



Universiteit
Leiden
The Netherlands

De-identifying clinical texts using biomed-clinical BERTs and comprehensive risk assessment

Shaji, D.; Paul, A.; Han, L.; Del-Pinto, W.; Nenadic, G.; Verberne, S.

Citation

Shaji, D., Paul, A., Han, L., Del-Pinto, W., Nenadic, G., & Verberne, S. (2025). De-identifying clinical texts using biomed-clinical BERTs and comprehensive risk assessment. *2025 Ieee 13Th International Conference On Healthcare Informatics (Ichi)*, 683-684.
doi:10.1109/ICHI64645.2025.00098

Version: Publisher's Version

License: [Licensed under Article 25fa Copyright Act/Law \(Amendment Taverne\)](#)

Downloaded from: <https://hdl.handle.net/1887/4295563>

Note: To cite this publication please use the final published version (if applicable).

De-identifying Clinical Texts using Biomed-Clinical BERTs and Comprehensive Risk Assessment

Dhivin Shaji^{†,1}, Angel Paul^{†,1}, Lifeng Han^{*,1,2}, Warren Del-Pinto¹, Goran Nenadic¹ and Suzan Verberne²

¹ The University of Manchester, UK ² LIACS & LUMC, Leiden University, NL

dhivinshaji, angelpaulm17@gmail.com warren.del-pinto, g.nenadic@manchester.ac.uk

[†]Co-first Authors ^{*}Corresponding Author l.han@lumc.nl | s.verberne@liacs.leidenuniv.nl

Abstract—This work focuses on the de-identification and risk assessment of clinical text using Transformer-based biomedical models BioBERT and ClinicalBERT. The primary objective is to accurately *identify* and *mask* sensitive entities while maintaining data utility for research. We conduct a detailed *comparative* analysis between BioBERT, ClinicalBERT, and other models, including NER-BERT, NER-CRF, and LSTMs. We integrate BioBERT and ClinicalBERT into the MASK framework to facilitate effective de-identification (MASK-BioBERT / ClinicalBERT). We introduce a risk assessment tool that evaluates the potential re-identification risks of clinical data post-de-identification. The developed methods contribute to safeguarding patient *privacy* while supporting the usability of healthcare data for medical research. Our models and platform are publicly available at <https://github.com/angelpaulm17/DeIDClinic>

Index Terms— Clinical De-Identification, Patient Privacy, Deep Learning Models, BioBERT, ClinicalBERT, MASK Framework

I. INTRODUCTION AND BACKGROUND

In healthcare, data privacy is imperative, particularly when dealing with sensitive patient information found in clinical records. With the increasing digitisation of healthcare systems, the risk of exposing personally identifiable information (**PII**) has grown significantly. Clinical records, especially those in free-text form, contain highly sensitive personal information such as patient names, locations, dates, and medical histories [1]. Legislations like the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in Europe mandate stringent measures to protect patient data. Handling sensitive information requires secure methods to anonymise this data, ensuring it can be used for research without compromising privacy [2]. De-identification is a method used to reduce the risk of **re-identification** and to support data **privacy**, especially in sensitive domains like healthcare. It involves masking identifiable information, such as patient names, dates, and medical identifiers to protect privacy.

Structured data can be **anonymised** using conventional techniques, but free-text data, prevalent in clinical records, presents a unique challenge due to its unstructured nature. This work is motivated by the need to *develop robust methods for automatically detecting, and anonymising sensitive information in free-text clinical records, while maintaining the utility of the data*. One such method, **MASK**, offers a flexible framework specifically designed for the automatic de-identification of clinical texts. MASK enables users to choose from various

de-identification techniques, such as *redaction* and *replacement*, ensuring privacy while preserving the clinical context of the data [3].

