Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1098–1103
Marseille, 20-25 June 2022

Efficiently and Thoroughly Anonymizing a Transformer Language Model for Dutch Electronic Health Records: a Two-Step Method

Stella Verkijk, Piek Vossen
VU University Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam, Netherlands
stellaverkijk@outlook.com, p.t.j.m.vossen@vu.nl

Abstract

Neural Network (NN) architectures are used more and more to model large amounts of data, such as text data available online. Transformer-based NN architectures have shown to be very useful for language modelling. Although many researchers study how such Language Models (LMs) work, not much attention has been paid to the privacy risks of training LMs on large amounts of data and publishing them online. This paper presents a new method for anonymizing a language model by presenting the way in which MedRoBERTa.nl, a Dutch language model for hospital notes, was anonymized. The two step method involves i) automatic anonymization of the training data and ii) semi-automatic anonymization of the LM's vocabulary. Adopting the fill-mask task where the model predicts what tokens are most probable to appear in a certain context, it was tested how often the model will predict a name in a context where a name should be. It was shown that it predicts a name-like token 0.2% of the time. Any name-like token that was predicted was never the name originally presented in the training data. By explaining how a LM trained on highly private real-world medical data can be safely published with open access, we hope that more language resources will be published openly and responsibly so the community can profit from them.

Keywords: Anonymization, Language Model, Medical Text Data

1. Introduction

Deep learning with neural networks (NNs) has taken the machine learning world by storm, becoming the foundation of new AI-based services (Shokri and Shmatikov, 2015; Shickel et al., 2017). BERT by Google (Devlin et al., 2019) and RoBERTa by Facebook (Liu et al., 2019), Transformer-based NNs trained on the enormous amounts of language data available online like Wikipedia pages, have shown enormous success in the modelling of language. Such Language Models (LMs) now serve as the base for many specific NLP systems like Machine Translation, Concept Extraction, Question Answering and many more. However, a downside of these language models and NNs in general is that it is not entirely clear yet how they work. They are often referred to as black-box-like systems (Rogers et al., 2020). Among other concerns, this calls for vigilance for privacy breaches. Even if the training data is sourced online and is therefore not seen as protected information, a model trained on this data can be used to systematically search for specific information a malevolent user might be looking for. If the training data was not sourced online and does contain sensitive information, the model itself might be used to recover patterns the model has seen in the training phase.

Applying deep learning has also shown successful for clinical informatics tasks (Shickel et al., 2017). Notes containing information about the well-being of patients taken by hospital staff during the course of treatments (hospital notes) included in Electronic Health Records (EHRs) contain valuable information. If this information could be automatically and accurately analysed on a large scale, with the help of a specialized language model, this could lead to the production of real-world applications like clinical decision support systems, early symptom detection, prediction of re-hospitalization, fall prediction systems, and many other tools. However, since the data contained in EHRs is highly private, it is problematic to build and publish LMs for this type of data (Shokri and Shmatikov, 2015).

In this paper, we show how MedRoBERTa.nl (Verkijk and Vossen, 2022), a language model for Dutch hospital notes, was anonymized efficiently and accurately with a novel two-step method: i) automatic anonymization of the training data and ii) semi-automatic anonymization of the LM's vocabulary. The proposed method can be applied to any kind of LM. We discuss the risks of using and publishing neural networks trained on sensitive data and we offer solutions so that innovative NLP systems, among which those for the medical field, can be safely rely on such LMs and shared among researchers.

Section 2 will briefly explain how MedRoBERTa.nl was built. In Section 3, we address the risks of publishing Deep Learning models and identify possible privacy leaks when publishing a LM. We explain how MedRoBERTa.nl was anonymized in Section 4, after which we will provide results of a final anonymity test in Section 5. In Section 6, we will conclude, discussing some limitations and considerations for further research.

2. MedRoBERTa.nl

MedRoBERTa.nl (Verkijk and Vossen, 2022) is the first domain specific LM for the free text contained in Dutch Electronic Health Records (EHRs). It was trained on nearly 10 million hospital notes (comprising more than 13.2 GB of data) provided by the Amsterdam University Medical Centres (AUMC). The model was trained on the hospital's highly secured server. The training was done from scratch with random pre-initialisation and a new, specialized vocabulary. The vocabulary serves to link tokens the model encounters to its learned embeddings. The usage of a new vocabulary assures that the lexical variation of medical text is well represented in the model's learned embeddings (any token the model encounters but is not represented in the vocabulary has to be broken down in smaller bits, resulting in a less accurate embedding representation). Verkijk and Vossen (2022) show that the model performs better at medical NLP tasks than general language models for Dutch such as BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020).

