The Devil's in the Details: the Detailedness of Classes Influences Personal Information Detection and Labeling

Maria Irena Szawerna[†], **Simon Dobnik**[‡], **Ricardo Muñoz Sánchez**[†], **Elena Volodina**[†]
[†]Språkbanken Text, SFS, University of Gothenburg, Sweden
[‡]CLASP, FLoV, University of Gothenburg, Sweden
mormor.karl@svenska.gu.se
[†]{maria.szawerna,ricardo.munoz.sanchez,elena.volodina}@gu.se
[‡]simon.dobnik@gu.se

Abstract

In this paper, we experiment with the effect of different levels of detailedness or granularity — understood as i) the number of classes, and ii) the classes' semantic depth in the sense of hypernym and hyponym relations — of the annotation of Personally Identifiable Information (PII) on automatic detection and labeling of such information. We fine-tune a Swedish BERT model on a corpus of Swedish learner essays annotated with a total of six PII tagsets at varying levels of granularity. We also investigate whether the presence of grammatical and lexical correction annotation in the tokens and class prevalence have an effect on predictions. We observe that the fewer total categories there are, the better the overall results are, but having a more diverse annotation facilitates fewer misclassifications for tokens containing correction annotation. We also note that the classes' internal diversity has an effect on labeling. We conclude from the results that while labeling based on the detailed annotation is difficult because of the number of classes, it is likely that models trained on such annotation rely more on the semantic content captured by contextual word embeddings rather than just the form of the tokens, making them more robust against nonstandard language.

1 Introduction

Personal information is ubiquitous in many text genres, posing a unique challenge for those seeking to create and share corpora. While access to collections of texts is highly desirable from the perspective of researchers in fields such as linguistics, Natural Language Processing (NLP), or digital humanities, the potential presence of clues indicating the identity of the writer or other natural persons makes them fall under the General Data Protection Regulation (GDPR, Official Journal of the European Union, 2016). The GDPR itself suggests potential solutions to the problem: de-identification methods such as anonymization — the "[c]omplete and irreversible removal [...] of any information that, directly or indirectly, may lead to a subject's data being identified" — or pseudonymization, the "[p]rocess of replacing direct identifiers with pseudonyms or coded values," for which there must exist a mapping between the original data and the pseudonyms, which is securely stored separately from the pseudonymized texts (Lison et al., 2021).

Both of these privacy-preserving procedures presuppose a stage where the Personally Identifiable Information (PII) found in the data is detected. While this can be done manually, it is time-consuming. While automatic approaches for both anonymization and pseudonymization have been proposed (Lison et al., 2021), Szawerna et al. (2024a) show that there appears to be very little uniformity in how researchers and corpus creators choose to classify PIIs. The taxonomies range in terms of granularity or detailedness, understood as the number of classes that PIIs are divided into and their semantic depth in terms of hypernym and hyponym relations (as in WordNet (Miller, 1995)). For example, Pilán et al. (2022) utilize only one label, `PERSON`, to refer to elements such as names, surnames, nicknames, usernames, etc., which can be differentiated in other corpora (e.g. Volodina et al. 2016, 2019; Eder et al. 2020; Alfalahi et al. 2012). Very little work has been done on determining what level of granularity of PII annotation is the most suitable for subsequent removal or replacement of personal information.

It is worth noting that while the term *detection* often includes labeling in other research on

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
March 3-4, 2025

General category	Corresponding detailed categories
personal_name	firstname_male, firstname_female, firstname_unknown, initials, middlename, surname
institution	school, work, other_institution
geographic	area, city, geo, country, place, region, street_nr, zip_code, ~~foreign~~
transportation	transport_name, transport_nr
age	age_digits, age_string
date	date_digits, day, month_digit, month_word, year
other	phone_nr, email, url, personid_nr, account_nr, license_nr, other_nr_seq, extra, prof, edu, fam, sensitive, ~~gen~~, ~~def~~, ~~pl~~

Table 1: General and detailed categories in the SWELL PII taxonomy. Tags that can be combined with other categories and therefore were not included in the experiments are crossed out.

