# Swedish Learner Essays Revisited: Further Insights into Detecting Personal Information

**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

Personally Identifiable Information (PII) is pervasive in linguistic data, making open sharing thereof complicated from both the legal and ethical perspective. Simply redacting out the PIIs or replacing them with pseudonyms presupposes a detection step, where the personal information is identified. In this study, we expand the existing research on PII detection in unstructured data (learner essays) in Swedish, testing more Large Language Models (LLMs) on a larger amount of data. We compare three different LLMs, two Swedish (KB-BERT and AI Sweden's RoBERTa) and one multilingual (M-BERT). We found that KB-BERT tends to be better than the other models but that there is some overlap in their performance.

## 1 Introduction

A non-negligible portion of texts that could be used in research or for training language models contains Personally Identifiable Information (PIIs), i.e. elements that could lead to the reidentification of the data subject. As such, they are protected by various regulations; in the EU, the GDPR governs the usage and sharing of such data (Official Journal of the European Union, 2016). Being able to eliminate PIIs from the text enables the sharing of not only corpora intended for linguistic research or language model training, but also collections of texts relevant for research in broadly understood digital humanities or fields related to law and medicine.

For example, essays written by language learners often contain PIIs, as the students are commonly prompted to talk about themselves or their experiences, limiting the texts' shareability without any privacy-protecting measures in place. This can be an issue as this kind of data is essential in language acquisition research (e.g. Golden et al., 2017), for developing essay grading (e.g. Beigman Klebanov and Madnani, 2020; Wilkens et al., 2023; Lagutina et al., 2023), or grammatical error detection tools (e.g. Bryant et al., 2023; Grundkiewicz and Junczys-Dowmunt, 2019). There is, therefore, a strong need for protecting author identities in this specific domain, as has been underlined by Stemle et al. (2019).

The two most common ways of handling the presence of PIIs in the text are anonymization and pseudonymization. Lison et al. (2021) define the former as the "[c]omplete and irreversible removal from a dataset of any information that, directly or indirectly, may lead to a subject's data being identified," whereas the latter, according to them, consists of replacing direct identifiers with pseudonyms and retaining the mapping separately. Other researchers choose not to limit it to direct identifiers (Volodina et al., 2020). What connects both of these procedures is the step in which the personal elements are identified, which is why developing robust methods for PII identification is extremely relevant for both of these applications.

Our experiment is an extension of the one conducted by Szawerna et al. (2024) and is therefore also inspired by the work done by Grancharova and Dalianis (2021), where the ability of various Large Language Models (LLMs) to detect personal information in Swedish texts was tested. We set out to test the capabilities of three different LLMs, namely KB-BERT (Malmsten et al., 2020), AI Sweden's RoBERTa (AI Sweden), and Multilingual BERT (Devlin et al., 2018) and two versions of cross entropy loss: weighted and not weighted (Ansel et al., 2024). We use SweLL-pilot and SweLL-gold, corpora of essays written by learners of Swedish as a second language which contain PIIs, the presence of which has been manually annotated (Volodina et al., 2016; Wirén et al., 2018; Volodina et al., 2019).

While Szawerna et al. (2024) have already shown that LLMs can learn to simply distinguish between PIIs and other kinds of tokens, what we want to test in our version of this experiment is a) whether AI Sweden's RoBERTa performs better than the models tested by Szawerna et al. (2024) and b) if the performance changes noticeably with the improvements to the pre-processing and the addition of more training data.

## 2 Prior Research

While a lot of work on PII detection has already been conducted, much of it focuses on English and normative text, with the genres likely to include e.g. misspellings or nonstandard variation being underrepresented, and the bulk of the pseudonymization and anonymization efforts being focused on medical and legal data (Lison et al., 2021).

