Detecting Personal Identifiable Information in Swedish Learner Essays

Maria Irena Szawerna[†], **Simon Dobnik**[‡], **Ricardo Muñoz Sánchez**[†],
Therese Lindström Tiedemann[§], **Elena Volodina**[†]
[†]Språkbanken Text, SFS, University of Gothenburg, Sweden
[‡]CLASP, FLoV, University of Gothenburg, Sweden
[§]Department of Finnish, Finnougrian and Scandinavian Studies, University of Helsinki, Finland
mormor.karl@svenska.gu.se
[†]{maria.szawerna,ricardo.munoz.sanchez,elena.volodina}@gu.se
[‡]simon.dobnik@gu.se
[§]therese.lindstromtiedemann@helsinki.fi

Abstract

Linguistic data can — and often does — contain PII (Personal Identifiable Information). Both from a legal and ethical standpoint, the sharing of such data is not permissible. According to the GDPR, pseudonymization, i.e. the replacement of sensitive information with surrogates, is an acceptable strategy for privacy preservation. While research has been conducted on the detection and replacement of sensitive data in Swedish medical data using Large Language Models (LLMs), it is unclear whether these models handle PII in less structured and more thematically varied texts equally well. In this paper, we present and discuss the performance of an LLM-based PII-detection system for Swedish learner essays.

1 Introduction

While there is a constant need for linguistic data — fuelled recently by the advent of Large Language Models (LLMs) which require copious amounts of training data — legal and ethical sharing and use thereof is problematic. The EU Commission (2016) severely limits the use and sharing of data containing Personal Identifiable Information (PII). However, the regulation also presents a possible solution: pseudonymizing the data, defined as: "[...] the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person" (Art. 4 of EU Commission, 2016). Within the field of Natural Language Processing (NLP), this definition becomes more narrow — and while various researchers formulate it slightly differently, we understand pseudonymization as "the process of replacing an individual's personal data with a pseudonym, which is not related to the original data," with the same end goal as outlined in the GDPR (Volodina et al., 2023).

Naturally, conducting such a de-identification procedure manually is extremely time-consuming and costly, especially when the data in question is copious and very sensitive (Berg and Dalianis, 2020). It would therefore be beneficial to be able to automatize the process in a reliable and robust way. While there is existing research on automated de-identification systems, many of them are restricted to specific domains (especially healthcare), and not as much work has been conducted on less structured types of input, which we expect to be more problematic due to more varied types of personal information as well as a higher likelihood of various kinds of errors or non-standard forms (e.g. in terms of spelling, syntax, or semantics). We choose to work with L2 (second language) learner essays, as this kind of texts not only fulfills the requirement of larger structural and thematic variety but, as Volodina et al. (2020) show, the essays are also likely to contain PIIs. Since L2 corpora are relevant for various research applications, developing models that can handle PII detection and replacement in this kind of texts would be useful.

What exactly constitutes sensitive information can differ across domains, documents, or even paragraphs, and is heavily context-dependent. We believe that algorithms could learn something akin to human intuition about what is personal and/or sensitive in the data. With this in mind, we experiment with an approach where none of the PII and sensitive categories are labeled for their classes (e.g. name, city, etc.), but are binary (personal/sensitive or not). This distinction is formalized as inside-outside-beginning (IOB) classes, where non-sensitive tokens are labeled O (outside), while sensitive tokens or token spans are labeled with B (beginning) and, in the case of multi-token sensitive elements, I (inside) for every token after

Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization
CALD-pseudo 2024, pages 54–63

the first one. We replace manually assigned categories in our dataset of learner essays in Swedish (SweLL-pilot, Volodina et al. (2016)) with B(I) and O, and fine-tune two Large Language Models (LLMs, `KB/bert-base-swedish-cased` and `bert-base-multilingual-cased`) to distinguish between the two types of tokens (Malmsten et al., 2020; Devlin et al., 2018). While we are aware that pseudonym generation is likely to rely on a predicted PII class, we decide to focus on the detection step, which can precede classification – presuming that such a step is necessary in a pseudonymization pipeline. Our hypothesis is that the model will learn to distinguish between sensitive and non-sensitive information in a given context, and potentially even capture more types of personal information than we at the moment envisage, helping us identify new classes that can be added to the taxonomy or refine the existing ones. Simultaneously, we hope to assess the usefulness of fine-tuned LLMs for PII detection, especially in more free-flowing and error-prone genres such as learner essays.