Although LMs are not designed for the generation of data, such models can be used to predict single words if you feed it a sentence with a left-open space where a word should be. This is called the fill-mask task. For example, after feeding MedRoBERTa.nl a sentence like 'Mr *mask* has been diagnosed with Covid', it predicts which tokens are most likely to be in the masked spot (and vice versa: when giving it a name it could generate the diagnosis). Therefore, it is very important to anonymize the model in a way that ensures that no information can ever be linked to an individual.

3. Possible privacy leaks when publishing a Language Model

In their survey about privacy in deep learning, Mireshghallah et al. (2020) describe how private information can be discovered by querying published Machine Learning systems, and specifically neural networks. We use their analysis to identify possible threats.

Mireshghallah et al. (2020) identify two classes of threats: Direct Information Exposure (DIE) and Indirect (Inferred) Information Exposure (IIE). DIE applies to cases where unauthorized individuals and systems have direct access to the data from which a model is built. In the case of IEE, only the model itself is accessible. Having trained the model on an external, highly secured server and having assured that no data ever left the server, DIE was not a possible threat in the process of creating and publishing MedRoBERTa.nl. However, some forms of IIE were.

According to Mireshghallah et al. (2020), there are five types of IIE: Membership Inference, Model Inversion and Attribute Inference, Hyperparameter Inference, Parameter Inference and Property Inference. Hyperparameter Inference and Parameter Inference indicate that someone could steal the model, the intellectual property of the owner, by inferring how it was built. Since it was decided to publish MedRoBERTa.nl with open access[1] as well as the code that was used to build it[2], Hyperparameter Inference and Parameter Inference were no threats. Membership Inference entails attacks where a user applies techniques to infer whether some item, or in our case, a piece of text, has been part of the training data by analysing the confidence score for a queried input. During Model Inversion, the predictions of a model are analysed to reveal whether the prediction has literally been seen previously in the training data. Hence, Model Inversion is a form of Membership Inference. Property Inference attacks try to infer specific patterns of information from the target model by extracting out-of-distribution training data that the model has memorized. Membership Inference, Model Inversion and Property Inference are all attacks that aim to reveal private information by extracting parts of the training data. Mireshghallah et al. (2020) conclude that until now little attention has been paid to privacy leaks occurring during the inference phase in particular.

For a language model like MedRoBERTa.nl, this means the threat lies in the generative function the model has. MedRoBERTa.nl was trained with the MASK learning objective, where randomly chosen words are masked in the training input and the model learns to predict these words (Devlin et al., 2019). As a form of model inversion, the model can thus be used to predict masked words in sentences a user provides it with. Along with the predicted words, the model provides a probability score, which, when analysed, can be used to infer which words were part of the training data. Shokri et al. (2017) demonstrate how just a model's probability scores can be used for membership interference by turning it into a classification problem. They created several 'shadow models' that they trained on synthetic data generated from the target model itself (following the intuition that inputs that are classified by the target model with high confidence should be statistically similar to the target's training data set). These models imitate the behavior of the target model. They then trained an attack model on the labeled inputs and outputs of the shadow models, and show an overall precision of 0.895 on a membership inference attack against a Google-trained model.

In the following paragraphs, we will explain how we made sure no prediction in a so-called *fill-masked-name* task can ever be claimed to have been part of the training data.

[1] https://huggingface.co/CLTL/MedRoBERTa.nl

[2] https://github.com/cltl-students/verkijk_stella_rma_thesis_dutch_medical_language_model

4. Anonymizing MedRoBERTa

4.1. Introduction

An ample amount of research has been published on the anonymization of electronic patient records or other health data in order to make them available for researchers (Marimon et al., 2019; Uzuner et al., 2008; Stubbs et al., 2015). However, the anonymization of computational models, let alone a state-of-the-art language model, is a new problem that has not received much attention yet. A language model internally bases its decisions on contextual word embeddings. This means it attributes a word's definition to the words that tend to be surrounding it. In that way, it learns to make associations between words or phrases based on their contexts. To make sure the model never returns a name with which it has strong associations in the form of a specific piece of information, like a phrase or a word, we adopt a two-step anonymization method. First, the training data is anonymized as much as is possible with automatic methods by replacing names by anonymous labels, and then the model's vocabulary is anonymized as an additional step. In the following sections, we will explain these two steps individually, provide results of an anonymity test that was conducted, and end with some reflections on what this mean.

