

Experimental evaluation of bidirectional encoder representations from transformers models for de-identification of clinical document images

Ravichandra Sriram, Siva Sathya Sundaram, S. LourduMarie Sophie

Department of Computer Science, Pondicherry University, Pondicherry, India

Article Info

Article history:

Received Apr 26, 2024

Revised Jan 28, 2025

Accepted Mar 5, 2025

Keywords:

Clinical de-identification

Deep learning

Natural language processing

Protected health information

Support vector machines

Tesseract character recognition

ABSTRACT

Many health institutes maintain patients' diagnosis and treatment reports as scanned images. For healthcare analytics and research, large volumes of digitally stored patient information have to be accessed, but the privacy requirements of protected health information (PHI) limit the research opportunities. Particularly in this artificial intelligence (AI) era, deep learning models require large datasets for training purposes, which hospitals cannot share unless the PHI fields are de-identified. Manual de-identification is beyond possible, with millions of patient records generated in hospitals every day. Hence, this work aims to automate the de-identification of clinical document images utilizing AI models, particularly pre-trained bidirectional encoder representations from transformers (BERT) models. For the purpose of experimentation, a synthetic dataset of 550 clinical document images was generated, encompassing data obtained from diverse patients across multiple hospitals. This work presents a two-stage transfer learning approach, initially employing Tesseract character recognition (OCR) to convert clinical document images into text. Subsequently, it extracts PHI fields from the text for de-identification. For the purpose of extraction, BERT models were utilized; in this work, we contrasted six pre-trained versions of such models to examine their effectiveness and achieve the F1 score of 92.45%, thus showing better potential for de-identifying PHI data in clinical documents.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Ravichandra Sriram

Department of Computer Science, Pondicherry University

Pondicherry, India

Email: ravichandrasriram@gmail.com

1. INTRODUCTION

Patients' clinical files contain records of medications and therapies that have already been administered to them. These records play an important role in improving the accuracy of patient diagnoses and overall care. Furthermore, they are valuable resources in medical research, providing important insights into the patient's state. Prior to sharing these medical records with researchers for clinical studies, the protected health information (PHI) must be removed or anonymized. This comprises the patient's name, age, and contact information. This de-identification need stems from the sensitive nature of patient data. In the United States, HIPAA [1], or the Health Insurance Portability and Accountability Act, specifies eighteen different types of PHI that must be carefully de-identified before being shared with third parties. Furthermore, several artificial intelligence (AI) models used to analyze healthcare records require large amounts of data for training. Hospitals do not share clinical data due to privacy policies, and all patient

records contain PHI elements. These PHI components must be anonymized or de-identified before being provided for research purposes. However, only a limited number of healthcare professionals manually de-identify clinical document images using specialized tools for editing. This procedure is ineffective for both time and expense. Consequently, many digital copies of these healthcare documents remain unused.

This work presents a method that utilizes pre-trained bidirectional encoder representations from transformers (BERT) models for the automated detection and elimination of personally identifying information in clinical document images. Medical documents exhibit uneven organizational frameworks, integrating various data forms including large headings, logos, single-column and dual-column key-value pairs, as well as multi-column tables with and without borders. This diversity in document layouts presents challenges in accurately detecting PHI details within the documents. The language representation paradigm, BERT [2], is specifically developed to pre-train deep bidirectional representations from unlabeled text. The pre-training process involves simultaneous consideration of both the left and right contexts throughout all layers. As a result, the BERT model can be effectively fine-tuned by incorporating a singular output layer. This enables the development of state-of-the-art models for various tasks, such as question answering and language inference, without necessitating substantial modifications to the underlying architecture specific to each task.

The literature review conducted so far did not reveal any previous studies on de-identifying clinical documents stored as images. The aforementioned models have been subjected to testing utilizing publicly available clinical records datasets, including MIMIC– II [3], MIMIC– III [4], i2b2 Corpus 2006 [5], i2b2 Corpus 2014 [6], and CEGS N-GRID Corpus 2016 [7]. The datasets in question consist of various free-text and electronic health records (EHR) information, unlike the documents we have considered in this research, which are pure images. This research focuses on the de-identification and anonymization of PHI fields present in patients' clinical documents. Given the existing gaps in the literature, the research addresses the following key scientific challenges to tackle this problem effectively: i) extraction of content from scanned images, ii) identifying the PHI fields and their location within the scanned clinical documents/image, and iii) de-identification/anonymizing the PHI fields.

