings that are treated as outliers and a high homogeneity score. While the KB-BERT embeddings result in a large number of outliers, the number of detected clusters is the closest to the number of detailed labels in the human taxonomy and the lowest out of the three outcomes. Finally, FastText embeddings result in noticeably worse results than the contextual embeddings.

An interesting, but not entirely unexpected, observation can be made by comparing the scores relative to the detailed and general human-assigned labels. Overall, completeness and v-score are lower when the comparison is made to the general classes, as the number of clusters is always much larger than the 7 general classes, meaning that multiple clusters will consist of examples from one such class. Homogeneity improves noticeably for contextual embeddings when the comparison is made to general human-assigned labels instead of detailed. This indicates that even though not all clusters are pure, elements that belong to different detailed classes but the same overarching general classes are still grouped together. This does not hold for the FastText embeddings, implying that the clusters there have more random impurities.

This is further corroborated by the per-cluster entropy scores pictured in the histograms in Figure 1.[9]

[6]Entropy of a cluster X here is defined as $H(X) = -\sum_{x \in X} p(x) log_2 p(x)$ and it is normalized by the maximum possible entropy for the cluster, i.e. $-log_2(|X|)$.

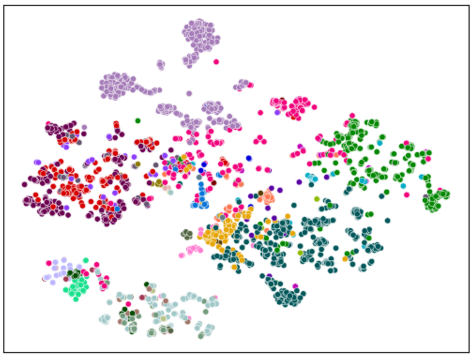
[7]Parameter details can be found in Appendix A.

[8]The detailed counts can be found in Appendix A.

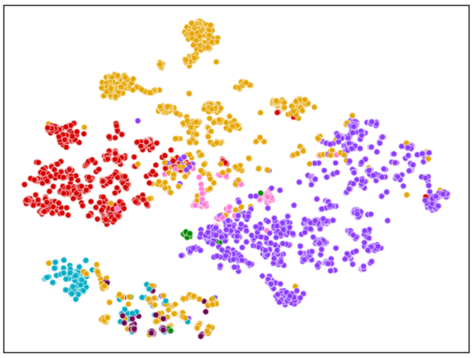
[9]Visualized using pandas (pandas development team, 2020), matplotlib (Hunter, 2007), and seaborn (Waskom, 2021)

A cluster being perfectly homogenous relative to ground truth means it has an entropy of 0, whereas an entropy of 1 means a maximally random assortment of human-assigned labels in the cluster. For FastText embeddings, there are relatively minor differences between the entropy scores for detailed and general labels. What can be noticed is that comparing to general labels leads to a small increase in the lower entropy scores, whereas comparing to detailed labels is what is responsible for entropy scores above 0.5. A similar pattern occurs in the case of the contextual embeddings, but the differences are more pronounced, especially in the case of XL-LEXEME, where the percentage of near-zero entropy clusters skyrockets when the comparison is made to general labels. This indicates that while the XL-LEXEME embeddings appear to be the best for clustering PI (with a relatively small number of outliers and high homogeneity), they do not permit the same fine-grained distinctions as the human annotation and instead group the information differently at that level of detailedness, though within the same general classes.

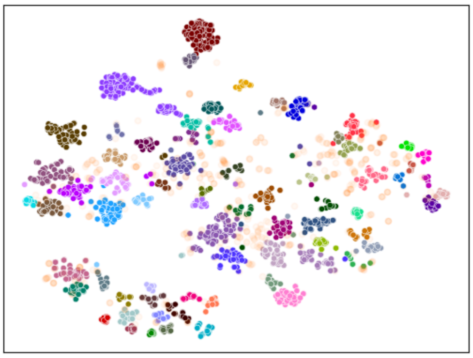
This can be visualized by reducing the dimensionality of the embeddings to 2 using UMAP and plotting the datapoints. Figure 2 shows this data for XL-LEXEME embeddings, colored according to detailed and general human annotation (Figure 2a, Figure 2b) and with the clusters assigned by HDBSCAN (Figure 2c). While the correspondence between colors and actual labels is not provided due to the number of labels, the differences between detailed human labels and the HDBSCAN clusters are quite apparent.[10] These results indicate that while some semantic similarity is captured by clustering, the original classes become fragmented. This could perhaps be mitigated both by tweaking how the embeddings are obtained (which layer, what context window) and by what the ground truth reference is.