2 Prior Research

As previously mentioned, pseudonymization, as we understand it, entails the replacement of sensitive tokens or groups of tokens with new and somewhat unrelated — but still contextually appropriate — surrogates. The replacement of PII with a pseudonym presupposes a step at which the sensitive data is detected and possibly classified; recently Eder et al. (2022) conceptualized the pseudonymization pipeline as consisting of the two aforementioned steps.

While Lison et al. (2021) consider the pseudonym generation step to be more of an open question than the detection of PIIs themselves, many previously presented detection systems do not account for, for example, misspellings or otherwise non-normative writing, which is essential when working with data like learner essays (Eder et al., 2019). Although Accorsi et al. (2012) highlights the issues stemming from spelling variation, these issues seem to mostly pertain to specific genres, which so far have been underrepresented in PII detection research, as the bulk of the existing research is focused on medical data.

As shown by Yogarajan et al. (2020), many of the well-performing systems for PII detection in medical data rely on neural or hybrid approaches. Recently, Pilán et al. (2022) have released a text anonymization benchmark corpus consisting of texts from the legal domain, and presented the results obtained by several models. While their custom metrics rely at least partly on there being more than one possible way to annotate a text, they do provide overall recall and precision as well, with the best model — a LongFormer model with a large window size — reaching 91.9% recall and 83.6% precision; however, an F1 score is not reported. Grancharova and Dalianis (2021), in turn, fine-tuned a Swedish BERT model for Named Entity Recognition and Classification (NERC) in Swedish medical texts. The NER categories in the corpus utilized in their experiment are actually PHI categories, which could be considered a type of PII, rendering this task sufficiently similar to warrant a comparison.

They report precision and recall scores for various models, with the best of them (KB-BERT trained and tested on data from the same source) reaching a weighted precision score of 92.26% and a weighted recall score of 92.20% (with the weighted F1 of 92.23%). They also reach relatively good scores on M-BERT (multilingual BERT) with the same data setup - 88.99% recall and 90.51% precision (and F1 of 89.74%). While Berg and Dalianis (2020) argue that high recall is more desirable in PII detection systems than high precision, the latter is also important, as it means that the model is not over-detecting the sensitive data and flagging innocent passages. We believe that the alterations to the text should be kept to a necessary minimum as any changes made to the linguistic data may affect its future usability in various types of research (e.g. linguistics or machine learning). While our experiment is meant to test an approach similar to that of Grancharova and Dalianis (2021), it is worth keeping in mind that the data we use is less structured and may contain a bigger variety of personal information, as described in Volodina et al. (2020), which may lead to a worse performance by the system.

3 Materials and Methods

In this experiment, we utilize 445 learner essays from the SweLL-pilot corpus, representing a wide variety of learner levels, topics, and types of writing (e.g. descriptive or argumentative essays) (Volodina et al., 2016). Some of the essays contain PIIs, and some do not, predominantly due to the variation in types of writing and the prompts (e.g. a

descriptive essay with the topic“about me" is much more likely to contain PIIs than an argumentative essay with the topic “stress in the modern society"). We use the unpseudonymized[1] versions of the texts. The essays have also been tokenized and reannotated with tags for PII categories using the SVALA tool according to the SweLL pseudonymization guidelines developed for the SweLL-gold corpus and the corresponding tagset (Wirén et al., 2019; Megyesi et al., 2021; Volodina, 2024). This annotation includes not only typically NER-like categories such as place names or surnames, but also e.g. names of professions or references to one's faith, with only the tokens deemed sensitive in a given context being annotated as such. In our experiments, we ignore the categories of sensitive information and only differentiate between sensitive and non-sensitive information. We transform the existing category annotation into an inside-outside-beginning (IOB) annotation to represent the difference between PIIs and non-PII tokens. Due to BERT-imposed input sequence limitations, we subdivide the essays into sections that are at most 100 tokens long,[2] resulting in a total of 651 such sections, out of which 165 contain at least one token of sensitive information.

The data is then balanced so that the data splits (training, testing, development) contain equally many passages with PIIs as passages without PIIs, meaning that these splits are composed of 165 fragments with PIIs and 165 randomly chosen fragments without PIIs out of 486 such fragments. Importantly, this does not mean that half of the tokens include sensitive information, and the actual distribution of sensitive and non-sensitive tokens can be seen in Table 1. We also calculate weights that represent class importance for later use with a weighted loss function using Scikit-learn's `compute_class_weight` function (Pedregosa et al., 2011). The class distribution and the calculated weights for the data used in the experiment are presented in Table 1.