Since 1996, numerous researchers have committed to investigating the potential for automating the process of de-identifying healthcare records. Several computational models have been devised, such as conditional random fields (CRFs), hidden Markov models (HMM), decision trees (DTs), and support vector machines (SVMs). These systems utilize machine-learning methodologies that are rule-based, knowledge-based, or data-driven. Deep learning models such as long short-term memory (LSTM), gated recurrent unit (GRU), and bidirectional encoder representations from transformers (BERT) have demonstrated more favorable outcomes. This section delves into various de-identification models employed on publicly accessible datasets. Reference [8] thoroughly examines the use of deep learning models in clinical text de-identification, offering a comprehensive assessment of the topic. This survey presents a selection of literary works for examination. The authors in [9] utilized a bidirectional LSTM model that strongly resembles the model established by [10] to achieve de-identification. The bidirectional LSTM model is highly prevalent in natural language processing (NLP) and demonstrates exceptional performance in a range of named entity recognition (NER) tasks. Previous research investigations [11], [12] have also utilized Bi-LSTM for the purpose of clinical de-identification. This approach allows for estimating probabilities associated with several potential labels for personally identifiable health information (PHI) for every token in the sequence.

Furthermore, the authors in [2] proposed a pre-trained deep bidirectional transformer for the language model, referred to as the BERT model. This model can simultaneously contextualize word embeddings by incorporating all adjacent contexts, thereby making major contributions to the field of NLP across multiple tasks. In another work by [13], fine-tuning of a bidirectional encoder representation model to attain the most advanced level of de-identification performance for electronic health data was performed. This model demonstrates a high level of efficacy in removing patient identifiers, including but not limited to name, age, and social security number. Though various techniques have been adopted, no single model could de-identify all the PHI fields, particularly from Clinical documents that are in the form of images. This forms the basis of this research, which tries to experimentally evaluate some of the popular pre-trained BERT models previously used for data extraction. This research aims to use BERT for the purpose of PHI identification and anonymization, which is not found in the existing literature. Since several versions of BERT are available, it is necessary to compare them with proper experimentation and conclude the efficacy of the best one so that future research can rely on it for the de-identification process. The performance of the methods is evaluated using generally established metrics such as positive predictivity and F1 Score.

The subsequent sections of the paper are structured in the following manner: section 2. introduces the proposed method, section 3 outlines the experimental setup and displays the obtained results, and section 4 concludes the work.

2. METHOD

The proposed system looks for prospective techniques for automatically de-identifying clinical documents stored as images. In this regard, several techniques were analyzed. Among them, conventional word-level vector representations, like word2vec [14], GloVe [15], and fastText [16], reduce all potential semantics of a word into a single vector, unable to discern different word senses depending on the context in which they appear. ELMo [17] and BERT [2] offer robust solutions for contextualized word representations compared to other language models. ELMo achieves context sensitivity by pre-training on extensive text corpora, generating unique embeddings for each word in a given sentence. These embeddings are subsequently utilized in downstream tasks. In contrast, BERT is a more complex model with more parameters, affording it enhanced representational capabilities. Notably, BERT goes beyond providing mere word embeddings; it can be seamlessly integrated into specific tasks and fine-tuned for enhanced task-specific performance. Overall, BERT has demonstrated superior performance compared to ELMo and non-contextual embeddings across various applications, including those in the medical field [18]. With these advancements, we focus exclusively on BERT [2] in this work, omitting ELMo [17] and non-contextual embedding methods. The framework for automatic de-identification of clinical document images is shown in Figure 1, comprising several steps which are explained in this section.