4.2. Anonymizing the training data

As the training data consisted of more than 13 GB of text, it was impossible to manually perform de-identification of patients as well as doctors. Therefore, spaCy (Honnibal and Montani, 2017) was used to automatically replace individuals' names with *PERSON* and countries', states' and cities' names with *GPE*. There are several Dutch spaCy models available for this task. For the anonymization of the training data, the *nl_core_news_lg* model was used. This model for Named Entity Recognition (NER) was trained on the training set of the LassySmall corpus (Van Noord et al., 2013) which spaCy augmented with NER annotations using NLP Town. The training set consisted of 6641 sentences.

This anonymization step is not perfect: some names will remain undetected by spaCy. However, it places the token *PERSON* in contexts that are normally filled by names of people for the vast majority of the time. This basically means that the model processes this '*PERSON*' as a single individual that occurs very often in the dataset. In this way, *PERSON* pushes away associations the model has with other names that may still be present. On top of that, tokens like *Mr* (mr.), *Mw*, (mrs.) and *pt* (patient), without the inclusion of a name or last name, are already much more frequent in the training data than names. This assures that the model will strongly associate all people-like information with *PERSON* or with tokens like the ones mentioned before more than with any other token. Any other named entities, like names of organizations and locations, were not anonymized, because preliminary experiments showed that spaCy confuses locations' and organizations' names very often with names of illnesses, medicines or other medical terms. The absence of these terms in the pre-training data was seen as too much of a loss of information for the medical language model, while the privacy of these organisations was considered less delicate.

4.3. Anonymizing the vocabulary

As mentioned in Section 2, a language model uses a vocabulary to link the tokens it encounters in the data to an embedding. Any words it encounters that are not in the vocabulary will be broken down into smaller pieces (bytes) for the model to be able to process the word. Regarding language generation, the model can only predict words that are in its vocabulary. This means that any name the model has seen in the training data but is not represented in the vocabulary can never be regenerated, for example when performing the fill-mask task. The vocabulary of MedRoBERTa.nl was created by gathering the 52.000 most frequent words in the training data that was anonymised using spaCy.

To find any names that were frequent enough to end up in the medical models' vocabulary despite of the anonymization of the training data, we looked for the words the model deemed most similar to *PERSON*. In order to do this, a development set of sentences was created to perform the *fill-masked-name* task with MedRoBERTa.nl. This was done as follows. First, two large datasets of 8k sentences were collected where *PERSON* occurred: one from seen data (the pre-training data) and one from unseen data. The unseen data was the validation data used during pre-training: data that came from the same source as the training data, but was not used to train the model on. From each of these large data sets, 100 sentences each were selected manually for which it was clear that there must have been a name that now was replaced by *PERSON*. We performed this manual selection because spaCy sometimes replaced other tokens than people's names (names of medicines, ilnesses etc.) with *PERSON*.

For each sentence in the combined data set of 200 items, PERSON was replaced with 'mask '. We used MedRoBERTa.nl to perform the *fill-masked-name task* on this development set. For each sentence, the top 40 predictions were gathered. By taking such a large sample of predictions per sentence, we made sure that not only tokens that are highly similar to *PERSON* would be returned (*mw* or *mr*, for example) but also those that are remotely similar [3]. The list of tokens that was gathered in this process therefore also contained many tokens that were not names. These tokens ranged from anything like *mr* (mister) and *pt* (patient) to *oxazepam* (a medicine) and even function words such as *dus* (so/therefore). Next, the list was checked manually

[3]The model provides a score for the full vocabulary, which means all words receive some probability score eventually.

several times to extract tokens that had the possibility of being names. This meant, for example, that all nouns with a capital letter were selected as possible names.
The resulting list of names from the predictions on the sentences from unseen data had a large amount of overlap with the list of names gathered from the predictions on the seen sentences. From the total amount of 144 names that were gathered from the predictions on the unseen sentences, 97 were also in the predictions for the seen sentences. This suggests that the combined collection of names is a solid representation of the names the model has made any associations with. The final set of 212 unique names was then taken out of the vocabulary by replacing each name with the token *unk* (from 'unknown') and a randomly generated number. In this way, anytime the model wants to generate a name that has been taken out of the vocabulary, it will generate *unk7783002*, for example.