[1]This version is used only within the context of the project that this experiment is conducted in and is unavailable for anyone except the project team. The released version of SweLL-pilot is anonymized and the access form is linked in Appendix A.

[2]While BERT's maximum input sequence length is 512, this applies to the sequence length after tokenization using the BERT tokenizer, which often divides words into sub-word units; since the sectioning of the essays occurred at a much earlier step than BERT tokenization due to the framework used, an arbitrary length was chosen to mitigate the impact of the BERT tokenizer and maximum sequence length.

	Instances (%)	Count	Weight
B	2.64%	1142	12.64419148
I	0.20%	86	167.90310078
O	97.16%	42091	0.34305829

Table 1: The proportions of token instances of classes in the data used in the experiment and the corresponding calculated class weights.

The PII-detection system used in this paper is based on modified code for token classification included in the `transformers` library (see Appendix A) (Wolf et al., 2020). This code allows for the fine-tuning of a model of choice hosted by HuggingFace for a token classification task like NER (Named Entity Recognition) or part-of-speech (POS) tagging; in our case, we have chosen to work with the BERT model for Swedish (`KB/bert-base-swedish-cased`, KB-BERT)[3] developed by the National Library of Sweden (Kungliga Biblioteket, KB) as well as a multilingual BERT model (`bert-base-multilingual-cased`, M-BERT)[4] (Malmsten et al., 2020; Devlin et al., 2018). This was done to mirror the setup utilized by Grancharova and Dalianis (2021) for an easier comparison of results; simultaneously, our hope is that using a multilingual model may help mitigate the effect the foreign tokens found in learner essays may have on the performance of the system, since those tokens may then be parsed as something other than an unknown word. Additionally, having an insight into whether multilingual models can be used for this type of task could be useful when working with languages that are only featured in multilingual models.

We have fine-tuned the models on 80% of our data (after balancing the set) twice, once with a standard `CrossEntropyLoss` loss function, and once with a weighted version thereof, with the intent of accounting for the class imbalance in a task of this type[5]. We have also reduced the batch size to 8 since due to the length of the samples we did not have the computational resources to process that much data in one batch. Aside from that, we have proceeded with the default settings for the script (notably, 3 epochs and AdamW optimizer

[3]https://huggingface.co/KB/bert-base-swedish-cased

[4]https://huggingface.co/bert-base-multilingual-cased

[5]Similarly to regular NER tasks, sensitive and not sensitive tokens are not equally prominent in the data, with the majority of the tokens being not sensitive.

with a learning rate of 5e-05). The fine-tuning process also makes use of another 10% of our data for evaluation between the epochs (development set).

The fine-tuned model has been tested on the held-out test set (another 10% of the data). The Transformers evaluation code calculates average evaluation metrics but here we also additionally calculate per-class metrics. Additionally, tokens misclassified by the models relative to the manually annotated gold standard have been extracted with their contexts and analyzed manually to see if any patterns of what the model struggled with could be identified.

4 Results and Discussion

4.1 Evaluation Metrics

The standard KB-BERT model (using an unweighted cross-entropy loss function) performs better in terms of accuracy, with the standard M-BERT and weighted KB-BERT only slightly behind. Surprisingly, the M-BERT model with a weighted loss function performs drastically worse, as shown in Table 2. However, it is important to remember that accuracy is not a weighted metric and that the O class outnumbers the other two.

Accuracy			
Standard model		Weighted model	
KB	M	KB	M
99.11%	97.78%	97.73%	29.16%

Table 2: The models' accuracy.

Due to the aforementioned class imbalance, we find it important to inspect measures like per-class recall and precision instead of just accuracy in order to gain a better understanding of the performance of the models. We additionally follow the example of Grancharova and Dalianis (2021) and provide combined scores for the "sensitive" classes (B and I).

	Recall			
	Standard model		Weighted model	
	KB	Multi	KB	Multi
B	82.57%	38.53%	92.66%	74.31%
I	14.29%	0.00%	57.14%	0.00%
O	99.67%	99.46%	97.93%	28.05%
B+I[6]	77.79%	35.83%	90.17%	69.11%

Table 3: The models' per-class recall.