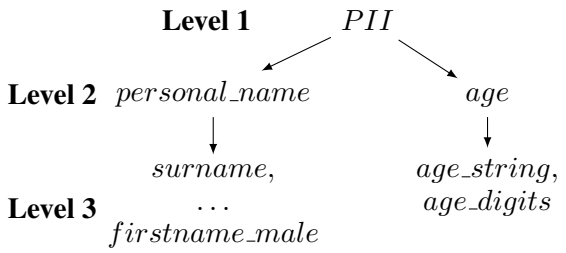

Figure 1: Hierarchical ontological structure of PII categories on the example of selected SWELL categories.

this topic, we choose to differentiate between the two: *PII detection* is the process of determining whether a text span constitutes a piece of Personally Identifiable Information, while *PII labeling* is assigning a PII span a specific class which describes the type of PII it is (this procedure often, by default, detects and assigns a specific PII class at the same time).

In this paper, we set out to investigate what the effect of the class granularity on PII detection and labeling in the learner essay domain. We run our experiments on a set of Swedish texts that are PII-annotated at varying levels of detailedness. A related notion is that of the categories' ontological structure. As shown in Figure 1, the categories used in this experiment can be hierarchically arranged from the most general (level 1) to the most specific (level 3). Simultaneously, e.g. level 2 categories are semantically broader (include more semantically varied elements) than the more specific level 3 categories. How varied the contents of a category are could have an impact on how easy it is to automatically detect. While we make an initial assumption that having a larger number of more specific labels means that they will be less internally diverse, labels in one tagset are not necessarily equally internally coherent.

In addition, we are curious to see how various factors pertaining to the class divisions (e.g. the class's frequency) or the word tokens themselves (e.g. being ungrammatical) influence the performance. While improvement in terms of PII detection on the data with more specific annotation relative to the general one has been previously observed (Sierro et al., 2024), we expect multi-class classification to be more prone to error.

2 Prior Research

Data for research or training language models needs to be free from personal information to protect those who generate it, and the work on automatic de-identification methods, especially for texts belonging to domains other than medical or legal, has gained much traction in the recent years (Lison et al., 2021).

Much research has gone into testing what kinds of models perform best for PII detection or labeling. Eder et al. (2022) evaluate 9 different model architectures and embedding combinations on the PII-annotated corpus of German e-mails, CODE ALLTAG, reaching the best performance with a Transformer-based architecture and embeddings, optionally combined with noncontextual word embeddings. Papadopoulou et al. (2022, 2023) successfully utilize a combination of a generic Named Entity Recognition (NER) model with a gazetteer to detect and classify PIIs in English (the TAB CORPUS and a set of annotated

Wikipedia biographies) and employ privacy risk estimation methods to determine whether a span should be anonymized or not. Grancharova and Dalianis (2021) frame the closely related task of Protected Health Information (PHI) detection as a Named Entity Recognition and Classification (NERC) task and obtain good results on it using two BERT-type language models on Swedish medical data from the STOCKHOLM EPR PHI CORPUS. Szawerna et al. (2024b) also use models of this kind to detect PIIs in the SWELL corpus, a collection of learner essays in Swedish. Notably, they forego the labeling step, differentiating only between PII and non-PII tokens.

It is worth noting that all of the previously mentioned PII or PHI detection or labeling studies utilized different data, and only the texts used by Papadopoulou et al. (2022, 2023) — representing a vastly different domain and a more general tagset than the texts we work with — are openly available with the original PIIs in place. Additionally, all of the papers employed different categories for the labeling task. As Szawerna et al. (2024a) point out, differences between PII taxonomies employed in the de-identification of corpora can be quite considerable, not only in terms of class granularity but also class overlap. This may be motivated by the specific characteristics of the de-identified domains or the end goal: taxonomies used for pseudonymization seem to feature more classes than those intended for anonymization, likely because the class of the PII is later used to generate a suitable pseudonym. This leads to the results not being fully comparable. The TAB CORPUS features fewer, semantically more general classes (grouping together many different concepts into one category); it also lexical or grammatical correction annotation[1]. This makes it unsuitable for addressing our research questions without a considerable amount of time going into manual reannotation.

However, it remains unclear how and to what extent the types of classes used in personal information detection affect the detection step itself. In Szawerna et al. (2024a) we consider a more detailed taxonomy more favorable, but we do not test that. We do, however, point out that what is personal is context-dependent and may vary between domains, so the choice of the labels can also depend on the domain. To the best of our knowledge, the only study that investigated whether a more diverse class division facilitates better PII detection is the one by Sierro et al. (2024). In this case, the authors adapted the TAB CORPUS by automatically translating it into Spanish and projecting the PII categories back into the text. They later re-annotated the corpus with refined, less ambiguous classes, leading to an increase in the number of classes. Notably, they also discard the `MISC` class, which is used to annotate very semantically diverse elements. They note an increase in performance on the dataset annotated using the refined tagset, which could be due to the new tagset being easier for their models to train on, but also due to manual re-annotation being more reliable than projection, and some information not being as revealing after translation.