Advances in Natural Language Processing (**NLP**) and machine learning (**ML**) techniques offer new opportunities for tackling this challenge. Transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) and its specialised variants, **BioBERT** and **ClinicalBERT**, have demonstrated significant promise in identifying sensitive information within clinical texts. This work investigates if the accuracy of de-identifying free-text clinical data can be improved by leveraging MASK alongside domain-specific Transformer-based models while maintaining the data's utility for research. This work is based on our preprint [4] where we detailed the system architecture and the open-source software.

II. METHODOLOGY AND IMPLEMENTATION

The system architecture, as depicted in the diagram (Figure 1), is divided into three major sections: Entity Recognition, Masking, and Risk Assessment. The clinical letters serve as the input to the system, and they are processed through multiple pipelines before final redacted or replaced documents are produced. Key steps are as follows: 1) **Data Ingestion**: Clinical letters are fed into the system via the user interface (UI), allowing users to upload documents in bulk for de-identification. 2) **Entity Recognition**: Several techniques were employed to identify sensitive information from the clinical letters, such as names, professions, dates, ages, and locations. To accommodate the complexity of clinical data, multiple approaches were integrated: dictionary lookup, rule-based search, and PLM models Bio/ClinicalBERT. 3) **Union of Entities**: Once the various methods of entity recognition are applied, the system combines the identified entities into a unified list for further processing, ensuring comprehensive entity recognition. 4) **Masking Strategies**: Users can choose between *Redaction* and *Replacement*. Redaction involves replacing the identified entities with placeholders (e.g., "XXX-Name"). Replacement, on the other hand, substitutes sensitive information with synthetically generated or random replacements to maintain the structure of the original document. Replacement is more suitable for contexts where document coherence needs to be preserved, such as for clinical research purposes. The Replacement mapping is stored and later used for reference in the risk