	Precision			
	Standard model		Weighted model	
	KB	Multi	KB	Multi
B	86.54%	64.62%	58.38%	2.58%
I	100.00%	0.00%	18.18%	0.00%
O	99.41%	98.28%	99.78%	97.46%
B+I[8]	87.48%	60.09%	55.57%	2.40%

Table 4: The models' per-class precision.

The scores presented in Table 3 show that using a weighted loss function in KB-BERT models has improved the detection of the two classes that are used to denote the sensitive information (B and I), at the cost of a small drop in the recall for O.

Simultaneously, while it helps with the detection of the B class in M-BERT models, it has no effect on the detection of I and causes a drastic drop in the detection of O. It is rather clear that the models are struggling with the detection of I-tags, likely due to them being extremely infrequent in the data, with most of the sensitive data being restricted to single tokens. Comparing this with the results obtained by Grancharova and Dalianis (2021) for their sensitive data detection models for the medical domain, we achieve 90.17% recall on the sensitive data in our best model compared to their 92.20%, leading us to the conclusion that in terms of recall, our weighted KB-BERT model is performing rather well, especially taking into account the fact that the types of PII present in learner essays are more diverse and potentially harder to detect than those found in medical data (a more narrow domain). However, the same cannot be said about any of the M-BERT models which fail much more noticeably when trained with the current hyperparameters: our 69.11% for the weighted M-BERT model is much lower than 88.99% reported in the aforementioned research. This is further illustrated in Figure 1, Figure 2, Figure 3, and Figure 4, which depict normalized confusion matrices for the models' predictions, where the numbers on the main diagonal correspond to per-class recall.[7]

[6]Weighted average of scores for the two sensitive classes.

[7]Please note that any value differences stem from different rounding in the table than in the confusion matrices.

[8]Weighted average of scores for the two sensitive classes.

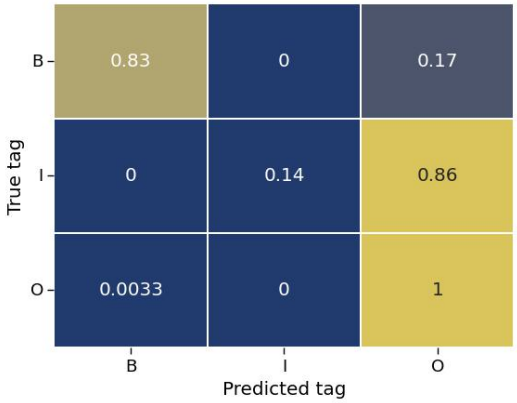

Figure 1: Normalized confusion matrix for PII detection with KB-BERT with the standard `CrossEntropyLoss`.

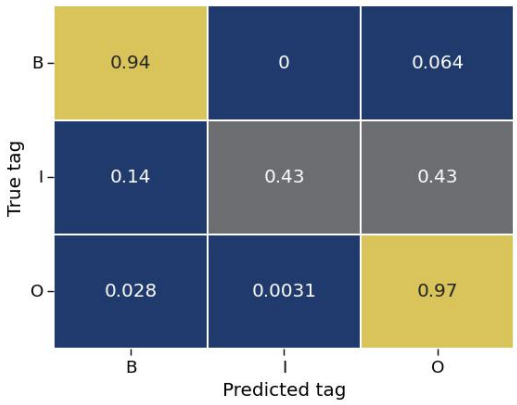

Figure 2: Normalized confusion matrix for PII detection with KB-BERT with the weighted `CrossEntropyLoss`.

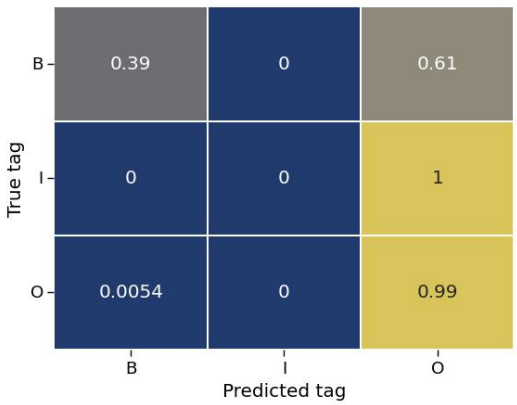

Figure 3: Normalized confusion matrix for PII detection with M-BERT with the standard `CrossEntropyLoss`.