When it comes to Swedish, a significant amount of work was done on medical data, including using rule-based approaches, machine learning, and fine-tuning LLMs for the task (Dalianis, 2019; Berg et al., 2019; Berg and Dalianis, 2019, 2021; Grancharova and Dalianis, 2021). Many valuable insights pertaining to the handling of PIIs also stem from the creation of the SweLL corpus, where both manual and automatized, rule-based methods were used to detect and replace personal information (Volodina et al., 2020). The data from a pilot version of this corpus was further utilized by Szawerna et al. (2024) to fine-tune and test several models.

Since the goal of this experiment is to test the performance of an array of fine-tuned LLMs on PII detection, it is worth reviewing the results reported by Grancharova and Dalianis (2021) and Szawerna et al. (2024). Grancharova and Dalianis (2021)'s best performing model (KB-BERT fine-tuned on original, in-domain data) reaches 0.923 precision, 0.922 recall, and 0.922 F1 on the task of PII detection in the medical domain. In Szawerna et al. (2024), KB-BERT is also the basis for the best performing models; here, however, its versions fine-tuned with and without a weighted loss function excel at different aspects of the task in learner-written texts. The model without a weighted loss function has the highest precision (0.875) and F1 (0.803), whereas the one with a weighted loss function has the highest recall (0.902).

## 3 Materials and Methods

The setup of this experiment follows closely the one of Szawerna et al. (2024), albeit with a number of significant changes. While they originally used only the SweLL-pilot corpus (Volodina et al., 2016), we also include the SweLL-gold corpus[1] (Volodina et al., 2019; Wirén et al., 2018). This has doubled the number of essays and nearly tripled the instances of PIIs, as seen in Table 1[2]. The data in both of the aforementioned corpora consists of essays written by learners of Swedish as a second language, of varying levels of proficiency, but also varied in terms of topic or genre. Following the original experiment, we disregard the detailed PII type annotation, focusing only on whether a token belongs to a PII passage or not, and assigning the appropriate inside-outside-beginning (IOB) tag. It is important to note that the annotation of personal information in the two SweLL corpora was conducted by different people, though they did follow the same guidelines (Megyesi et al., 2021).

| | SweLL-pilot | SweLL-pilot +<br>SweLL-gold |
|---|---|---|
| B | 1142 | 3111 |
| I | 86 | 237 |

Table 1: The counts of the instances of B and I PII classes in Szawerna et al. (2024) and in our experiment.

The two major improvements relative to the original experiment concern the preprocessing of the samples. During their construction, we pre-tokenize them using the respective LLM's tokenizer in order to be able to obtain samples with as much context as possible, i.e. as close to 512 sub-word tokens as possible without the sample ending in the middle of a word (whereas previously samples had the maximum length of 100 tokens). Since this is dependent on the LLM tokenizer used, the number of non-PII tokens varies between models (see Table 4, Table 5, Table 6 in Appendix A). The samples obtained from the same essay are bound to be in the same data split.

We ensure that an equal number of samples that include personal information and ones that do not do that are included in our data splits. We calculate the class weights using Scikit-learn (Pedregosa et al., 2011). The exact class counts, proportions, and weights are provided in Appendix A. In the final step of the pre-processing, we perform a 5-fold cross-validation in order to obtain a better overview of their performance.

[1] SweLL-gold was originally pseudonymized, we reintroduced the PIIs into the texts in order to use this corpus in our experiment.

[2] For details concerning the types of PIIs that can be found in the data, consult Megyesi et al. (2021).

In this experiment we fine-tune three different BERT-based LLMs, using 5-fold cross-validation. The LLMs in question are the BERT model for Swedish developed by the National Library of Sweden[3] (KB-BERT, Malmsten et al. (2020)), the multilingual BERT[4] (M-BERT, Devlin et al. (2018)), as in Szawerna et al. (2024)), and – unlike in the original experiment – AI Sweden's RoBERTa[5] (AI-Sweden's RoBERTa, AI Sweden). The fine-tuning process is conducted in the same way as in Szawerna et al. (2024), utilizing modified code from the Transformers library (Wolf et al., 2020) and the same settings: a batch size of 8, 3 epochs, and the AdamW optimizer (learning rate: 05e-05).