3 Materials and Methods

3.1 Data

The data used in our experiments comes from the SWELL-PILOT (480 texts) and SWELL-GOLD corpora (502 texts) (Volodina et al., 2016, 2019; Språkbanken Text, 2024b,a), consisting of essays written by adult learners of Swedish as a second language (L2) at varying proficiency levels, with varied essay genres and topics. We chose to work with this data mainly because it is already PII-annotated with a hierarchical PII tagset and because its subset, SWELL-GOLD, features correction annotation which denotes e.g. grammatical variation in the text. The correction annotation was only used in evaluation, and our models were never overtly given that information.

While the released versions of the SWELL corpora[2] are pseudonymized, we utilize the texts in their original form with the unaltered PIIs in place. We preserve the aforementioned annotation of what spans contain personal information and of what kind. This annotation is done following the SWELL taxonomy (Megyesi et al., 2018), which consists of 38 types of PIIs (it also includes functional or morphosyntactic tags which we disregard for the sake of this experiment). Every PII token gets assigned an appropriate class and a number used for coreference resolution, which also helps

[1] This kind of annotation indicates that a token is in some way at odds with the standard for a given language, e.g. it is misspelled, the wrong word is used, the wrong grammatical form is used, or it is a part of a grammatical construction that is unacceptable from the standard point of view.

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

Class	Bs	Is	Total
firstname_male	234	0	234
firstname_female	289	0	289
firstname_unknown	49	0	49
initials	0	0	0
middlename	1	0	1
surname	49	2	51
school	44	25	69
work	2	0	2
other_institution	65	24	89
area	0	0	0
city	564	23	587
geo	17	0	17
country	400	1	401
place	93	19	112
region	37	2	39
street_nr	21	0	21
zip_code	7	2	9
transport_name	5	1	6
transport_nr	14	0	14
age_digits	82	0	82
age_string	12	0	12
date_digits	30	14	44
day	27	0	27
month_digit	9	0	9
month_word	46	0	46
year	53	0	53
phone_nr	7	0	7
email	10	0	10
url	0	0	0
personid_nr	0	0	0
account_nr	0	0	0
license_nr	0	0	0
other_nr_seq	169	1	170
extra	37	3	40
prof	12	2	14
edu	6	1	7
fam	464	3	467
sensitive	256	114	370

Table 2: Class counts for the detailed PII classes.

Class	Bs	Is	Total
personal_name	622	2	624
institution	111	49	160
geographic	1139	47	1186
transportation	19	1	20
age	94	0	94
date	165	14	179
other	961	124	1085

Table 3: Class counts for the general PII classes.

to define the edges of a PII span. These PII categories can be grouped into 7 general classes (as shown in Table 1). Therefore, the data can have the original SWELL classes (Specific), the overarching SWELL categories (General), or an even more general binary distinction whether the element is personal or not can be made (Basic; this corresponds more to a task of PII detection). It is worth noting that not all of the detailed SWELL classes are present in the data, and some were just theorized by the tagset creators to be possible. Many of the classes are also unlikely to span more than one token. The annotation can be modified to follow the inside-outside-beginning (IOB) schema or not include the distinction between beginning and inside (though the non-PII tokens are still marked as O in that case). This yields six different sets of classes that can be tested (henceforth Specific IOB, Specific, corresponding to Level 3 in Figure 1; General IOB, General, corresponding to Level 2; Basic IOB, Basic, corresponding to Level 1; see also Appendix A for a practical example).

When constructing our samples, we want to include as much context as possible, as we believe that the personal nature of a text span is context-dependent. Many of the essays exceed the maximum input size allowed by the BERT model that we are using.[3] We therefore split such essays into several chunks. Such a chunk has a maximum size of 512 BERT sub-word tokens. We ensure that our data consists of equally many samples containing at least one token belonging to a PII category as samples without any and that chunks of the same essay always appear in the same data split. This yields a collection of samples with 217,430 non-PII tokens and 3,348 PII tokens (3,111 B-tokens and 237 I-tokens). The exact counts for the Specific and General class sets can be found in Table 2 and Table 3, respectively. It is worth noting that some classes in the detailed set are not present in the data at all, and are only theoretically permitted by the taxonomy. Having considered discarding some of the data to balance the classes, we have decided against that, since our dataset is small as is, and we are curious to see how the prevalence of certain PII classes influences their labeling.

[3]Unfortunately, Longformer or a similar model is not available for Swedish.