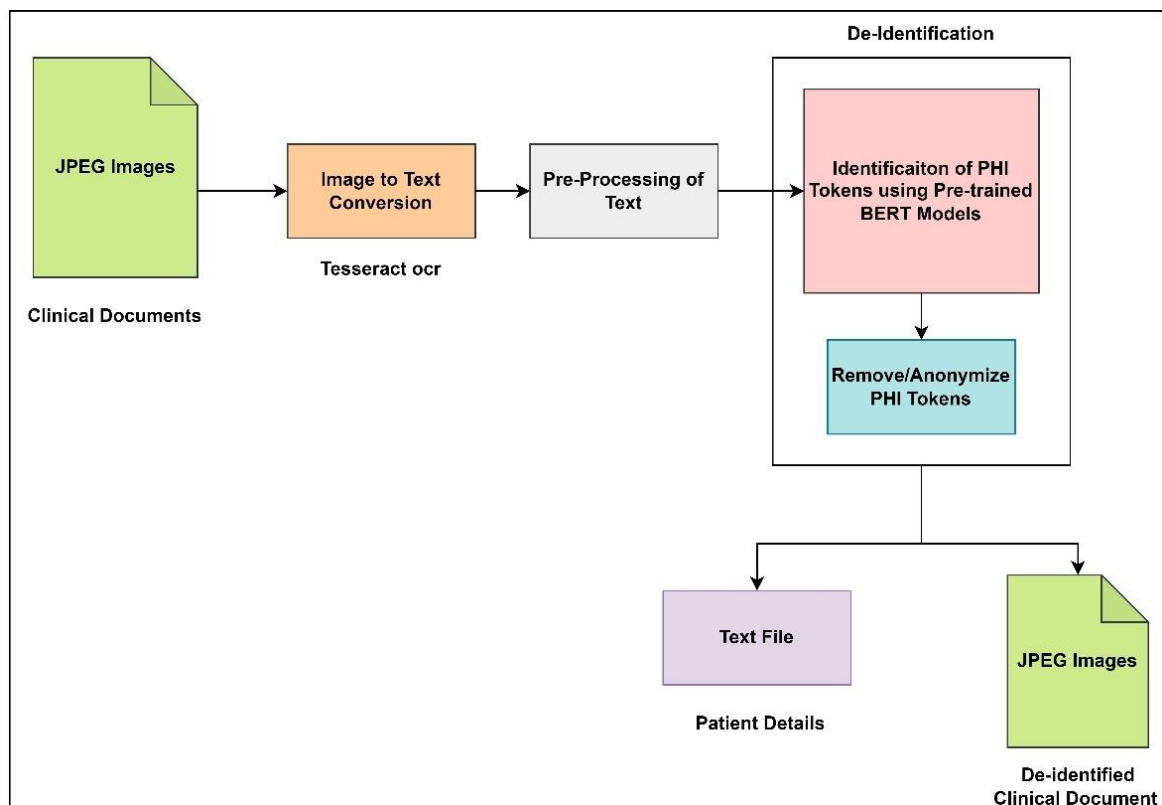


Figure 1. Framework of clinical document image de-identification

2.1. Image-to-text conversion

BERT models process only text data, so the clinical document image must be converted to text using Tesseract OCR. Tesseract OCR converts the image to text and provides text as tokens, providing each token location in the input image. These token locations are useful while removing tokens from the image if tokens are identified as PHI tokens.

2.2. Pre-processing

Various text pre-processing tasks are performed in this step to clean the converted text. The tasks performed are removing punctuation marks from the text, such as commas, periods, question marks, etc. Punctuation removal helps to eliminate unnecessary symbols and make the text more clean and simple. Normalization of text by performing case conversion helps reduce text variability and make it more

consistent. Correcting spelling errors or typos helps improve text quality and readability and avoids confusion or misunderstanding.

2.3. De-identification

Here, we adopted the model architecture of BERT with additional components. To identify PHI tokens, the result of the final layer of BERT is given to a fully connected dense layer with seven outputs for performing named entity recognition, where seven include the six PHI categories as specified in Table 2 plus one non-PHI type. Our experiment aims to evaluate the performance of various pre-trained BERT models in identifying PHI Tokens of clinical documents. Table 1 lists the various BERT models utilized in this work. The sequence of pre-processed text tokens is fed to these pre-trained BERT models to recognize whether a token is one of the six PHI categories or a non-PHI type. If a token is identified as PHI, it is removed from the clinical document image with the help of the Token location stored at the time of image-to-text conversion. These identified PHI tokens are stored in a text file for future re-identification.