## 4 Results and Discussion

While performing a 5-fold cross-validation of the models allows us to gain better insights into the performance of the respective LLMs on the task of PII detection, we do not have enough versions per model to provide a reliable statistical analysis. This is due to the computational requirements of fine-tuning these models. Therefore, in Table 2 and Table 3 we only report the mean scores across the folds, alongside the standard deviation (Numbers in bold indicate the highest score across both tables). We highlight the best performances in bold. We have also selected to focus our analysis on the weighted averages of precision, recall, F1, and F2[6] scores only across the PII classes (disregarding the scores for the non-PII tokens) – following Grancharova and Dalianis (2021) and Szawerna et al. (2024). Focusing only on the PII classes allows us to compensate for the class imbalance of the non-PII vs. PII classes. We chose to report F2 since it gives more importance to recall, which is essential for reflecting how many of the PIIs were detected (and, therefore, how successful the model is at protecting the data subjects). Nevertheless, precision is important as well, since we want to tamper with the data as little as necessary. Detailed per-model results and a wider selection of measures are provided in Appendix A.

In terms of per-PII-class precision, Szawerna et al. (2024) report 0.8748 as the highest score, obtained by their KB-BERT model fine-tuned without a weighted loss function. In our case, the best performance is also obtained by the same model and loss function combination, but the actual score drops to 0.857 with a standard deviation of 0.036.

As far as the per-PII-class recall is concerned, Szawerna et al. (2024)'s best model is KB-BERT with a weighted loss function, which scores 0.902. Among our models, once again the same model prevails, with 0.883 recall and an STD of 0.037.

We do, however, note that our best F1 score is higher than that reported by Szawerna et al. (2024). Theirs was of KB-BERT without a weighted loss function at 0.803, while ours – for the same model and loss function – is 0.810, with a standard deviation of 0.039. It is worth pointing out, however, that the STD for that score is quite big (0.054), meaning that there is larger variety between recall scores.

| KB-BERT | | | | |
|---|---|---|---|---|
| | Precision | Recall | F1 | F2 |
| Average | **0.857** | 0.784 | **0.810** | 0.793 |
| STD | 0.036 | 0.054 | 0.039 | 0.047 |
| M-BERT | | | | |
| | Precision | Recall | F1 | F2 |
| Average | 0.831 | 0.775 | 0.794 | 0.782 |
| STD | 0.027 | 0.033 | 0.026 | 0.030 |
| AI-Sweden's RoBERTa | | | | |
| | Precision | Recall | F1 | F2 |
| Average | 0.690 | 0.653 | 0.665 | 0.657 |
| STD | 0.387 | 0.367 | 0.372 | 0.368 |

Table 2: Measures across only the PII classes for the models without a weighted loss function.

| KB-BERT | | | | |
|---|---|---|---|---|
| | Precision | Recall | F1 | F2 |
| Average | 0.619 | **0.883** | 0.727 | **0.813** |
| STD | 0.032 | 0.037 | 0.031 | 0.032 |
| M-BERT | | | | |
| | Precision | Recall | F1 | F2 |
| Average | 0.625 | 0.858 | 0.721 | 0.797 |
| STD | 0.036 | 0.043 | 0.032 | 0.035 |
| AI-Sweden's RoBERTa | | | | |
| | Precision | Recall | F1 | F2 |
| Average | 0.261 | 0.354 | 0.300 | 0.330 |
| STD | 0.360 | 0.486 | 0.413 | 0.453 |

Table 3: Measures across only the PII classes for the models with a weighted loss function.

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

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

[5] https://huggingface.co/AI-Sweden-Models/roberta-large-1160k

[6] $F_\beta = (1+\beta^2) * \frac{precision*recall}{(\beta^2*precision)+recall}$ where $\beta = 2$

While the F2 score has not previously been reported for this task, we note that if we consider recall more important than precision, then this combined score elevates KB-BERT with a weighted loss function, as it achieved F2 equal to 0.813 (STD: 0.032).