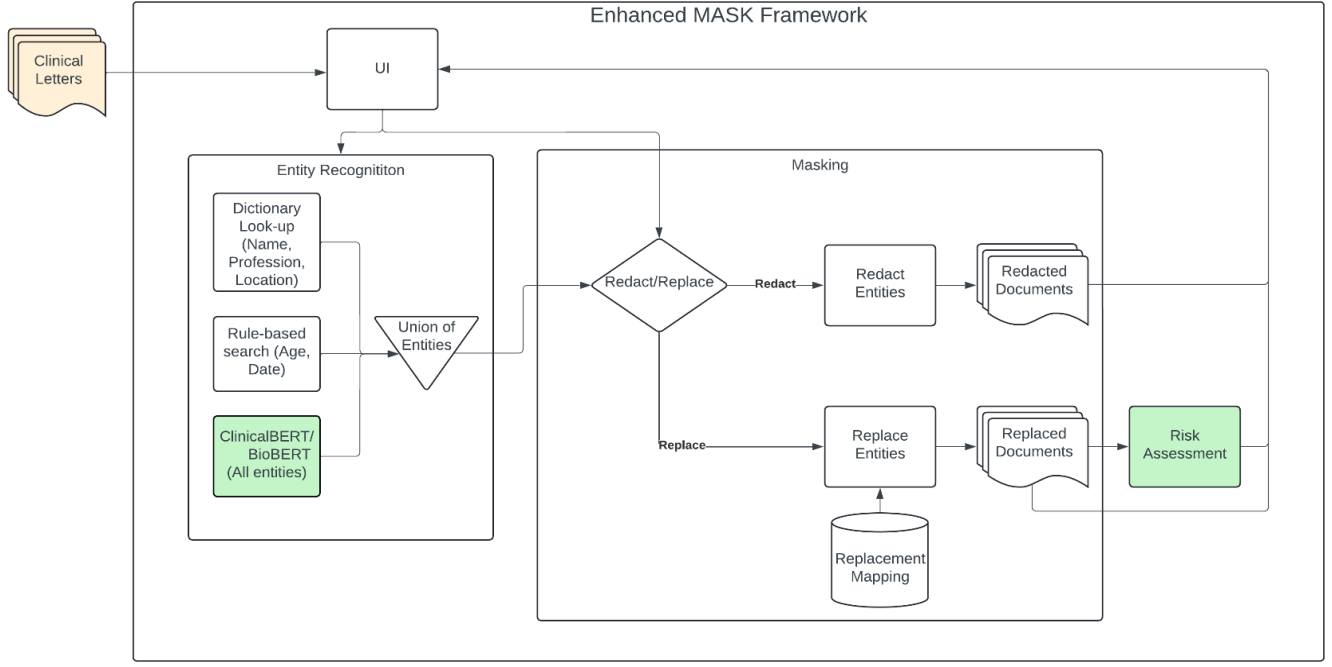


Fig. 1. DEIDCLINIC Architecture of Enhanced MASK framework with Bio/ClinicalBERT

Entity	BioBERT			ClinicalBERT		
	P	R	F1	P	R	F1
NAME	0.963	0.960	0.985	0.970	0.970	0.970
DATE	0.974	0.974	0.974	0.960	0.982	0.971
ID	0.948	0.973	0.988	0.986	0.993	0.989
AGE	0.956	0.978	0.967	0.930	0.974	0.951
LOCATION	0.933	0.903	0.928	0.919	0.922	0.921
CONTACT	0.955	0.928	0.941	0.947	0.960	0.954
PROFESSION	0.810	0.630	0.748	0.844	0.643	0.730
Micro Avg	0.959	0.964	0.965	0.955	0.963	0.959
Macro Avg	0.817	0.793	0.804	0.820	0.805	0.811
Weighted Avg	0.958	0.961	0.964	0.953	0.963	0.958

TABLE I

EVALUATION OUTCOMES WITH HIGHER SCORES BOLDDED

assessment stage. This dual masking strategy aims to provide users with flexibility in how sensitive information is treated. 5) **Risk Assessment:** The probability of re-identification of the de-identified data is evaluated, ensuring that the transformation process sufficiently anonymises sensitive information. 6) **Final Output:** Depending on the chosen masking strategy, the system generates either redacted or replaced documents. These documents are returned to the user via the UI, alongside a risk assessment report. We show the evaluation results in Table I, for which the dataset used is the data set i2b2 / UTHealth Deidentification and Heart Disease Risk Factors, specifically the 2014 PHI Gold Set 1 and 2, which is part of the National NLP Clinical Challenges (n2c2) initiative (<https://n2c2.dbmi.hms.harvard.edu>).

III. CONCLUSIONS AND FUTURE WORK

This work set out to address the critical challenge of protecting patient privacy in clinical free-text data through intro-

ducing the domain-tuned and adaptable *Biomed-Clinical BERT* models for de-identification, i.e. MASK-BioBERT and MASK-ClinicalBERT. A novel implementation of *risk assessment* was also introduced, allowing the system to flag documents with higher re-identification risks. Our fine-tuned models and platform are named as DeIDClinic. Future work includes a detailed error analysis on these two Transformer models.

ACKNOWLEDGEMENT

Funded by the European Union under Horizon Europe Work Programme 101057332. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. The UK team are funded under the Innovate UK Horizon Europe Guarantee Programme, UKRI Reference Number: 10041120.

REFERENCES

- [1] K. El Emam, F. K. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J.-P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, and J. Bottomley, "A globally optimal k-anonymity method for the de-identification of health data," *Journal of the American Medical Informatics Association*, vol. 13, no. 5, pp. 556–569, 2006.
- [2] P. Voigt and A. Von dem Bussche, *The EU General Data Protection Regulation (GDPR). A Practical Guide*, 1st ed. Springer International Publishing, 2017.
- [3] N. Milosevic, G. Kalappa, H. Dadafarin, M. Azimae, and G. Nenadic, "Mask: A flexible framework to facilitate de-identification of clinical texts," *arXiv preprint arXiv:2005.11687*, 2020.
- [4] A. Paul, D. Shaji, L. Han, W. Del-Pinto, and G. Nenadic, "Deidclinic: A multi-layered framework for de-identification of clinical free-text data," 2024. [Online]. Available: <https://arxiv.org/abs/2410.01648>