Table 1. Pre-trained models employed in this work

Pre-trained model	Description
BERT _{base} [2]	Base BERT is pre-trained with 110 million parameters with uncased tokens, also pre-trained with Books Corpus (800M words) [19] and English Wikipedia (2.5 B words)
BioBERT [20]	It is pre-trained with PubMed abstracts and Central articles
SciBERT [21]	It is pre-trained with academic corpus from the semantic scholar
ClinicalBERT [22]	It is pre-trained with MIMIC-III v 1.4 database
RoBERTa [23]	Pre-trained with CC-NEWS [24], OpenWebText, and Stories [25]
BERT by Johnson [13]	Pre-trained with i2b2 2006 Corpus [5], i2b2 2014 Corpus [6], PhysioNet Corpus [26], and Demoncourt lee Corpus datasets [9]

3. RESULTS AND DISCUSSION

3.1. Clinical document image dataset

A synthetic dataset has been created for experimental purposes, consisting of clinical document images obtained from various hospital patients. The reports were acquired with the patient's explicit permission; nonetheless, the dataset remains inaccessible to the general public owing to the confidential nature of the data. The dataset used in this work is a synthetic collection consisting of 550 images obtained from 46 individuals. This dataset encompasses a variety of document types, such as discharge summaries, radiology reports, biochemistry, hematology, laboratory findings, and more. The files are accessible as JPEG images, employing the RGB color system. The structure of the clinical document image exhibits a considerable degree of complexity. The layout of each image is distinct and contingent upon the specific type of report, with reports originating from the same organization exhibiting diverse patterns. Implementing HIPAA-compliant substitutions and deleting personally identifiable health information (PHI) identifiers inside the dataset enhances its appropriateness for automated de-identification in NLP research. Table 2 presents the dataset's probability distribution of PHI.

Table 2. PHI distributions of clinical document image dataset

Category	Sub-category	No of tokens	Category	Sub-category	No of tokens
Name	Patient	550	Age	Age	550
	Doctor	495	Date	Date	784
	Hospital	378	Contact	Phone	245
	Organization	172		Mobile	438
Location	Dno	126		Fax	47
	Street	412		Email	389
	City	550	ID	Patient ID	241
	State	316		UHID	243
	Country	149		IPNO	388
	Pincode	478			

3.2. Evaluation

The performance of the different models is assessed by computing essential metrics such as F1 Score, precision, and recall. Of the metrics considered, sensitivity is regarded as the most critical indicator for patient de-identification since it shows the proportion of accurately identified individuals that the model appropriately annotates. The pre-processing and evaluation of individual pre-trained BERT models were conducted using the Clinical Document Images dataset, with all computations performed on a single

NVIDIA Quadro GV 100 24 GB GPU. The evaluation of various pre-trained models, including BERT_{base}, BioBERT, SciBERT, ClinicalBERT, RoBERTa, and the BERT-based model proposed by [13], is conducted using a dataset of clinical document images. The findings of clinical document image de-identification are depicted in Figure 2. The tabulation of the performance outcomes of the de-identification approach may be seen in Table 3. BERT by Johnson [13] model achieved F1 Score of 92.4. As stated in the literature review, no researcher has yet to de-identify clinical document images. Previously, all research focused on de-identifying clinical free text or electronic records. With this proposed model, we can de-identify clinical document images and use them for clinical research.