Notably, with the changes that we have introduced we no longer see the catastrophic drop in the performance of the M-BERT model fine-tuned with a weighted loss function reported by Szawerna et al. (2024). Interestingly, though, we do note that a similar effect can be observed in AI-Sweden's RoBERTa when using weighted loss, where the scores revolve around 30% on average, with large standard deviations. When inspecting the per-model scores, we noted that three out of five runs seem to have completely stopped predicting the PII classes. This is also true for one out of five RoBERTa runs without a weighted loss function (resulting in a large STD for those as well). This could very well be due to the large imbalance of the PII classes versus the non-PII tokens, but further experiments would be needed to confirm whether that is the case.

What is worth noting is that the different folds for AI-Sweden's RoBERTa lead to very inconsistent performances, and while on the results of some of those models on their own are quite high, the fine-tuning of this LLM is not reliable with the current setup.

## 5 Conclusions

Within this experiment we attempted to improve the PII detection model introduced by Szawerna et al. (2024), increase the amount of training data, and evaluate one more Large Language Model's performance. When it comes to the best performing models, we did not note any changes that are likely to be statistically significant. Our changes and improvements have, however, led to eliminating the issues with M-BERT with a weighted loss function.

We have also observed that the current setup leads to very inconsistent results when fine-tuning AI-Sweden's RoBERTa, but we note that singular results from those models exceed those of any other model. This suggests that better results can be obtained using the latter model, but that requires us to eliminate the issues leading to the over-detection of the “outside" class.

This experiment solidifies the previous findings that simplifying the personal information detection task down to whether a token is personal or not given the context (but disregarding further subclassification) is a valid method, and hopefully contributes to the efforts of building a pipeline for anonymizing or pseudonymizing a wider variety of Swedish texts.

## 6 Future Work

Given the results reported for AI-Sweden's RoBERTa, a natural continuation of this experiment would be to identify and neutralize the issues causing the model's inconsistent performance. Since the model is only trained on one domain of data (learner essays), it would be interesting to see how it performs on other kinds of data, or how mixing data from various domains will affect the performance. Another idea would be to compare IOB-based PII detection to more detailed classification.

## Limitations

We would like to draw the attention to the limitations concerning the fine-tuning of the models that are present in our experiment. With a total of 6 model and loss function combinations, it would have been very computationally expensive to fine-tune more versions; however, this has a negative impact on our ability to make statistically relevant comparisons. Additionally, better scores are likely possible with some hyper-parameter tweaking.

## Ethics Statement

When working with personal information, the safety and privacy of our data subjects is paramount. Since our training data contains such information, we can share neither the data, nor the fine-tuned models. For the same reason we can only use models which can be run locally, without uploading the data to any third parties.

## Acknowledgments

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

AI Sweden. AI-Sweden-models/Roberta-large-1160K.

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

Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.

Hanna Berg, Taridzo Chomutare, and Hercules Dalianis. 2019. Building a de-identification system for real Swedish clinical text using pseudonymised clinical text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 118–125, Hong Kong. Association for Computational Linguistics.

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

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

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical Error Correction: A Survey of the State of the Art. *arXiv preprint arXiv:2211.05166*.

Hercules Dalianis. 2019. Pseudonymisation of Swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 16–23, Turku, Finland. Linköping Electronic Press.

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*.

Anne Golden, Scott Jarvis, and Kari Tenfjord. 2017. *Crosslinguistic influence and distinctive patterns of language learning: Findings and insights from a learner corpus*, volume 118. Multilingual Matters.

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.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2019. Minimally-augmented grammatical error correction. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 357–363, Hong Kong, China. Association for Computational Linguistics.

Nadezhda Stanislavovna Lagutina, Kseniya Vladimirovna Lagutina, Anastasya Mikhailovna Brederman, and Natalia Nikolaevna Kasatkina. 2023. Text classification by CEFR levels using machine learning methods and BERT language model. *Modelirovanie i Analiz Informatsionnykh Sistem*, 30(3):202–213.