5. Testing Anonymity

A final test was performed to assess how big the chance is that the model will still predict a name after the two-step method was completed. From the two larger development sets of 8k sentences, two new test sets were created by manually selecting relevant sentences (in the same way the smaller developments sets were created). For some examples of these sentences, see examples 1-5[4]. We made sure that these were different sentences than the ones in the development sets. For the combined data set of 200 sentences from seen and unseen data, the first 20 predictions for the *fill-masked-name* task were collected. This resulted in a total of 4000 predictions. Of these 4000 predictions, 8 tokens were name-like. 4 of these tokens appeared in the predictions for the unseen sentences, and 4 in the predictions for the seen sentences. For 192 out of 200 sentences, no name was ever predicted. The 8 names that were predicted were never predicted more than once and they were never the first, most probable prediction. The highest ranking name-like token for the unseen sentences was in the 14th position. For the seen data, the highest ranking name-like token was in the 6th position. For the four sentences in the unseen data where a name-like token was predicted, the original sentences could be retrieved from the original training data (from before anonymizing it with spaCy). For all four sentences, the name predicted by the model was not the name that originally occurred. For an overview of all predictions that were made four times or more by the model, see Table 1 (predictions on the seen data) and Table 2 (predictions on the unseen data).

(1) <mask> vertelt dat hele gezin heeft al de buikgriep gehad.
(<mask> says the whole family already had the stomach flu.)

(2) Zekerheidshalve werd <mask> naar de oogarts verwezen.
(To be on the safe side, <mask> was referred to the ophthalmologist.)

(3) Met vriendelijke groet, <mask>, Logopedist, logopedist.
(Kind regards, <mask>, Speech Therapist, speech therapist.)

(4) PATIENTGEGEVENS Naam <mask>.
(PATIENT RECORDS Name <mask>.)

(5) Mob: <mask>was erg vermoeid, wilde niet mobiliseren, dus niet aan toegekomen om dhr te wegen.
(Mob: <mask>was very tired, didn't want to mobilize, so didn't get around to weighing mr.)

We presented the model to the privacy office of the Amsterdam University Medical Centre (AUMC) to consider any privacy risks. We documented the procedure, the test results and the source code for anonymizing the data and building the model. We had several meetings with the privacy office and provided them with a thorough report. After very careful consideration, the privacy office of the AUMC granted permission for open access publication of MedRoBERTa.nl, which is now available on the huggingface.com platform for building fine-tuned models in the medical field, as well as for the anonymization test set that consists of sentences that were not part of the pre-training data of MedRoBERTa.nl.

6. Conclusion

We have shown how a language model trained on highly sensitive information can be anonymized in such a way that it can be released to the public. The two-step method of firstly anonymizing the training data and later the LM's vocabulary can be applied to any state-of-the-art Transformer-based language model.
Although the manual inspection of the data in order to anonymize the vocabulary can be seen as non-efficient, the manual work is severely reduced by anonymizing the pre-training data beforehand. Since even the best packages for automatic anonymization are not perfect, combining automatic anonymization and manual inspection leads to more secure results.
For future research, we hope that more attention is paid to the risks of privacy leaks in neural networks and especially language models. It would also be interesting to investigate if or to what extent anonymization leads to a loss of predictive power. We hope that our paper can help other researchers to share their work safely with the scientific community.

[4]See `https://github.com/cltl-students/verkijk_stella_rma_thesis_dutch_medical_language_model/tree/master/src/anonymization` for the complete test data set of unseen sentences

Token	Times predicted t
PERSON	179
GPE	113
Mw; mw; dhr *(Mrs; mrs; sir)*	$60 \leq t \leq 75$
Patiënte; Zij; Patient; Patiënt; Pte; Pt; Dhr; Hij; hr; mevr; Hr; Mevrouw *(Patient; She; Patient; Patient; Pt; Pt; Sir; He; mr; mrs; Mr; Madam)*	$30 \leq t \leq 60$
van; Dochter; unk6817259958247; U; PERSONGPE; PERSONPERSON; ja; pte; ORG*; Ze; pt; MW; Partner; Mvr; Vader; Heer; Meneer; patient *(from; Daughter; unk; U; PERSONGPE; PERSONPERSON; yes; pt; ORG*; She; pt; MRS; Partner; Mrs; Father; Mister; Mister; patient)*	$10 \leq t \leq 30$
unk2996650053260; unk274788874542; unk2072351197669; Me; G; nefroloog; patiente; Het; Dr; V; unk3373371847810; en; Patiente; unk8794409766950; patiënt; Er; Ik; ,; huisarts; .; ze; Ja; Alg; unk1856888768377; unk1964271219986;hij; Moeder; Arts; Echtgenote; M; moeder; unk3195551421068 *(unk; unk;unk; Me; G; nephrologist; patient; It; Dr; V; unk; and; Patient; unk; patient; There; I; ,; general practitioner; .; she; Yes; Gen; unk; unk; Mother; Doctor; Wife; M; mother; unk)*	$4 \leq t \leq 10$