(a) Detailed human labels



(b) General human labels



(c) HDBSCAN labels

Figure 2: XL-LEXEME scatterplots. Due to the number of classes the legend is not provided. For HDBSCAN, the outliers are marked with transluscent orange points.

5. Conclusions and Future Work

In this paper, we explored the effectiveness of automatic clustering of authentic PI spans from Swedish texts, represented using three different types of embeddings, in order to increase our understanding of the semantics of PI labels and their alignment with computational representations. We observed that in both non-contextual and contextual embeddings, a certain number of PI instances are hard to cluster, but a specialized word-in-context model struggled less with this issue. Clustering algorithms tend to identify more clusters than the human-assigned detailed PI classes. The boundaries for those clusters sometimes align with the human annotation, but that depends on the embedding and annotation type. Impurities in clusters identified for the contextual embeddings tend to stem from semantically similar concepts being grouped together (e.g., different types of geographical information).

In the future, it could be interesting to use this approach to try to identify which models' representations (and from which layers) are sensitive to the differences between PI types and non-personal information with the goal of establishing which models are worth trying to train for PI detection and labeling, and whether the performance in experiments such as ours correlates with that and what

[10] See Appendix A for plots for FastText and KB-BERT.

effect model fine-tuning has on these representations. Further investigating what sets the separate clusters containing the same human-annotated class apart could be an interesting way to potentially help refine the taxonomy used for annotating PI. Comparing which human-assigned labels have the lowest inter-annotator agreement and which kinds of personal information are the hardest to cluster could bring more nuance to an analysis of this type. Finally, semantic relatedness between various PI clusters could perhaps be exploited in studies on semantic coherence of pseudonyms used to replace personal information.

Limitations

A natural limitation of this research is that it is conducted only on one genre of texts. However, to the best of our knowledge, there exists no other PI-annotated corpus in Swedish or another corpus annotated with the same categories as SweLL that is possible for us to access, which would be a very valuable comparison allowing us to generlize our observations about the nature of personal information. Similarly, the use of authentic PI data severely limits the reproducibility of this study; however, it shows our compliance with legal and ethical standards. We believe that this methodology can be successfully applied to any other PI-annotated dataset.

Another limitation is that this comparison only includes three models from which embeddings are obtained. While still fewer than for some other languages, there are many models that can, to some extent, handle Swedish text and that could be included in a larger-scale comparison.

Our experiment does not clearly assess the usefulness of the embedding representations for PI detection (i.e., telling it apart from the surrounding non-personal context), but only for its subsequent classification.

Since a part of the representations stand for multiword expressions, the way in which they are calculated (a mean of the constituent embeddings for FastText and KB-BERT) could make them harder to cluster and result in them being rejected as outliers.

Any more qualitative analysis of the purity of the identified clusters was hindered by the sheer number of clusters and the fact that we were comparing three such results.

Ethical Considerations

Research about PI and de-identification is, in large part, fueled by ethical considerations and legal requirements when it comes to processing language data. Continued exploration of such questions can contribute to a better understanding of the effectiveness and the consequences of de-identification, as well as help improve the methods employed; in the case of this paper, it could inform the choice of model for PI detection tasks and perhaps assist with the development of refined PI taxonomies.

As we are using authentic PI, which is not available in the current release of the corpus that we are working with, we cannot go into a too detailed analysis of clusters (we cannot provide specific, authentic PI span examples), nor can we share the data or the embeddings used in the analysis. We assess the risks of information leakage from the results that we provide to be low, as they are only shown aggregated and without any references back to the texts that the phrases are extracted from, and all the experiments were conducted locally.

6. Bibliographical References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

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.