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. *Preprint*, arXiv:2007.01658.

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.

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).

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.

Egon W Stemle, Adriane Boyd, Maarten Jansen, Therese Lindström Tiedemann, Nives Mikelić Preradović, Alexandr Rosen, Dan Rosén, and Elena Volodina. 2019. Working together towards an ideal infrastructure for language learner corpora. *Widening the Scope of Learner Corpus Research*.

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.

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, 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.

Rodrigo Wilkens, Alice Pintard, David Alfter, Vincent Folny, and Thomas François. 2023. TCFLE-8: a corpus of learner written productions for French as a foreign language and its application to automated essay scoring. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3465, Singapore. Association for Computational Linguistics.

Mats Wirén, Arild Matsson, Dan Rosén, and Elena Volodina. 2018. SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. In *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018*, Linköpings universitet. Linköping University Electronic Press, Linköpings universitet.

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. *Preprint*, arXiv:1910.03771.

# A Appendix: Class Counts, Weights, and Detailed Results

| | Instances (%) | Count | Weight |
|---|---|---|---|
| B | 1.41% | 3111 | 23.66 |
| I | 0.11% | 237 | 310.52 |
| O | 98.48% | 217430 | 0.33 |

Table 4: The proportions of token instances of classes in the data used in the experiment and the corresponding calculated class weights for experiments with KB-BERT.

| | Instances (%) | Count | Weight |
|---|---|---|---|
| B | 1.60% | 3111 | 20.78 |
| I | 0.12% | 237 | 272.82 |
| O | 98.27% | 190626 | 0.34 |

Table 5: The proportions of token instances of classes in the data used in the experiment and the corresponding calculated class weights for experiments with M-BERT.

| | Instances (%) | Count | Weight |
|---|---|---|---|
| B | 1.79% | 3111 | 18.59 |
| I | 0.14% | 237 | 244.013 |
| O | 98.07% | 170145 | 0.34 |

Table 6: The proportions of token instances of classes in the data used in the experiment and the corresponding calculated class weights for experiments with AI-Sweden's RoBERTa.

|  | Accuracy | Precision | Recall | F1 | F2 | MCC | Sensitive Precision | Sensitive Recall | Sensitive F1 | Sensitive F2 |
|---|---|---|---|---|---|---|---|---|---|---|
| Model 1 | 0.995294 | 0.995074 | 0.995294 | 0.994897 | 0.995117 | 0.798232 | 0.862433 | 0.737319 | 0.777104 | 0.751789 |
| Model 2 | 0.995719 | 0.995696 | 0.995719 | 0.995707 | 0.995714 | 0.861316 | 0.857771 | 0.856305 | 0.856995 | 0.856572 |
| Model 3 | 0.994029 | 0.993586 | 0.994029 | 0.993698 | 0.993886 | 0.785669 | 0.796362 | 0.745710 | 0.763545 | 0.752142 |
| Model 4 | 0.994902 | 0.994634 | 0.994902 | 0.994638 | 0.994774 | 0.820049 | 0.887983 | 0.754545 | 0.812867 | 0.776487 |
| Model 5 | 0.995202 | 0.995131 | 0.995202 | 0.994900 | 0.995057 | 0.856825 | 0.880130 | 0.825338 | 0.837538 | 0.828830 |
| K-fold mean | 0.995029 | 0.994824 | 0.995029 | 0.994768 | 0.994910 | 0.824418 | 0.856936 | 0.783843 | 0.809610 | 0.793164 |
| K-fold STD | 0.000631 | 0.000788 | 0.000631 | 0.000721 | 0.000667 | 0.033978 | 0.036062 | 0.053501 | 0.039416 | 0.047343 |