Annotation type	Precision	Recall	F1	F2
Specific IOB	0.794 ± 0.028	0.709 ± 0.059	0.748 ± 0.042	0.724 ± 0.052
Specific	**0.867 ± 0.020**	0.733 ± 0.053	0.793 ± 0.036	0.756 ± 0.047
General IOB	0.788 ± 0.049	0.770 ± 0.061	0.770 ± 0.043	0.770 ± 0.053
General	0.858 ± 0.026	***0.803 ± 0.059***	***0.828 ± 0.037***	***0.813 ± 0.050***
Basic IOB	0.842 ± 0.021	0.796 ± 0.050	0.808 ± 0.037	0.800 ± 0.045
Basic	0.857 ± 0.019	**0.817 ± 0.045**	**0.836 ± 0.028**	**0.824 ± 0.038**

Table 4: Mean results ± standard deviation over the runs evaluated as detection (whether the token was detected as any PII class). Bold indicates the overall best scores. Italicized elements in bold are the best scores if the basic type of annotation were disregarded.

Annotation type	Precision	Recall	F1	F2
Specific IOB	0.497 ± 0.090	0.539 ± 0.083	0.498 ± 0.086	0.519 ± 0.085
Specific	0.591 ± 0.051	0.569 ± 0.062	0.550 ± 0.065	0.558 ± 0.063
General IOB	0.719 ± 0.041	0.727 ± 0.057	0.714 ± 0.049	0.720 ± 0.054
General	***0.806 ± 0.039***	***0.761 ± 0.062***	***0.770 ± 0.053***	***0.763 ± 0.059***
Basic IOB	0.842 ± 0.021	0.796 ± 0.050	0.808 ± 0.037	0.800 ± 0.045
Basic	**0.857 ± 0.019**	**0.817 ± 0.045**	**0.836 ± 0.028**	**0.824 ± 0.038**

Table 5: Mean results ± standard deviation over the runs evaluated as labeling (whether the token was assigned the right class). Bold indicates the overall best scores. Italicized elements in bold are the best scores if the basic type of annotation were disregarded.

3.2 Model and Code

We take the model from Szawerna et al. (2024b) that reports the best results, the Swedish BERT developed by the National Library of Sweden[4](Malmsten et al., 2020), which is based on the BERT architecture (Devlin et al., 2019), with a regular cross-entropy loss. This is confirmed by our own preliminary testing. Due to the model's relatively small size and short fine-tuning time, it is possible to conduct cross-validation.

In order to fine-tune KB-BERT we utilize the code for token classification[5] included in the Transformers library together with the model hosted on HuggingFace (Wolf et al., 2020). This code makes use of HuggingFace's Trainer class to fine-tune a BERT model for classification by discarding its head and replacing it with a new classification head, which is what is trained for the classification task at hand, while other pre-trained knowledge does not get altered. The only notable change that we make to the default settings of this classification set-up is decreasing the batch size to 8. For each of our 6 sets of data (which differ by annotation type) we conduct a 10-fold cross-validation.

For the rest of the preprocessing and evaluation we expand the code provided by Szawerna et al. (2024b) for working with SWELL data. [6]

3.3 Evaluation

For each of the runs, we obtain predictions on the held-out fold. We report the mean and the standard deviation across the 10 separate runs for each type of data. Due to the overwhelming prevalence of the non-PII tokens and following the example of e.g. Grancharova and Dalianis (2021), we report the means and standard deviations of the weighted averages of precision, recall, F1, and F2[7] across all of the PII classes (excluding the scores for non-PII tokens). Consequently, precision reflects the models' ability to avoid falsely flagging a word token as some PII class, whereas recall illustrates how well PII tokens can be detected instead of slipping through the cracks. The rationale behind reporting an F2 score is that it gives more weight to recall, and Berg and Dalianis (2020) consider recall to be a more important measure (as it reflects how many PII tokens were actually detected, which is a pri-

[4] KB/bert-base-swedish-cased, henceforth KB-BERT.

[5] https://github.com/huggingface/transformers/tree/main/examples/legacy/token-classification

[6] https://github.com/mormor-karl/the-devils-in-the-details

[7] $F_2 = (1 + 2^2) * \frac{precision * recall}{(2^2 * precision) + recall}$

Annotation type	Correction annotated	Misclassified	Correction-annotated and misclassified	% of misclassified tokens that are correction-annotated
Specific IOB	14014	405	47	11.61%
Specific	14014	407	52	12.78%
General IOB	14014	334	64	19.16%
General	14014	294	64	21.77%
Basic IOB	14014	277	67	24.19%
Basic	14014	255	65	25.49%

Table 6: Counts of the correction-annotated tokens, tokens misclassified during testing (in the labeling task), and the overlap of the two groups per type of PII annotation. Note that the number of correction-annotated tokens does not change across the PII annotation types and that these results concern only the data from SWELL-GOLD, as SWELL-PILOT does not include correction annotations.

ority).