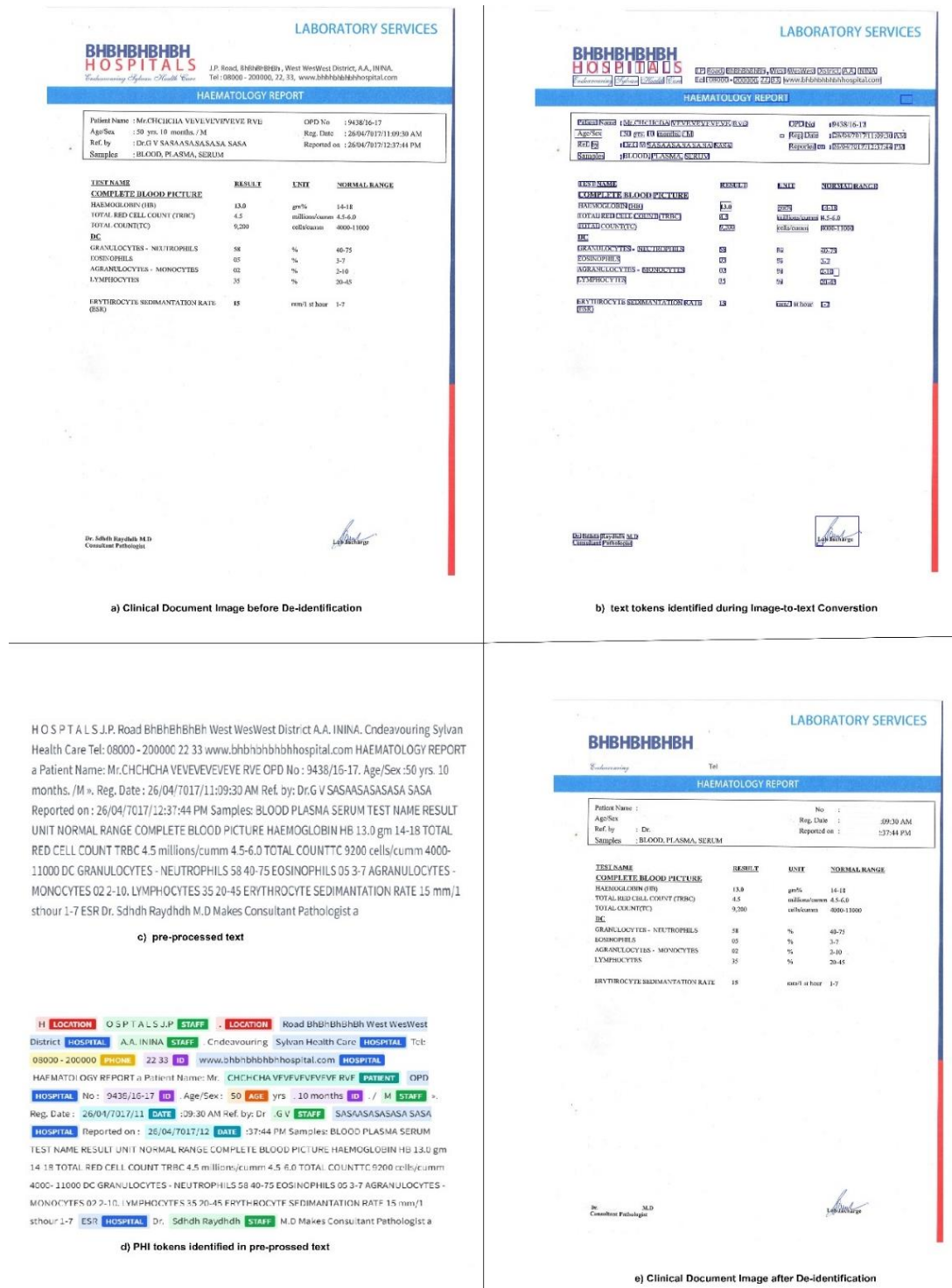


Figure 2. Sample outputs of clinical document image de-identification

Experimental evaluation of bidirectional encoder representations ... (Ravichandran Sriram)

Table 3. Performance of de-identification on clinical document images dataset

Pre-trained model	Precision	Recall	F1 Score
BERT _{base} [2]	84.56	83.15	84.78
BioBERT [20]	86.87	85.21	86.05
SciBERT [21]	91.4	87.4	89.36
ClinicalBERT [22]	88.9	92.1	90.45
RoBERTa [23]	90.75	89.5	90.12
BERT by Johnson [13]	93.89	91.06	92.45

3.4. Discussion

The overall performances of all pre-trained models except BERT by Johnson [13] are nearby and have a difference of 1% with respect to all metrics. BERT by Johnson [13] model has a high F1 Score of 92.45%. Compared to other pre-trained models, BERT by Johnson [13] is a trained model on top of BERT_{base} by [2] with de-identification datasets of i2b2 2006 corpus [5], i2b2 2014 corpus [6], Physionet corpus [26] and Dernoncourt-lee corpus [9]. This clinical-related specific training gave the best performance. But BERT by [13] had achieved an F1 Score of 98.82 on the i2b2 2014 challenge [6] test set. The clinical information on these document images is not uniformly structured. The layouts of these documents change from one health institute to another, and so do the types of reports. When images are converted to text, the information is not exactly transferred. The text in these documents may have different font styles and sizes. Also, the text is not uniformly aligned. Some issues in capturing physical documents into images include the degradation of a few documents due to age. The header and footer of these document images have the health institute's name, address, contact details, and logo, with different font styles and sizes. Capturing information from these locations is the major challenge. The performance of BERT by Johnson [13] for individual entities of PHI categories is in Table 4. It shows the reason for the low performance of BERT by Johnson [13] compared to the i2b2 2014 challenge. Identifying location and contact entity types is less accurate due to the nature of these clinical document images.