Table 7: Detailed results for the KB-BERT model without a weighted loss function

|  | Accuracy | Precision | Recall | F1 | F2 | MCC | Sensitive Precision | Sensitive Recall | Sensitive F1 | Sensitive F2 |
|---|---|---|---|---|---|---|---|---|---|---|
| Model 1 | 0.990770 | 0.993043 | 0.990770 | 0.991582 | 0.990991 | 0.707881 | 0.580454 | 0.847826 | 0.688978 | 0.776190 |
| Model 2 | 0.989598 | 0.993134 | 0.989598 | 0.990823 | 0.989900 | 0.752192 | 0.617264 | 0.932551 | 0.741469 | 0.844452 |
| Model 3 | 0.989888 | 0.992136 | 0.989888 | 0.990684 | 0.990103 | 0.718842 | 0.597128 | 0.845554 | 0.699533 | 0.780258 |
| Model 4 | 0.991518 | 0.993345 | 0.991518 | 0.992142 | 0.991680 | 0.770238 | 0.659364 | 0.898485 | 0.760498 | 0.837663 |
| Model 5 | 0.990039 | 0.992246 | 0.990039 | 0.990780 | 0.990223 | 0.765317 | 0.642115 | 0.888069 | 0.745001 | 0.824611 |
| K-fold mean | 0.990362 | 0.992781 | 0.990362 | 0.991202 | 0.990580 | 0.742894 | 0.619265 | 0.882497 | 0.727096 | 0.812635 |
| K-fold STD | 0.000777 | 0.000551 | 0.000777 | 0.000636 | 0.000741 | 0.028024 | 0.032134 | 0.036602 | 0.031047 | 0.032244 |

Table 8: Detailed results for the KB-BERT model with a weighted loss function

|  | Accuracy | Precision | Recall | F1 | F2 | MCC | Sensitive Precision | Sensitive Recall | Sensitive F1 | Sensitive F2 |
|---|---|---|---|---|---|---|---|---|---|---|
| Model 1 | 0.994317 | 0.993833 | 0.994317 | 0.994007 | 0.994183 | 0.785353 | 0.809423 | 0.739130 | 0.769499 | 0.750605 |
| Model 2 | 0.993883 | 0.993799 | 0.993883 | 0.993810 | 0.993850 | 0.823782 | 0.822239 | 0.816716 | 0.817711 | 0.816901 |
| Model 3 | 0.994167 | 0.993734 | 0.994167 | 0.993742 | 0.993981 | 0.799651 | 0.830278 | 0.748752 | 0.776084 | 0.758689 |
| Model 4 | 0.994300 | 0.994203 | 0.994300 | 0.994044 | 0.994178 | 0.834744 | 0.877261 | 0.798571 | 0.826560 | 0.808725 |
| Model 5 | 0.992593 | 0.992039 | 0.992593 | 0.992019 | 0.992343 | 0.812006 | 0.818028 | 0.772448 | 0.780969 | 0.774885 |
| K-fold mean | 0.993852 | 0.993522 | 0.993852 | 0.993524 | 0.993707 | 0.811107 | 0.831446 | 0.775123 | 0.794164 | 0.781961 |
| K-fold STD | 0.000725 | 0.000849 | 0.000725 | 0.000851 | 0.000775 | 0.019459 | 0.026694 | 0.032703 | 0.026045 | 0.029631 |