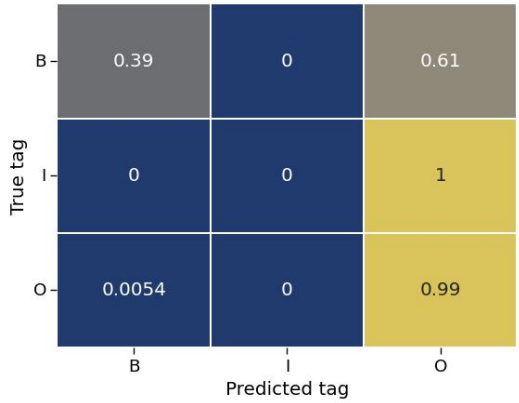

Figure 4: Normalized confusion matrix for PII detection with M-BERT with the weighted `CrossEntropyLoss`.

Contrary to the results for recall, precision for the sensitive classes is much better for the models without the weighted loss function, as shown in Table 4. Once again, the M-BERT models are overall performing worse, with the weighted version thereof achieving the worst result. This indicates that the weighted models are noticeably over-detecting tokens as sensitive — so although they now correctly identify more of the originally sensitive passages, they are also marking completely non-sensitive tokens as sensitive. While it is more important to correctly detect as many PIIs as possible, we are of the opinion that for the data to be useful for downstream tasks, such as semantic meaning extraction or information retrieval, it should be altered only as much as necessary, meaning that high precision would also be desirable.

One way to reconcile the need for considering both recall and precision in model evaluation is to look at the F1 score. One drawback of this score is that it assigns equal importance to its constituent parts, which is less ideal in the current scenario, where recall is considered to be more important. However, it is a widely used metric and it still allows us to compare the models to each other and to results from other research. Table 5 contains the per-class f1 scores alongside a weighted average of that score for the two sensitive classes.

In terms of the F1 score, the KB-BERT model with the standard cross-entropy loss function performs best on two out of three classes and as far as the combined B and I score is concerned. The KB-BERT with a weighted loss function slightly outperforms it on the I class. Both of the M-BERT models display significantly worse performance.

The standard KB-BERT model achieves the best

	F1			
	Standard model		Weighted model	
	KB	Multi	KB	Multi
B	84.51%	48.28%	71.63%	4.99%
I	25.00%	0.00%	27.59%	0.00%
O	99.54%	98.86%	98.85%	43.56%
B+I[9]	80.34%	44.89%	68.55%	4.64%

Table 5: The models' per-class F1 score.

result here - 87.48% weighted precision for the sensitive classes, which is still somewhat below 92.26% reported by Grancharova and Dalianis (2021); it is also not fair to compare only the highest results, as they are not achieved by the same model; the one with the best recall score only achieved 55.57% precision. When it comes to the F1 score, our best model (KB-BERT with a standard loss function) with a score of 80.34% on the sensitive classes is about 12 percentage points behind the best model for medical data, which is reported to have achieved 92.23% F1. This disparity stems from our model's decidedly lower precision.

Judging by all of the discussed metrics, KB-BERT models perform better than the multilingual BERT models. With the current hyper-parameters, the standard models suffer from relatively low recall, especially for the rarest class; weighted models, in turn, are over-detecting sensitive data, leading to lower precision. Nevertheless, the results seem to indicate that all the models except M-BERT with a weighted loss function are capable of distinguishing between sensitive and non-sensitive passages with a reasonable level of correctness. Importantly, the underdetection of the I class by all of the models suggests that they struggle with detecting multi-token spans of sensitive data.

It is also important to mention that we consider the task of learning to simply distinguish between sensitive and non-sensitive tokens or sequences of tokens to be more difficult than distinguishing specific classes of PIIs or PHIs, which is also reflected in the notably low precision of most of the models that we have trained. However, the results promisingly suggest that LLMs are indeed capable of learning, to some extent at least, what makes data sensitive in a given context.

[9]Weighted average of scores for the two sensitive classes.

4.2 Qualitative Prediction Analysis

A qualitative analysis of the predictions made by the models allows us to investigate what types of data marked as sensitive during manual annotation are particularly problematic for the models — and what kinds of generalizations lead to over-detection of PIIs they make. Importantly, due to the sensitive nature of the data used in this experiment sharing specific examples raises ethical concerns. We have decided to address this issue twofold: we manually pseudonymize the sensitive tokens in the examples and we provide the examples only in English (while simultaneously trying to mirror any kinds of learner errors).