Table 4. PHI category-wise performance of pre-trained BERT by Johnson [13]

Entity Type	Precision	Recall	F1 Score
Name	96.4	98.5	97.4
Location	83.56	78.71	81.05
Age	97.25	95.55	96.38
Date	96.45	98.65	97.5
Contact	86.5	83.35	84.9
ID	98.4	96.67	97.5

4. CONCLUSION

This work introduces a comprehensive framework for de-identifying clinical document images utilizing multiple pre-trained BERT models. The methodology consists of three essential steps: first, transforming images into text; second, preparing the text to enhance its format for analysis; and third, employing a pre-trained BERT model to detect and redact PHI tokens. We generated a synthetic dataset of 550 clinical document images from many healthcare organizations to validate the experiment. Our comparative analysis of publicly available pre-trained BERT models revealed notable variations in performance, with the BERT by Johnson model achieving the highest F1 Score of 92.45%. This performance is particularly significant when contrasted with results from established datasets used in the i2b2 2006 and i2b2 2014 de-identification challenges. The superior performance of these datasets suggests that the BERT by Johnson model may offer enhanced capabilities for handling the specific challenges presented by clinical document images. However, the complex layouts typical of clinical documents hindered converting images to text, which often led to incomplete or inaccurate text extraction. This limitation underscores a critical area for future research. Subsequent efforts should aim to enhance the OCR phase to better manage the intricacies of varied document layouts, ensuring a more accurate text extraction. Additionally, future work will explore advanced techniques for automatically extracting and de-identifying PHI from these complex documents, potentially involving integrating AI-driven layout analysis tools with BERT-based text processing methodologies. By advancing these areas, the field can move closer to fully automated, highly accurate systems for the secure handling of sensitive health information, thereby expanding the possibilities for clinical research while rigorously safeguarding patient privacy.

FUNDING INFORMATION

No funding was received.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ravichandra Sriram	✓	✓	✓					✓						
Siva Sathya Sundaram				✓		✓			✓		✓		✓	
S. LourduMarie Sophie					✓	✓	✓			✓				

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author, Ravichandra Sriram. The data, which contain information that could compromise the privacy of research participants, are not publicly available due to certain restrictions.

The data that support the findings of this study are available from the corresponding author, Ravichandra Sriram, upon reasonable request.




REFERENCES

- [1] GovInfo, "Health insurance portability and accountability act of 1996. Public Law 104-191," 1996.
- [2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North*, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
- [3] M. Saeed *et al.*, "Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database," *Critical Care Medicine*, vol. 39, no. 5, pp. 952–960, May 2011, doi: 10.1097/CCM.0b013e31820a92c6.
- [4] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, p. 160035, May 2016, doi: 10.1038/sdata.2016.35.
- [5] O. Uzuner, Y. Luo, and P. Szolovits, "Evaluating the state-of-the-art in automatic de-identification," *Journal of the American Medical Informatics Association*, vol. 14, no. 5, pp. 550–563, Sep. 2007, doi: 10.1197/jamia.M2444.
- [6] A. Stubbs and Ö. Uzuner, "Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus," *Journal of Biomedical Informatics*, vol. 58, pp. S20–S29, Dec. 2015, doi: 10.1016/j.jbi.2015.07.020.
- [7] A. Stubbs, M. Filannino, and Ö. Uzuner, "De-identification of psychiatric intake records: overview of 2016 CEGS N-GRID shared tasks Track 1," *Journal of Biomedical Informatics*, vol. 75, pp. S4–S18, Nov. 2017, doi: 10.1016/j.jbi.2017.06.011.
- [8] R. Sriram, S. S. Sundaram, and S. L. Sophie, "Deep learning models for automatic de-identification of clinical text," in *International Conference on Computer, Communication, and Signal Processing*, 2023, pp. 116–127. doi: 10.1007/978-3-031-39811-7_10.
- [9] F. Demoncourt, J. Y. Lee, O. Uzuner, and P. Szolovits, "De-identification of patient notes with recurrent neural networks," *Journal of the American Medical Informatics Association*, vol. 24, no. 3, pp. 596–606, May 2017, doi: 10.1093/jamia/ocw156.
- [10] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 260–270. doi: 10.18653/v1/N16-1030.
- [11] T. Ahmed, M. M. Al Aziz, and N. Mohammed, "De-identification of electronic health record using neural network," *Scientific Reports*, vol. 10, no. 1, p. 18600, Oct. 2020, doi: 10.1038/s41598-020-75544-1.
- [12] R. Catelli, V. Casola, G. De Pietro, H. Fujita, and M. Esposito, "Combining contextualized word representation and sub-document level analysis through Bi-LSTM+CRF architecture for clinical de-identification," *Knowledge-Based Systems*, vol. 213, p. 106649, Feb. 2021, doi: 10.1016/j.knsys.2020.106649.
- [13] A. E. W. Johnson, L. Bulgarelli, and T. J. Pollard, "Deidentification of free-text medical records using pre-trained bidirectional transformers," in *Proceedings of the ACM Conference on Health, Inference, and Learning*, Apr. 2020, pp. 214–221. doi: 10.1145/3368555.3384455.
- [14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Neural information processing systems*, vol. 1, pp. 1–9, 2006.
- [15] J. Pennington, R. Socher, and C. Manning, "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.