Simon Dobnik and John D. Kelleher. 2013. Towards an automatic identification of functional and geometric spatial prepositions. In *Proceedings of PRE-CogSsci 2013: Production of referring expressions - bridging the gap between cognitive and computational approaches to reference*.

Simon Dobnik and John D. Kelleher. 2014. Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In *Proceedings of the Third Workshop on Vision and Language*, pages 33–37, Dublin, Ireland. Dublin City University and the Association for Computational Linguistics.

Brendan J. Frey and Delbert Dueck. 2007. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976.

Keinosuke Fukunaga and Larry Hostetler. 1975. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40.

Mila Grancharova and Hercules Dalianis. 2021. Applying and Sharing pre-trained BERT-models for Named Entity Recognition and Classification in Swedish Electronic Patient Records. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 231–239, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Niclas Hertzberg, Robin Cooper, Elina Lindgren, Björn Rönnerstrand, Gregor Rettenegger, Ellen Breitholtz, and Asad Sayeed. 2022. Distributional properties of political dogwhistle representations in Swedish BERT. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 170–175, Seattle, Washington (Hybrid). Association for Computational Linguistics.

John D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

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.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Beáta Megyesi, Lisa Rudebeck, and Elena Volodina. 2021. Swell pseudonymization guidelines. *GU-ISS Forskningsrapporter från Institutionen för svenska, flerspråkighet och språkteknologi*, GU-ISS 2021-02.

Daniel Müllner. 2011. Modern hierarchical, agglomerative clustering algorithms.

The pandas development team. 2020. pandas-dev/pandas: Pandas.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. *Computational Linguistics*, 48(4):1053–1101.

Andrew Rosenberg and Julia Hirschberg. 2007. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic. Association for Computational Linguistics.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Claude E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

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 Two 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.

Luc Steels and Tony Belpaeme. 2005. Coordinating perceptually grounded categories through language: a case study for colour. *Behavioral and Brain Sciences*, 28(4):469–489.

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.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252, Marseille, France. European Language Resources Association.

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.

Elena Volodina, Simon Dobnik, Therese Lindström Tiedemann, Ricardo Muñoz Sánchez, Maria Irena Szawerna, and Lisa Södergård. 2025. Towards Shared Standards for Pseudonymization of Research Data. In *Proceedings of the 2nd Huminfra Conference*.

Michael L. Waskom. 2021. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

7. Language Resource References

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

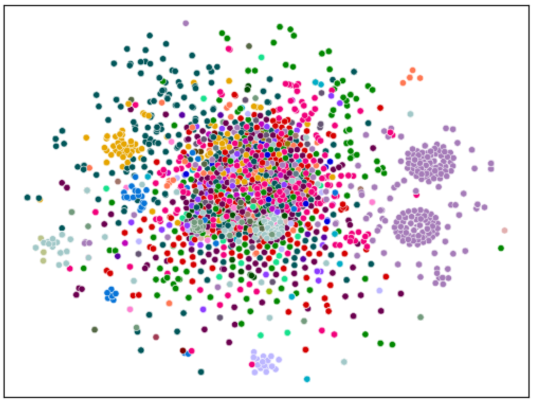
Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT.

Språkbanken Text. 2025a. Kubord-fasttext - dagens nyheter 2010–2024 - token.

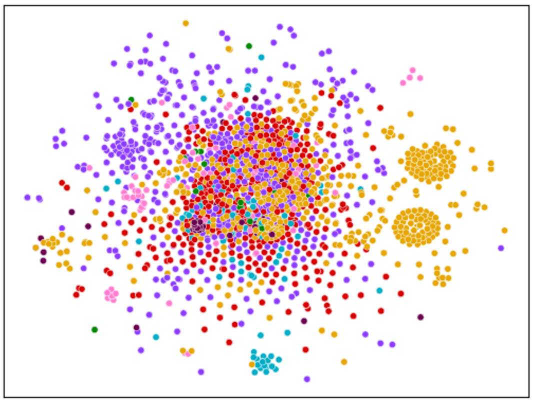
Språkbanken Text. 2025b. Swell.

Elena Volodina. 2024. On two swell learner corpora – swell-pilot and swell-gold. In *Proceedings of the Huminfra Conference (HiC 2024), 10-11 January, 2024, Gothenburg, Sweden*, pages 83–94, Linköping. Linköping University Press.