The weighted M-BERT has failed to learn to differentiate between sensitive and non-sensitive data, as it does not mark some words with regular spelling (common Swedish given names, names of languages), and instead classifies words such as pronouns, determiners, some verbs as sensitive, in contexts where they with a great degree of certainty are not sensitive, as Examples 1 and 2 in Table 6 show. Simultaneously, some clearly sensitive tokens do not get recognized as such (Example 4). There are also instances of misspelled tokens being assigned the wrong category, but sometimes it is unclear whether the cause for the misclassification was the spelling or the model's disagreement as to what private data is, as in Example 3, where one could argue that *reltivs* "relatives" is a word denoting family members which could potentially be sensitive. This could be due to a language-specific model like KB-BERT being better at capturing specific semantic knowledge and being better able to generalize over e.g. street or place names; alternatively, it could be that while we have expected a multilingual model to improve the results since it would have representations for foreign language tokens, it actually struggled more with misspellings. While we did not explicitly notice that in our results, it is also possible that a multilingual model may have issues with tokens that have two separate meanings in two different languages.

The M-BERT model with the standard loss function, which has achieved low recall but somewhat higher precision appears to make more interpretable decisions: there are instances where this classification could be up for debate, and perhaps the token should have been marked as such by the annotator. This can be seen in Example 5, where *Stockholm* is not where the author lives, but

№	Token	Token in context	Prediction	Ground truth
M-BERT WITH A WEIGHTED LOSS FUNCTION				
1	was	Historically, stress **was** a [...]	B	O
2	me	**me** and johnny at school sit	B	O
3	reltivs	Other **reltivs** have come	B	O
4	Alice	[...] they are called Sally, **Alice** and Sam.	O	B
M-BERT WITHOUT A WEIGHTED LOSS FUNCTION				
5	Stockholm	We came to **Stockholm** city from Cairo directly	B	O
6	Germany	[...] one stress muc more in **Germany**.	B	O
7	Malmö	$$$$$[10] $$$ $$ $$$$s in **Malmö**. Later w$	O	B
8	Nobel street[11]	I live on **Nobel street**.	O	B
KB-BERT WITH A WEIGHTED LOSS FUNCTION				
9	sweden	tim lives in the family in **sweden**	B	O
10	novmber	wynter is four months from **novmber** to February	B	O
11	small	because I have a **small** family here.	O	B
12	family	because I have a small **family** here.	B	I
KB-BERT WITHOUT A WEIGHTED LOSS FUNCTION				
13	dad	and my **dad** was dizzy always	B	O
14	Cairo	**Cairo** has a verybig airport	B	O
15	Pierogi	they eat **Pierogi** which are traditional fud	O	B
16	%olis%	I am **$olis$**. We $$$$$$ $$$ $$	O	B
17	don't work[12]	I **don't work**.	O	B, I

Table 6: Examples of errors made by the M-BERT model with a weighted loss function.

constitutes an intermediate point in their travel, or in Example 6, where one can guess that someone writing about the reality of living in a given country in an argumentative essay has likely been born and raised there, or at least lived there for a longer period of time. We believe it is likely that in this case, the model has learned to classify all cities and countries that it has recognized as sensitive; this effect could at least partly be attributed to a possible imbalance between instances where such entities are not sensitive versus when they are sensitive.

When it comes to KB-BERT, the model with the weighted loss function provides even more examples of the model overgeneralizing certain entity types to always be sensitive — in the SweLL annotation, *Sweden* was not considered to be sensitive (as it was certain that all of the essays came from people living in Sweden), and yet in Example 9 the model predicts it to be sensitive. Similarly, *november* in Example 10 does not refer to a specific event in the author's life, but rather to a description of the climate, rendering it rather non-sensitive. Another interesting example here comes from two subsequent words in a sentence – since we differentiate between the start and the continuation of a sensitive passage, misclassifying the first token as non-sensitive, but classifying the second one as sensitive still leads to two errors, as in the case of the second error the class should be I, not B. Nevertheless, this suggests that a small fraction of the errors made by the model could be attributed to such cases, meaning that the model's performance is slightly better than the evaluation metrics may show.