However, we want to highlight that a high precision score is important as well, as avoiding flagging innocuous tokens as PII is essential for preserving as much of the original text as possible, which affects its later usability in linguistic research or NLP applications. Additionally, we evaluate the results both in terms of labeling – whether a token was assigned the correct class – and detection – whether the token was correctly identified as non-PII or any of the PII classes. In the case of the basic-type annotation, these two evaluations are equivalent.

We conduct further analysis, the purpose of which is to study two aspects of label selection: (i) whether grammatical and lexical divergence from the standard has an effect on the labeling of personal information, (ii) how the number of labels used and the depth of their semantics affects the labeling. We approximate the former by analyzing the raw counts of correction-annotated tokens that were misclassified and what percentage of all misclassified tokens they constitute within each annotation type.

4 Results

The mean detection results and the standard deviation over the runs are presented in Table 4. The same is shown for labeling in Table 5.

Both tables show that the IOB-type annotation appears to be more difficult to predict. This is likely due to relatively few PIIs spanning more than one token, leading to the classifiers having more issues determining those boundaries; in our case both the IOB component and the class label have to match for a token to be counted as correctly classified. Yet another aspect worth mentioning is that in many cases (`firstname_male`, `month_word`, etc.) the original SWELL annotation is intended to describe only one token, whereas other classes (e.g. `school`) are likely to consist of more than one token (see Table 2 and Table 3). This means that the effect of the IOB annotation is negligible for most classes.

When it comes to PII detection (Table 4), two different kinds of annotation excel in different metrics. Using the Specific annotation leads to the best precision. However, in terms of recall and the F-scores, the Basic annotation performs better.

In the labeling task, the basic type of annotation excels in all evaluation metrics. In the case of Basic annotation, detection and labeling are the same. If we consider the types of annotation where these two tasks are different, then the runs with the best recall, F1, and F2 for detection are the ones fine-tuned on the general type of annotation. In the case of labeling, these runs would be the best on all of the evaluation metrics. This partly contradicts the findings of Sierro et al. (2024), who report that more detailed classes facilitate better PII detection and labeling. However, the detailed classes in their experiment were refined based on what they found to be ambiguous in the original tagset. In our case, we utilized an existing hierarchical tagset. The difference in granularity between General and Specific classes is also much larger, as Sierro et al. (2024) split the original classes into at most 2 new classes, while in our data one General class can correspond to as many as 12 Specific classes.

We can only examine the interplay between the identification of PIIs and the tokens that were labeled for grammatical or lexical errors in different tagsets (Table 6) on the basis of SWELL-GOLD, as SWELL-PILOT does not include any correc-

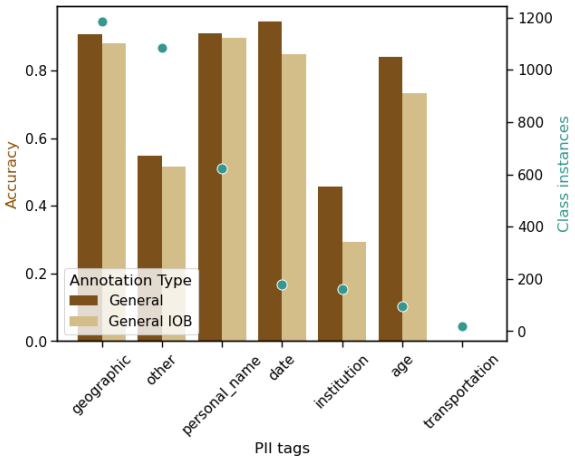

Figure 2: Per class prediction accuracy for the General and General IOB annotations (I and B is merged). The points illustrate the classes' raw frequencies.