- [16] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, Dec. 2017, doi: 10.1162/tacl_a_00051.
- [17] M. Peters *et al.*, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 2227–2237. doi: 10.18653/v1/N18-1202.
- [18] Y. Si, J. Wang, H. Xu, and K. Roberts, "Enhancing clinical concept extraction with contextual embeddings," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1297–1304, Nov. 2019, doi: 10.1093/jamia/ocz096.
- [19] Y. Zhu *et al.*, "Aligning books and movies: towards story-like visual explanations by watching movies and reading books," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 19–27. doi: 10.1109/ICCV.2015.11.
- [20] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: a pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3613–3618. doi: 10.18653/v1/D19-1371.
- [21] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.
- [22] E. Alsentzer *et al.*, "Publicly available clinical BERT embeddings," *arXiv preprint arXiv:1904.03323*, 2019.
- [23] Y. Liu *et al.*, "RoBERTa: a robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019, [Online]. Available: <http://arxiv.org/abs/1907.11692>
- [24] J. Mackenzie, R. Benham, M. Petri, J. R. Trippas, J. S. Culpepper, and A. Moffat, "CC-news-en," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, Oct. 2020, pp. 3077–3084. doi: 10.1145/3340531.3412762.
- [25] A. Gokaslan and V. Cohen, "Open web text corpus," *GitHub*, 2019. <https://skylion007.github.io/OpenWebTextCorpus/> (accessed Jan. 16, 2024).
- [26] I. Neamatullah *et al.*, "Automated de-identification of free-text medical records," *BMC Medical Informatics and Decision Making*, vol. 8, no. 1, p. 32, Dec. 2008, doi: 10.1186/1472-6947-8-32.

BIOGRAPHIES OF AUTHORS






Ravichandra Sriram    is presently a research scholar in the Department of Computer Science at Pondicherry University, Puducherry. He completed his B.Tech in Information Technology from Swarnandhra College of Engineering and Technology, Narsapur in 2007 and M.Tech in Computer Science and Technology at Andhra University, Visakhapatnam in 2010. He worked as an assistant professor at Shri Vishnu Engineering College for Women, Bhimavaram from 2011 to 2019. His research interests include computer vision, natural language processing, and deep learning. He has authored and co-authored some publications which include journals and international conferences in the field of computer science. He can be contacted at ravichandrasriram@gmail.com.



Siva Sathya Sundaram    completed her M.Tech and Ph.D. in Computer Science and Engineering from Pondicherry University. She has 25 years of teaching experience and specializes in evolutionary and bio-inspired computing, spatio-temporal data mining, VANET and natural language processing. She is UGC NET qualified and has published several research articles in reputed journals. She is the recipient of the Naari Shakthi Award 2017 from the President of India for her mobile Innovation "MITRA" for Women safety. She has also received the Chairman's Distinction award in South Asia's mBillionth Mobile Innovation contest. She can be contacted at ssivasathya@pondiuni.ac.in or ssivasathya@gmail.com.



S. LourduMarie Sophie    is presently a research scholar in the Department of Computer Science and Engineering at Pondicherry University, Puducherry. She completed her B.Tech in Computer Science and Engineering from Manakula Vinayagar Institute of Technology, Puducherry in 2015 and her M.Tech. in Computer Science and Engineering at Pondicherry University, Puducherry in 2019. She worked as an Assistant System Engineer at Tata Consultancy Service, Chennai from 2015 to 2017. She also has a year of teaching experience as a guest faculty at Pondicherry University in 2019-2020. She has qualified for the UGC NET examination in 2019. Her research interests include natural language processing, machine learning, and deep learning. She has authored and co-authored more than 10 publications which include journals and international conferences in the field of computer science. She can be contacted at lourdumariesophie15@gmail.com.