The highest-scoring model in terms of evaluation metrics, KB-BERT without a weighted loss function, still has examples of the issue of overgeneralization (Example 14). However, it also illustrates that in some cases the annotators may have missed data that should be considered sensitive — like in Example 13, where the word for a specific family relation was not annotated as sensitive when it should have been according to the guidelines. Understandably, the model struggles with half-unintelligible tokens, such as in Example 16,

[10] $ is used to designate unintelligible handwriting.

[11] Names of streets are often just one token in Swedish.

[12] In Swedish the negation comes after the verb in the main clause, so in the original the I tag would refer to the negation, and the B tag to the verb. We have decided to display the two tokens together in the table for the sake of simplicity.

where a human annotator is perhaps better able to guess that the token refers to a nationality, while the model has very little to go off of, not just in the token, but also in the context. Finally, Example 15 shows that not all foreign-looking named entities get classified as sensitive, and that at least in the case of this sentence the model is not able to guess that a token would be sensitive just from the surrounding presence of the word "traditional" which describes it.

Both for M-BERT and KB-BERT, the models seem to run into difficulties when it comes to determining the sensitivity of data in cases where the tokens are misspelled, foreign, or surrounded by misspelled or unintelligible tokens, as in Examples 3, 7, or 10. While the model with a weighted loss function tends to flag more passages as sensitive (such as the ones in Examples 1–3, 9, and 10), the standard one errs on the side of caution in that regard (as in Examples 7 and 8, as well as 15).

One more notable feature shared by some of the under-detected PIIs is span. Most of the annotation in the data marks distinct tokens (e.g. a given name is separated from the surname, only the number of a bus or tram is marked as sensitive, etc.), and the multi-token instances are often somewhat longer passages that could be considered sensitive but do not fit into any of the categories in the annotation guidelines, e.g. talking about a political event or work status (e.g. being unemployed), as in Example 17. This shows how difficult detecting PII and determining what that concept means is, especially in the case where contextual information is essential for resolving whether a token is sensitive.

5 Conclusions

Within this paper we have presented the results of an investigation into the performance of LLMs on PII detection in learner essays, framing it as a task similar to Named Entity Recognition. We have shown that a finetuned `KB/bert-base-swedish-cased` model is capable of learning how to distinguish between sensitive and non-sensitive information in this kind of data, reaching up to 90.17% recall, suggesting that LLMs are able to approximate a human intuition when it comes to discerning what is sensitive in a given context, although they may struggle with overdetecting such data. We are also of the opinion that some of the model's disagreements with the original PII annotation could be informative when it comes to refining manual PII annotation, though perhaps not to the extent we would have wished for (the models did not discover any new kinds of PII).

While the current performance of the models is behind the ones presented by Grancharova and Dalianis (2021) (although they are relatively close in terms of recall) and the one discussed by Pilán et al. (2022) (comparing our top two models, one is slightly ahead in precision, but much worse in recall, while the other one has a similar recall with much worse precision), they are promising for PII detection in unstructured and non-standard texts in Swedish, and — with some improvements — a fine-tuned system like this could constitute a part of a pseudonymization pipeline. The current challenge is optimizing the model's hyperparameters so as to maximize the recall at the least possible cost to precision. In its current form, a weighted loss function does not seem to perform its function, but some method of accounting for class imbalance is necessary given the models' low performance on the I class.

Simultaneously, when discussing the performance of our models in relation to the ones reported by Grancharova and Dalianis (2021) we consider it relevant to mention that the latter were trained and tested on various medical datasets. We consider the medical domain to be much more regular in terms of the kinds of PII it may include (corresponding, in large part, to what the authors of that paper described as named entities), as well as less likely to include errors of various kinds. Therefore, PII detection in learner essays seems to us to be a more difficult task than PII detection in medical data.

6 Future Work

Aside from trying to optimize the model for this particular kind of data, we would like to see how well a model trained on our data would perform on other PII datasets for Swedish like the Stockholm EPR PHI Corpus, which consists of medical records or data from social media, which would also allow us to see what kinds of PII are present across domains, and what kinds are more domain-specific (Velupillai et al., 2009; Dalianis and Velupillai, 2010). Unfortunately, the TAB corpus mentioned earlier in the paper is in English, and therefore not suitable for such a comparison (Pilán et al., 2022).