Table 1: Tokens that were predicted four times or more on the final anonymization test set with **seen** sentences. *ORG is the sum of various hospital names that were predicted

Token	Times predicted t
Person	180
GPE	118
Mw; mw *(Mrs; mrs)*	$60 \leq t \leq 75$
Meneer; Dhr; Patiënt; Hij; dhr; Mevrouw; Hr; mevr; hr *(Mister; Sir; Patient; He; sir; Madam; Mr; mrs; mr)*	$30 \leq t \leq 60$
Mvr; pte; S; unk3209426582247; Dochter; hij; U; mevrouw; PERSONPERSON; ja; Vader; patiënt; meneer; ORG*; moeder; PERSONGPE; Partner; MW; dhr; pt; Ze; Heer; Patiënte; Pt; patient; Pte; Zij; Patient *(Mrs; pt; S; unk; Daughter; him; YOU; madam; PERSONPERSON; yes; Father; patient; mister; ORG*; mother; PERSONGPE; Partner; MRS; sir; pt; She; Mister; Patient; Pt; patient; Pte; She; Patent)*	$10 \leq t \leq 30$
man; Vandaag; Zoon; Het; patiënte; M; unk4489617618388; :; ze; zij; hij; K; unk8229495895701; F; V; Er; Hr; unk6235465096554; unk1341059269723; ,; Ik; Echtgenoot; Ja; Echtgenote; Moeder; patiente; van *(husband; Today; Son; It; patient; M; unk; ;; she; she; he; K; unk; F; V; There; Mr; unk; unk; ,; I; Husband; Yes; Wife; Mother; patient; from)*	$4 \leq t \leq 10$

Table 2: Tokens that were predicted four times or more on the final anonymization test set with **unseen** sentences.
*ORG is the sum of various hospital names that were predicted

7. Acknowledgements

We would like to thank the privacy office of the AUMC for taking the time to meet with us, taking an interest in our research and scrutinizing the process. Also, we would like to thank Edwin Geleijn for taking part in the meetings we had with the privacy office. The GPUs used for the creation of MedRoBERTa.nl were financed by the NWO Spinoza Project assigned to Piek Vossen (project number SPI 63-260).

8. Bibliographical References

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Marimon, M., Gonzalez-Agirre, A., Intxaurrondo, A., Rodriguez, H., Martin, J. L., Villegas, M., and Krallinger, M. (2019). Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In *IberLEF@ SEPLN*, pages 618–638.

Mireshghallah, F., Taram, M., Vepakomma, P., Singh, A., Raskar, R., and Esmaeilzadeh, H. (2020). Privacy in deep learning: A survey. *arXiv preprint arXiv:2004.12254*.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in bertology: What we know about how

bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Shickel, B., Tighe, P. J., Bihorac, A., and Rashidi, P. (2017). Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604.

Shokri, R. and Shmatikov, V. (2015). Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE.

Stubbs, A., Kotfila, C., and Uzuner, Ö. (2015). Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics*, 58:S11–S19.

Uzuner, Ö., Sibanda, T. C., Luo, Y., and Szolovits, P. (2008). A de-identifier for medical discharge summaries. *Artificial intelligence in medicine*, 42(1):13–35.

9. Language Resource References

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., and Nissim, M. (2019). BERTje: A Dutch BERT model. *ArXiv*, abs/1912.09582.

Delobelle, P., Winters, T., and Berendt, B. (2020). RobBERT: a Dutch RoBERTa-based language model. In *EMNLP*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Van Noord, G., Bouma, G., Van Eynde, F., De Kok, D., Van der Linde, J., Schuurman, I., Sang, E. T. K., and Vandeghinste, V. (2013). Large scale syntactic annotation of written dutch: Lassy. In *Essential speech and language technology for Dutch*, pages 147–164. Springer, Berlin, Heidelberg.

Verkijk, S. and Vossen, P. (2022). MedRoBERTa.nl: A language model for Dutch electronic health records. *Computational Linguistics in the Netherlands Journal*, 11.