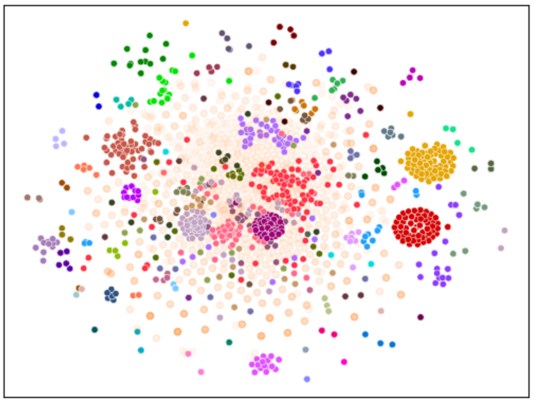
A. Appendix



(a) Detailed human labels

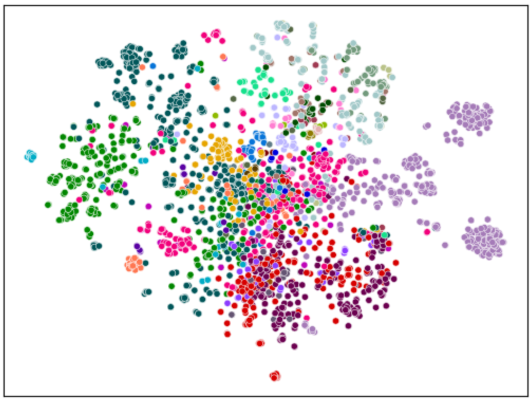


(b) General human labels

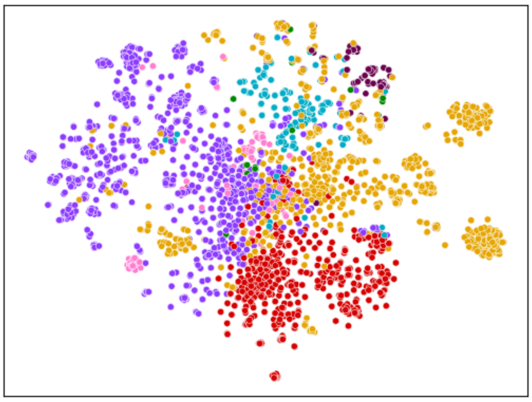


(c) HDBSCAN labels

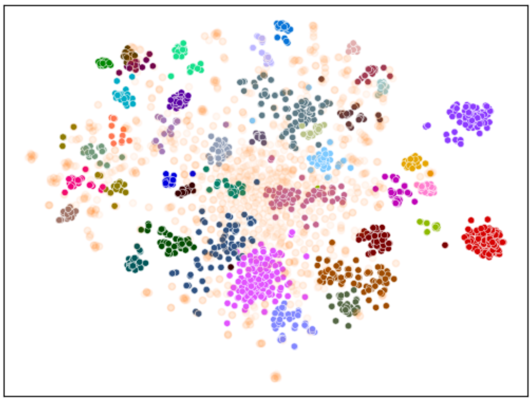
Figure 3: FastText scatterplots. Due to the number of classes the legend is not provided. For HDBSCAN, the outliers are marked with transluscent orange points.



(a) Detailed human labels



(b) General human labels



(c) HDBSCAN labels

Figure 4: KB-BERT scatterplots. Due to the number of classes the legend is not provided. For HDBSCAN, the outliers are marked with transluscent orange points.

Category	Tokens	Phrases
personal_name	**624**	**612**
firstname_male	234	228
firsname_female	289	287
firstname_unknown	49	47
middlename	1	1
surname	51	49
geographic	**1186**	**1135**
city	587	561
geo	17	17
country	401	399
place	112	94
region	39	36
street_nr	21	21
zip_code	9	7
institution	**160**	**111**
school	69	44
work	2	2
other_institution	89	65
transportation	**20**	**19**
transport_name	6	5
transport_nr	14	14
age	**94**	**94**
age_digits	82	82
age_string	12	12
date	**179**	**165**
day	27	27
month_digit	9	9
month_word	46	46
year	53	53
date_digits	44	30
other	**1085**	**940**
phone_nr	7	6
email	10	10
other_nr_seq	170	168
extra	40	32
prof	14	12
edu	7	5
fam	467	453
sensitive	370	254
TOTAL	**3348**	**3076**

Table 2: Token and phrase (MWE) counts for the PII spans in our data. General categories, given in **bold**, appear above the corresponding detailed labels.