Another step could be investigating to what extent the data from various domains like this can be combined in the fine-tuning process, possibly in

a semi-supervised fashion, in order to produce a more universal PII detection model. The insights from the analysis of model predictions could help determine how to annotate data for sensitivity. In terms of the differences between KB-BERT and M-BERT it would be interesting to see whether the poor performance of the latter was indeed due to it being worse at handling misspelled tokens. It would also be really interesting to be able to utilize a Swedish version of the LongFormer architecture in order to see if more contextual information helps with PII detection — but, unfortunately, no such model exists as of now (Beltagy et al., 2020).

Finally, we aim to follow up this experiment with a pseudonym generation task where we intend to have LLMs simply generate suitable replacements for the passages flagged as sensitive, without the intermediate PII classification step, with only the surrounding context to inform the prediction.

Limitations

This paper presents only a short study, where we are not really striving to create the best possible model but we are instead more focused on exploring what personal information is and how it can be detected, with the only change from the default settings of the fine-tuning script being the use of a weighted loss function and smaller batch size (a technical constraint). Therefore, hyperparameter tuning may lead to a much better performance than the presented results.

While this approach may work well, it is not a universal solution, especially cross-linguistically, as it relies on a large language model like BERT, which need not be available for all the languages in the world.

Ethics Statement

Various kinds of linguistic data are likely to contain personal information, which has implications on how the data can be used in terms of ethics and even legality. This paper aims to investigate the use of pre-existing language models and small amounts of annotated data in a pseudonymization pipeline, possibly leading to an alleviation of this challenge.

Written consent was obtained for all the collected essays; the data was processed in accordance with the GDPR requirements and is made available individually on signing an agreement for use. At the moment of corpus release, no requirement for ethical review was relevant. The original, non-pseudonymized data is used strictly within the project, with real names never being disclosed, which is why we can share neither the data used in this paper nor the fine-tuned models.

Acknowledgements

This work has been possible thanks to the funding of two grants from the Swedish Research Council.

The project *Grandma Karl is 27 years old: Automatic pseudonymization of research data* has funding number 2022-02311 for the years 2023-2029.

The Swedish national research infrastructure Nationella Språkbanken is funded jointly by contract number 2017-00626 for the years 2018-2024, as well 10 participating partner institutions.

References

Pierre Accorsi, Namrata Patel, Lopez Cédric, Rachel Panckhurst, and Mathieu Roche. 2012. Seek&hide: Anonymising a french sms corpus using natural language processing techniques. *Linguisticae Investigationes*, 35:163–180.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

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.

Hercules Dalianis and Sumithra Velupillai. 2010. De-identifying swedish clinical text - refinement of a gold standard and experiments with conditional random fields. *Journal of biomedical semantics*, 1:6.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2019. De-identification of emails: Pseudonymizing privacy-sensitive data in a German email corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 259–269, Varna, Bulgaria. INCOMA Ltd.

Elisabeth Eder, 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.

EU EU Commission. 2016. *General data protection regulation.* Official Journal of the European Union, 59, 1-88.

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, Lisa Rudebeck, and Elena Volodina. 2021. SweLL pseudonymization guidelines.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. 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.

Sumithra Velupillai, Hercules Dalianis, Martin Duneld, and Gunnar Nilsson. 2009. Developing a standard for de-identifying electronic patient records written in swedish: Precision, recall and f-measure in a manual and computerized annotation trial. *International journal of medical informatics*, 78:e19–26.

Elena Volodina. 2024. On two SweLL learner corpora – SweLL-pilot and SweLL-gold. In *Proceedings of the Huminfra Conference (HiC 2024)*, HiC 2024. Linköping University Electronic Press.

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, and Xuan-Son Vu. 2023. Grandma Karl is 27 years old – research agenda for pseudonymization of research data.

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. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), May 23-28, 2016, Portorož, Slovenia*.

Mats Wirén, Arild Matsson, Dan Rosén, and Elena Volodina. 2019. SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. In *Selected papers from the CLARIN Annual Conference 2018*, Selected papers from the CLARIN Annual Conference 2018. Linköping University Electronic Press.

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.

Vithya Yogarajan, Bernhard Pfahringer, and Michael Mayo. 2020. A review of automatic end-to-end de-identification: Is high accuracy the only metric? *Applied Artificial Intelligence*, 34(3):251–269.

A Appendix

- GitHub repository
- `transformers` code for token classification
- Application for access to the sanitized SweLL data