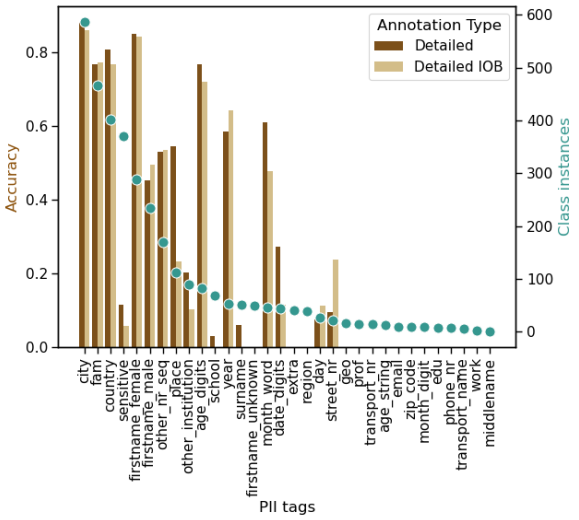

Figure 3: Per class prediction accuracy for the Specific and Specific IOB annotations (I and B is merged). The points illustrate the classes' raw frequencies.

tion annotation. The correction annotations were not visible to the classifier during training, and instead we use them to identify the tokens that were judged to belong to a grammatically or lexically non-standard span. What appears to be influenced by the annotation type is the number of total misclassifications, the percentage of those that consists of correction-annotated tokens, and the raw counts of correction-annotated misclassified tokens. It is clear that the more diverse types of annotation lead to more misclassifications in general; however, there is a reverse trend when it comes to what percent of the misclassified tokens is also correction-annotated. It follows that more diverse annotation is less affected by errors than more general annotation. This could mean that the poorer performance noted for more detailed annotation is caused by the multi-class classification during labeling being inherently more difficult given the number of classes, but that the models learn to connect the more specific tokens better with the word embeddings and their contexts that represent the semantics of the text to determine that the span is a part of some PII. This is also partly reflected in the major improvement of the scores when the predictions are reinterpreted from labeling into detection (as the scores for Specific and Specific IOB then jump by 15 to 30 percentage points).

Figure 2 and Figure 3 show the per-class accuracy (disregarding the I and B distinction). Points indicating the number of instances of the respective class in the data are overlaid atop the accuracy bar charts.

Figure 2 shows these statistics for the General and General IOB tagsets. For some of the classes, prevalence in the data correlates with accuracy – nearly 1200 tokens belong to the `geographic` class, which has high accuracy, while `institution`, with around 200 tokens, shows worse results and the extremely infrequent `transportation` class practically never gets correctly predicted. However, there are classes that diverge from this trend: despite having almost as many instances as `geographic`, `other` has noticeably lower accuracy, implying that they are difficult to predict. Less frequent classes like `personal_name`, `date`, and `age` achieve high accuracy scores despite not being as numerous as some other classes, indicating that they are easier to predict.

A similar phenomenon can be observed in Figure 3, which represents the Specific and Specific IOB tagsets. Classes like `city`, `fam`, and `country`, have high frequency and high accuracy. Many of the infrequent classes practically never get correctly predicted, and classes with intermediate frequency, like `firstname_male` or `other_nr_seq` have mediocre accuracy. Once again there are also frequent classes with low accuracy (`sensitive`) and less frequent

classes with high accuracy (e.g. `age_digits`, `month_words`) — which, once again, suggests that some classes can be easy or difficult to predict regardless of their frequency.

These results suggest that while having many examples helps the models to learn to predict a given class, some classes are much easier or much more difficult to predict than others. The performance of some classes is high because of their high frequency in the dataset, whereas some other classes are easy to predict despite not being all that frequent. It therefore appears that it is not only class frequency, but also class semantics that influence the accuracy of predictions, with frequent classes and classes with little internal variation in meaning performing better.

This might also explain why Sierro et al. (2024) observed an improvement with a larger number of classes, as their increase in the number of classes happened once they split (and subsequently narrowed down the semantics of) vague classes and disregarded the `MISC` class, their equivalent of our `sensitive` or `other`. This confirms that identifying semantically distinct classes for annotation is crucial for the success of the annotation scheme and its application in classification tasks. Such labeling requires a good understanding and knowledge of the domain.

While the results show what kind of annotation facilitates the best *detection* or *labeling*, the results of the experiments do not allow us to identify the overall best type of PII annotation, as this depends on the subsequent steps. For example, if the final corpus should contain more specific labels for anonymized spans, then it may be worth to split the process into detection followed by labeling, as detection outperforms labeling at this level of tagset detail; there are some results from other tasks which may suggest that such a separation could be beneficial, e.g. Park and Fung (2017). Another related observation is that PII entities tend not to appear directly adjacent to other PII elements belonging to the same class, which suggests that such boundaries (i.e. IOB-type annotation) need not be included, but it may vary for different labels and domains.