Embedding	**Clustering**	**Best parameters**	**Silhouette**
FastText	HDBSCAN	cluster_selection_method='leaf', min_cluster_size=12	0.83
KB-BERT	HDBSCAN	cluster_selection_method='leaf', min_cluster_size=17	0.68
XL-LEXEME	HDBSCAN	cluster_selection_method='eom', min_cluster_size=14	0.72

Table 3: Best clustering algorithm and parameters per embedding type.

CATEGORY	FASTTEXT	KB-BERT	XL-LEXEME
personal_name	**356 (58.17%)**	**216 (35.29%)**	**87 (14.22%)**
firstname_male	135 (59.21%)	105 (46.05%)	39 (17.11%)
firsname_female	159 (55.40%)	75 (26.13%)	23 (8.01%)
firstname_unknown	28 (59.57%)	22 (46.81%)	11 (23.40%)
middlename	1 (100.00%)	-	-
surname	33 (67.35%)	14 (28.57%)	14 (28.57%)
geographic	**327 (32.78%)**	**594 (52.33%)**	**246 (21.67%)**
city	166 (29.59%)	273 (48.66%)	88 (15.69%)
geo	11 (64.71%)	12 (70.59%)	5 (29.41%)
country	123 (30.89%)	211 (52.88%)	118 (29.57%)
place	40 (42.55%)	51 (54.26%)	17 (18.09%)
region	19 (52.78%)	27 (75.00%)	10 (27.78%)
street_nr	10 (47.62%)	17 (80.95%)	6 (28.57%)
zip_code	3 (42.86%)	3 (42.86%)	2 (28.57%)
institution	**42 (37.84%)**	**49 (44.14%)**	**22 (19.82%)**
school	13 (29.55%)	23 (52.27%)	4 (9.09%)
work	2 (100.00%)	2 (100.00%)	1 (50.00%)
other_institution	27 (41.45%)	24 (36.92%)	17 (26.15%)
transportation	**9 (47.37%)**	**16 (84.21%)**	**9 (47.37%)**
transport_name	4 (80.00%)	5 (100.00%)	1 (20.00%)
transport_nr	5 (35.71%)	11 (78.57%)	8 (57.14%)
age	**21 (22.34%)**	**33 (35.11%)**	**3 (3.19%)**
age_digits	21 (25.61%)	26 (31.71%)	3 (3.66%)
age_string	-	7 (58.33%)	-
date	**81 (49.09%)**	**57 (34.55%)**	**41 (24.85%)**
day	13 (48.15%)	7 (25.93%)	7 (25.93%)
month_digit	4 (44.44%)	4 (44.44%)	3 (33.33%)
month_word	14 (30.43%)	23 (50.00%)	21 (45.65%)
year	31 (58.49%)	17 (32.08%)	9 (16.98%)
date_digits	19 (63.33%)	6 (20.00%)	1 (3.33%)
other	**251 (26.70%)**	**339 (36.06%)**	**153 (16.28%)**
phone_nr	2 (33.33%)	-	-
email	5 (50.00%)	-	1 (10.00%)
other_nr_seq	28 (16.67%)	92 (54.76%)	14 (8.33%)
extra	16 (50.00%)	17 (53.12%)	13 (40.62%)
prof	9 (75.00%)	3 (25.00%)	8 (66.67%)
edu	4 (80.00%)	5 (100.00%)	3 (60.00%)
fam	63 (13.91%)	83 (18.32%)	24 (5.30%)
sensitive	124 (48.82%)	139 (54.72%)	90 (35.43%)

Table 4: Outlier counts per embedding type (for its best associated clustering method) by original human-assigned class. The value in brackets is the % of all the phrases of this type that the outliers constitute. General classes are given in bold.