Table 9: Detailed results for the M-BERT model without a weighted loss function

|  | Accuracy | Precision | Recall | F1 | F2 | MCC | Sensitive Precision | Sensitive Recall | Sensitive F1 | Sensitive F2 |
|---|---|---|---|---|---|---|---|---|---|---|
| Model 1 | 0.989526 | 0.991487 | 0.989526 | 0.990166 | 0.989676 | 0.691246 | 0.571514 | 0.809783 | 0.666948 | 0.745238 |
| Model 2 | 0.988573 | 0.992005 | 0.988573 | 0.989773 | 0.988885 | 0.750397 | 0.623565 | 0.913490 | 0.740544 | 0.835067 |
| Model 3 | 0.991328 | 0.992283 | 0.991328 | 0.991608 | 0.991388 | 0.747359 | 0.664426 | 0.818636 | 0.728981 | 0.779445 |
| Model 4 | 0.989250 | 0.991456 | 0.989250 | 0.990032 | 0.989466 | 0.752611 | 0.649001 | 0.877143 | 0.746002 | 0.819511 |
| Model 5 | 0.986974 | 0.989899 | 0.986974 | 0.987942 | 0.987206 | 0.742949 | 0.616482 | 0.870849 | 0.720684 | 0.803483 |
| K-fold mean | 0.989130 | 0.991426 | 0.989130 | 0.989904 | 0.989324 | 0.736912 | 0.624998 | 0.857980 | 0.720632 | 0.796549 |
| K-fold STD | 0.001578 | 0.000923 | 0.001578 | 0.001309 | 0.001507 | 0.025784 | 0.035587 | 0.043258 | 0.031590 | 0.035300 |

Table 10: Detailed results for the M-BERT model with a weighted loss function

|  | Accuracy | Precision | Recall | F1 | F2 | MCC | Sensitive Precision | Sensitive Recall | Sensitive F1 | Sensitive F2 |
|---|---|---|---|---|---|---|---|---|---|---|
| Model 1 | 0.994879 | 0.994370 | 0.994879 | 0.994500 | 0.994714 | 0.828151 | 0.852875 | 0.780037 | 0.808646 | 0.790642 |
| Model 2 | 0.993938 | 0.994188 | 0.993938 | 0.994050 | 0.993979 | 0.849293 | 0.832115 | 0.866084 | 0.848433 | 0.858830 |
| Model 3 | 0.981818 | 0.963966 | 0.981818 | 0.972810 | 0.978195 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| Model 4 | 0.994358 | 0.994217 | 0.994358 | 0.993958 | 0.994171 | 0.846429 | 0.910588 | 0.789781 | 0.833562 | 0.805941 |
| Model 5 | 0.993385 | 0.993212 | 0.993385 | 0.993181 | 0.993286 | 0.848825 | 0.852647 | 0.830189 | 0.836233 | 0.831862 |
| K-fold mean | 0.991675 | 0.987991 | 0.991675 | 0.989700 | 0.990869 | 0.674540 | 0.689645 | 0.653218 | 0.665375 | 0.657455 |
| K-fold STD | 0.005538 | 0.013438 | 0.005538 | 0.009454 | 0.007104 | 0.377180 | 0.386632 | 0.366762 | 0.372236 | 0.368444 |

Table 11: Detailed results for the AI-Sweden's RoBERTa model without a weighted loss function

|  | Accuracy | Precision | Recall | F1 | F2 | MCC | Sensitive Precision | Sensitive Recall | Sensitive F1 | Sensitive F2 |
|---|---|---|---|---|---|---|---|---|---|---|
| Model 1 | 0.988704 | 0.991199 | 0.988704 | 0.989603 | 0.988955 | 0.708627 | 0.591056 | 0.837338 | 0.692827 | 0.772809 |
| Model 2 | 0.990878 | 0.993067 | 0.990878 | 0.991648 | 0.991081 | 0.808050 | 0.715271 | 0.931587 | 0.807780 | 0.876952 |
| Model 3 | 0.981818 | 0.963966 | 0.981818 | 0.972810 | 0.978195 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| Model 4 | 0.980282 | 0.960952 | 0.980282 | 0.970520 | 0.976354 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| Model 5 | 0.977331 | 0.955176 | 0.977331 | 0.966127 | 0.972818 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| K-fold mean | 0.983803 | 0.972872 | 0.983803 | 0.978142 | 0.981480 | 0.303335 | 0.261266 | 0.353785 | 0.300121 | 0.329952 |
| K-fold STD | 0.005751 | 0.017876 | 0.005751 | 0.011669 | 0.008065 | 0.416844 | 0.360438 | 0.485585 | 0.412963 | 0.453304 |

Table 12: Detailed results for the AI-Sweden’s RoBERTa model with a weighted loss function