5 Conclusions

We have compared the performance of KB-BERT-based classifiers on detecting and classifying Personally Identifiable Information distinguished by a different number of classes and the semantic depth or specificity of these classes. We have found that for PII detection, Basic, non-IOB annotation yields the best results. When it comes to labeling, more specific classes do not ensure better results, possibly due to some of those classes being underrepresented, since frequency does appear to play some role in how well various classes are detected. An IOB-style annotation also results in a decrease in performance versus not differentiating between beginnings and insides of spans.

We have also found that models fine-tuned on more basic annotation tend to misclassify words that are misspelled, misplaced, or syntactically incorrect more often than models fine-tuned with more specific classes. We have also observed that it is not only class imbalance and a low frequency of a number of the classes, but also the classes' semantics that influence the accuracy of the predictions. Semantically less coherent or less constrained classes make it much more difficult for the models to make correct predictions, pointing to the need for well-defined classes. This emphasizes the role of understanding the domain for which the annotation scheme is designed and raises an important issue concerning the cross-domain transfer of annotation schemes as different classes will have different frequencies and semantics across these classes.

While the choice of PII taxonomy is likely to depend on the needs of the specific case, the results suggest that using over-detailed classes for automatic PII detection and labeling may not lead to optimal performance, at least not without a large dataset for the model to learn from. The same applies to the differentiating between the beginning and the inside of a PII span in IOB-type annotation, which does not lead to better performance, and therefore should only be included if required in the specific case.

In these experiments we have shown what kinds of annotation facilitate PII detection and labeling, the final choice also depends on the subsequent task, such as generating pseudonyms or removing PII spans. As long as the classes are required by the subsequent steps in a pipeline (e.g. pseudonym generation) or desired in the final version of the text (e.g. as placeholders in the anonymized text), there is a need for a more detailed annotation than the basic one utilized in our experiment. This also signals a need for investigating whether the label-

ing step can be separated from the detection step, and how the performance of such a setup compares to classifying the PIIs in a single step.

The overt class imbalance (including the lack of any PIIs of certain kinds of Specific labeling in the data) highlights the need for well-curated training datasets that feature a sufficient number of PIIs of each kind, either by collecting more data or adjusting the annotation; alternatively, one could also opt to combine machine learning and rule-based detection methods (many of the absent Specific classes, such as `account_nr`, could be more easily identified using e.g. regular expressions).

6 Future Work

To strengthen our results, these methods should be applied to larger amounts of training data, potentially resolving issues pertaining to some of the classes being very difficult for our models to learn to predict due to their low frequency. Since we also observe that the semantic vagueness of certain classes is problematic for the models, it would be interesting to split those classes into more coherent subclasses and examine what effect that has – however, this requires manual re-annotation of those tokens. Equally, we would like to see how these results compare for different domains where labels have different distributions in the text or are entirely different.

The question related to the variability of data (here in terms of non-standard spelling in the form of grammatical errors but also other variability such as unconstrained communication) and its interaction with the selected annotation scheme is also open for further exploration. An alternative route would be trying to utilize synthetic data, and, especially, comparing the performance of models trained on larger amounts of synthetic data with models that were only trained on a smaller corpus of authentic data. An intermediate step would be augmenting the training data using e.g. manually or automatically pseudonymized versions of the same texts.

It can also be worth exploring whether the same trends occur when using other BERT-type models for this task — although KB-BERT has been shown to perform the best on PII detection in Swedish texts, perhaps other models do not show the same trends as it does in these experiments.

We also aim to construct PII detection and PII labeling models which we plan to release without any privacy risks. Comparing an approach where we separate detection and labeling versus where they are combined in a single step is also an interesting path. Since more granular tagsets seem to be used for pseudonym generation in many cases, we consider it worth exploring alternative methods for pseudonym generation that are not as dependent on the PII taxonomy used, e.g. using language models.

Acknowledgments

This work was possible thanks to the funding of several grants from the Swedish Research Council.

All of the authors are supported by the research environment project *Grandma Karl is 27 years old: Automatic pseudonymization of research data* with the funding number 2022-02311 for the years 2023-2029.

The first, third, and fourth authors are also receiving support from the Swedish national research infrastructure *Språkbanken*, which is jointly funded by its 10 partner institutions and the Swedish Research Council (2018–2024; dnr 2017-00626).

The second author is also supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the *Centre for Linguistic Theory and Studies in Probability (CLASP)* at the University of Gothenburg.

This work has also been aided by the Swedish national research infrastructure *Huminfra*, funded for the years 2022-2024, contract 2021-00176, and the participating partner institutions.

Limitations

One major limitation in our experiments is the relatively small amount of training data. However, the particular hierarchical PII taxonomy that we analyze is only used in the SWELL corpora, and SWELL-GOLD's correction annotation sets it apart from other corpora with hierarchical annotation, such as CODE ALLTAG (Eder et al., 2020). Unfortunately, SWELL-PILOT is not correction-annotated, meaning that we can only conduct certain result analyses on a subset of our data.

Despite the small amount of data, a qualitative analysis of the errors made by the models was deemed to be beyond the scope, as it would require a manual inspection of almost 1000 texts in six different annotation versions.

Since it takes a considerable amount of time to train a BERT-based classifier, we trained on 6 different kinds of annotation, we limited ourselves to 10 runs per annotation type, which does not satisfy the requirements of applying statistical tests on the overall performance results.

Ethical Considerations

Since the data that we use to fine-tune our models includes Personally Identifiable Information, it cannot be openly shared. We choose not to share our models to avoid any risks of leakage of personal information. However, we provide the code (see subsection 3.2) from which the results can be generated provided one has access to the data in the appropriate SWELL format.

References

Alyaa Alfalahi, Sara Brissman, and Hercules Dalianis. 2012. Pseudonymisation of Personal Names and other PHIs in an Annotated Clinical Swedish Corpus. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC 2012*.

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.

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.

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.

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.

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

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.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Official Journal of the European Union. 2016. Consolidated text: 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).

Anthi Papadopoulou, Pierre Lison, Mark Anderson, Lilja Øvrelid, and Ildikó Pilán. 2023. Neural text sanitization with privacy risk indicators: An empirical analysis.

Anthi Papadopoulou, Yunhao Yu, Pierre Lison, and Lilja Øvrelid. 2022. Neural text sanitization with explicit measures of privacy risk. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 217–229, Online only. Association for Computational Linguistics.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45,

Vancouver, BC, Canada. Association for Computational Linguistics.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Maria Sierro, Begoña Altuna, and Itziar Gonzalez-Dios. 2024. Automatic detection and labelling of personal data in case reports from the ECHR in Spanish: Evaluation of 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.

Språkbanken Text. 2024a. SweLL-gold.

Språkbanken Text. 2024b. SweLL-pilot.

Maria Irena Szawerna, Simon Dobnik, Therese Lindström Tiedemann, Ricardo Muñoz Sánchez, Xuan-Son Vu, and Elena Volodina. 2024a. Pseudonymization categories across domain boundaries. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13303–13314, Torino, Italia. ELRA and ICCL.

Maria Irena Szawerna, Simon Dobnik, Ricardo Muñoz Sánchez, Therese Lindström Tiedemann, and Elena Volodina. 2024b. 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.

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and Mats Wirén. 2019. The SweLL Language Learner Corpus: From Design to Annotation. *Northern European Journal of Language Technology*, 6:67–104.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016. SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), May 23-28, 2016, Portorož, Slovenia*, Paris. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing.

A Appendix

Example (1), shows what all the annotation schemes used in this paper look like on sample text. The annotation schemes a-f correspond to the Specific IOB, Specific, General IOB, General, Basic IOB, and Basic annotations, respectively.

(1) a. My name is **Maria** . I
O O O B-firstname_female O O
come from **Wroclaw** (that is in
O O B-city O O O O
Poland) . I work at the
B-country O O O O O O
University of Gothenburg .
B-work I-work I-work O

b. My name is **Maria** . I come
O O O firstname_female O O O
from **Wroclaw** (that is in **Poland**) .
O city O O O O country O O
I work at the **University of**
O O O O work work
Gothenburg .
work O

c. My name is **Maria** . I come
O O O B-personal_name O O O
from **Wroclaw** (that is in
O B-geographic O O O O
Poland) . I work at the
B-geographic O O O O O O
University of Gothenburg
B-institution I-institution I-institution
.
O

d. My name is **Maria** . I come
O O O personal_name O O O
from **Wroclaw** (that is in **Poland**
O geographic O O O O geographic
) . I work at the **University**
O O O O O O institution
of Gothenburg .
institution institution O

e. My name is **Maria** . I come from
O O O B O O O O
Wroclaw (that is in **Poland**) . I
B O O O O B O O O
work at the **University of Gothenburg**
O O O B I I
.
O

f. My name is **Maria** . I come from
O O O S O O O O
Wroclaw (that is in **Poland**) . I
S O O O O S O O O
work at the **University of Gothenburg**
O O O S S S
.
O